MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Geografický ústav

Diplomová práce

Brno 2014 **Michal Zimmermann**

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta
Geografický ústav

Možnosti dolování a vizualizace dat ze sociálních sítí

Diplomová práce

Michal Zimmermann

Vedoucí práce: Mgr. Bc. Zdeněk Stachoň, Ph.D. **Brno 2014**

Bibliografický záznam

Autor:	Michal Zimmermann Přírodovědecká fakulta, Masarykova univerzita Geografický ústav

Název práce:	Možnosti dolování a vizualizace dat ze sociálních sítí
Studijní program:	Geografie a kartografie
Studijní obor:	Geografická kartografie a geoinformatika
Vedoucí práce:	Mgr. Bc. Zdeněk Stachoň, Ph.D.
Akademický rok:	2013/14
Počet stran:	$74 + 13$
Klíčová slova:	sociální sítě; dolování dat; API; Twitter; Foursquare; Instagram; časoprostorová kostka; open source

Bibliographic Entry

Author:	Michal Zimmermann Faculty of Science, Masaryk University Department of Geography
Title of Thesis:	Mining and Cartographic Visualization of Social Networks Data
Degree Programme:	Geography and Cartography
Field of Study:	Geographical Cartography and Geoinformatics
Supervisor:	Mgr. Bc. Zdeněk Stachoň, Ph.D.
Academic Year:	2013/14
Number of Pages:	$74 + 13$
Keywords:	social networks; data mining; API; Twitter; Foursquare; Instagram; space time cube; open source

Abstrakt

Diplomová práce představuje možnosti dolování prostorových dat z online sociálních sítí. Ta jsou získávána prostřednictvím veřejně dostupných aplikačních programovacích rozhraní těchto sítí. Důraz je kladen na automatizaci procesu, jehož výsledkem jsou vizualizace dat vytvořené za pomoci zdarma dostupných nástrojů. Nedílnou součástí práce je rovněž analýza získaných dat a návrh jejich možného využití.

Abstract

The thesis reveals possible ways of acquiring spatial data from online social networks. The data is obtained via public application programming interfaces of these services. We focus on making this process fully automated with our own code or with help of third party libraries. As a result, spatial data visualizations designed with free or open source software are presented. Data analysis, as well as its possible use in the future, is also covered.

MASARYKOVA UNIVERZITA
Přírodovědecká fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Akademický rok: 2013/2014

Ústav: Geografický ústav

Student: Bc. Michal Zimmermann

Program: Geografie a kartografie

Obor: Geografická kartografie a geoinformatika

Ředitel *Geografického ústavu* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s tématem:

Téma práce: Možnosti dolování a vizualizace dat ze sociálních sítí

Téma práce anglicky: Mining and Cartographic Visualization of Social Networks Data

Oficiální zadání: Práce zpracovává aktuální fenomén sociálních sítí z hlediska prostorových informací v nich obsažených. Na základě zjištěných informací bude navržen způsob jejich získávání a kartografické vizualizace. Zmíněna bude také problematika omezení vyplývající z právních a etických důsledků využití zmíněných dat. Navržené přístupy ke kartografické vizualizaci budou prakticky prezentovány na zvolených příkladech. Doporučená literatura: Campbell, S.G. et al, Blogscape: Cartography on Social Networks, dostupné na: http://terpconnect.umd.edu/~susanc/blogscape/BlogScape2.pdf Viégas F. and Donath J. Social Network Visualization: Can We Go Beyond the Graph?, dostupné na http://alumni.media.mit.edu/~fviegas/papers/viegas-cscw04.pdf

Jazyk závěrečné práce: čeština

Vedoucí práce: Mgr. Bc. Zdeněk Stachoň, Ph.D.

Datum zadání práce: 1. 10. 2012

V Brně dne: 11. 11. 2013

Souhlasím se zadáním (podpis, datum):

.....	.....	.....
Bc. Michal Zimmermann	Mgr. Bc. Zdeněk Stachoň, Ph.D.	doc. RNDr. Petr Dobrovolný, CSc.
student	**vedoucí práce**	**ředitel Geografického ústavu**

Poděkování

Na tomto místě bych chtěl poděkovat Mgr. Bc. Zdeňku Stachoňovi, Ph.D. za vstřícnost, cenné připomínky, rady a podněty.

Prohlášení

Prohlašuji, že jsem svoji diplomovou práci vypracoval samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno 6. května 2014 .....

Michal Zimmermann

Obsah

Přehled použitého značení

Pro snazší orientaci v textu zde čtenáři předkládáme přehled základního značení, které se v celé práci vyskytuje.

API	Application Programming Interface
CSS	Cascading Style Sheets
CSV	Comma-separated Values
HTML	HyperText Markup Language
JSON	JavaScript Object Notation
REST	Representational State Transfer
XML	Extensible Markup Language

Úvod

Sociální sítě jsou jedním z nejčastějších cílů uživatelů webu. Vzhledem k různému tematickému zaměření mohou hrát významnou roli při hledání práce, partnera nebo hudby - mají tedy velký potenciál zaujmout návštěvníka a učinit z něj svého stálého uživatele.

Boyd a Ellison (2007) za počátek éry sociálních sítí označují zhruba polovinu 90. let minulého století, kdy začaly vznikat první servery založené na provázané uživatelské komunitě. Od té doby již mnohé sociální sítě zanikly a byly nahrazeny novými, značně se rozšířily jejich možnosti a především se ještě více integrovaly do každodenního života návštěvníků webu.

Infografika publikovaná na serveru Visual.ly (2013) ukazuje, že sociální sítě disponují stamiliony uživatelů, kteří je pravidelně navštěvují. Facebook již překročil hranici jedné miliardy uživatelů, z nichž se na server denně připojí více než polovina. Twitter s více než 500 miliony uživateli zaujímá druhé místo, na Google+ je zaregistrováno 400 milionů uživatelů.

Vzhledem k povaze sociálních sítí, jejichž primárním cílem je sdílení či výměna informací jakékoliv povahy (text, fotografie, videa), na nich každý den vzniká množství dat, mezi nimi samozřejmě i data související s prostorem, v němž se uživatelé nacházejí.

Velký vliv na růst sociálních sítí má masové rozšíření chytrých telefonů, které uživatelům umožňují jednoduše přistupovat k jejich profilům, a především okamžitě sdílet data. Již zmiňovaná infografika uvádí, že 73 % uživatelů chytrých telefonů se alespoň jednou denně připojuje k nějaké sociální síti a že 93 % přenesených dat souvisí s aktivitou na sociálních sítích.

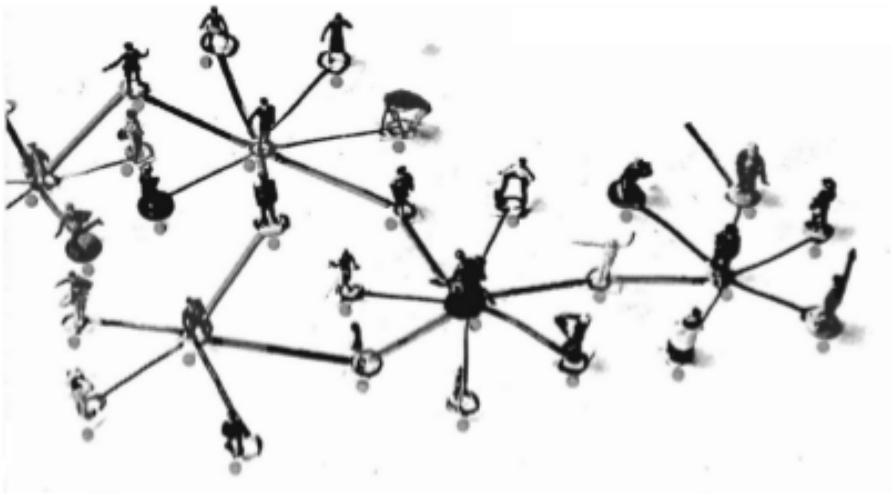
Východiska práce

Sociální sítě jsou zkoumány především ze sociologického hlediska: předmětem výzkumu bývají uživatelé a vztahy, které mezi nimi na sítích vznikají. Problematice vztahu jednotlivých členů sítě[1] se ve své práci věnoval již v 60. letech minulého století S. Milgram (1967), který empiricky zkoumal tzv. *small-world problem*, tedy předpoklad, že mezi každými dvěma členy sítě existuje konečný a poměrně malý počet členů, jejichž spojením vznikne řetězec, na jehož koncích se nacházejí právě původní dva členové. Prostřednictvím dopisů rozeslaných vybraným obyvatelům Spojených

[1] Zde v širším slova smyslu, nejedná se pouze o online sociální sítě.

států amerických se mu podařilo prokázat, že takové řetězce skutečně existují a mohou fungovat, aniž by je nějak výrazně ovlivňovala fyzická vzdálenost mezi prvním a posledním členem. Dnes je tento fenomén označován také jako *six degrees of separation*.[2] Právě propojení jednotlivých členů sítě hraje významnou roli v procesu šíření informací.

Výsledky Milgramova experimentu lze využít také opačně: pokud známe počet lidí s určitou charakteristikou (např. HIV pozitivní) v sociální síti jedince, můžeme odvodit celkový počet těchto „skrytých" populací. Výsledky těchto úvah však mohou být značně zavádějící, neboť zkoumané informace mohou být jednak nepravdivé (Kadushin 2012, s. 110), jednak jsou ovlivněny vybraným vzorkem populace. Jak dokazuje práce Zhenga, Salganika a Gelmana (2006), negativní charakteristiky mají tendence se shlukovat v určitých lokalitách.



Obr. 1: Zobrazení několikastupňové sociální sítě (převzato Milgram, 1967).

Milgram k vizualizaci vztahů v sítích používá grafy (obr. 1), jejichž uzly představují členy sítě a hrany vztahy mezi nimi. Tento způsob zobrazení vztahů je využíván i dnes a pomáhá odhalovat další zákonitosti sociálních sítí. Mislove et al. (2007) zkoumáním sociálních sítí Flickr, Youtube, LiverJournal a Orkut prokázali, že:

- v sítích existuje malé množství významných uzlů a velké množství nevýznamných uzlů[3],
- v sítích existuje velké množství shluků tvořených málo významnými uzly, které jsou spojeny prostřednictvím malého počtu významných uzlů,
- významné uzly navazují spojení mezi sebou a vytvářejí tak *jádro* sítě,
- významnou roli v síti hraje jádro, které je nezbytné pro propojení sítě[4] a skládá se z navzájem silně propojených uzlů.

[2] Ačkoliv z Milgramova pokusu vyplývá, že nejfrekventovanější jsou řetězce o délce pěti členů.

[3] Význam uzlů je přímo úměrný počtu uzlů, se kterými jsou spojeny.

[4] Mislove et al. (2007) uvádí, že při odstranění 10 % jádrových uzlů dochází k rozpadu sítě do milionů velmi malých shluků.

Autoři rovněž uvádějí, že uzly sociálních sítí se například od webových stránek, které můžeme chápat jako uzly v síti WWW, odlišují poměrem odchozích a příchozích spojení. Zatímco na webu existují jak hojně odkazované a zároveň málo odkazující stránky (např. zpravodajské servery), tak stránky málo odkazované a zároveň hojně odkazující, na sociálních sítích jsou počty odchozích a příchozích spojení vyrovnané. Tato teze však platí pouze pro ty sítě, na nichž jsou spojení reciproká - souhlas s příchozím spojením automaticky vytváří zpětné spojení k uzlu. Výjimku tvoří například Twitter (https://twitter.com), který využívá pouze jednostranných spojení.

Ačkoliv je vizualizace sítí pomocí grafů často užívanou metodou, trpí některými nedostatky. Viégas a Donath (2004) jako ty nejzásadnější zmiňují nepřehlednost grafu při zobrazení většího množství dat a omezený pohled na data, který podává informace pouze o topologii sítě. Kvůli překrývajícím se hranám či názvům uzlů může být navíc tato informace zkreslena. Autoři vedle tradičních grafů přicházejí s myšlenkou zasadit vývoj sociální sítě do časového rámce. Uživatel takovéto vizualizace může tedy zpětně vystopovat vznik nových vztahů či zánik těch stávajících (více viz PostHistory, 2004).

Efektivnějšího zobrazení sítě grafem lze podle autorů docílit také aplikováním principů známých z kartografie, z nichž za nejvhodnější považují:

- změnu měřítka *(adaptive zooming)* spojenou se změnou zobrazených dat - pro jednotlivé úrovně přiblížení uživatel dostane různě podrobná data,
- změnu tematické nadstavby *(multiple viewing modes)* - základem zobrazené informace zůstává topologie sítě, obsah nad ní se však může měnit.

Odlišným způsobem přistupují k výzkumu sociálních sítí Yardi a Boyd (2010). Cílem jejich práce je nalézt vztah mezi virtuálním a skutečným prostorem a kvantifikovat vliv událostí odehrávajících se ve skutečném prostoru na uspořádání prostoru virtuálního.

Předmětem výzkumu jsou dvě události veskrze *lokálního* významu: střelba v kansaské Wichitě a pád budovy v Atlantě. Autoři v této souvislosti zdůrazňují schopnost internetu nejen překonávat vzdálenosti mezi kontinenty, ale rovněž spojovat místní komunity, jejichž role při přenosu informací o takovýchto událostech může být klíčová.

Výzkum Yardi a Boyd (2010) rovněž vychází z předpokladu, že lidé žijící nedaleko od sebe sdílejí některé charakteristiky (např. socioekonomický status či věk) a zájmy, a že tedy existují objektivní důvody ke vzniku vazeb mezi nimi také na sociálních sítích.

Na základě těchto tezí autoři prověřovali počet příchozích a odchozích spojení a vzdálenost mezi uzly sítě. Výsledky zkoumání pak skutečně potvrdily původní předpoklady:

- hustota sociálních sítí vzniknuvších okolo lokálních událostí je vyšší než hustota náhodně vzniklé sítě, tzn. existuje zde vyšší počet spojení mezi jednotlivými uzly,
- nejvíce informací pochází od uživatelů, kteří jsou události nejblíže,

- uživatelé hledají informace o události v lokálních zdrojích, kterými mohou být jak lokální média, tak lokální sociální sítě.

S rostoucí hustotou sítě roste rychlost, kterou se informace šíří, problematická však může být jejich pravdivost, v reálném čase obtížně ověřitelná. Informace tak mohou podávat mylný či lživý obrázek o skutečnosti, čímž ovlivňují také způsob, kterým se k události staví široká veřejnost.

Metodika

Zmíněné práce a experimenty dokazují, že uživatelé sociálních sítí jsou ochotni sdílet data související s prostorem. Cílem této kapitoly je navrhnout obecný způsob identifikace, získání a následné vizualizace těchto dat.

Terminologie práce

V části zabývající se podstatou sítí z hlediska sociologie se autor snaží užívat českých termínů. Tam, kde panuje nejistota ohledně překladu, je anglický termín uveden kurzívou v závorce za českým překladem, případně je užito pouze anglického originálu. Rovněž některé z příloh jsou zpracovány v angličtině, neboť byly primárně určeny pro prezentaci převážně anglicky mluvícím uživatelům. Pro lepší přehled uvádíme v tab. 1 anglické termíny související s online sociálními sítěmi spolu s jejich významem.

Tab. 1: Anglické výrazy používané ve spojitosti s online sociálními sítěmi.

Síť	**Výraz**	**Význam**
Foursquare	badge	Virtuální odznak, který uživatel získá po splnění úkolu, například po určitém počtu check-inů ze stejného místa.
Foursquare	check-in	Přihlášení uživatele Foursquare na určitém místě (venue).
Foursquare	mayor	Uživatel, který má na venue nejvíce check-inů
Foursquare	venue	Místo definované v síti Foursquare, na kterém se uživatel může přihlásit (provést check-in).
Twitter	follower	Uživatel, který sleduje tweety daného uživatele.
Twitter	following	Uživatel, jehož tweety sleduje daný uživatel.
Twitter	hashtag	Slovo začínající znakem #, které je na Twitteru chápáno jako klíčové a lze podle něj vyhledávat.
Twitter	tweet	140 znaků dlouhá zpráva odeslaná na Twitter.
Twitter	retweet	Přeposlání tweetu cizího uživatele followerům.

Autorovi není známo, že by zmíněné výrazy měly své ustálené české ekvivalenty. Pomineme-li poměrně nápaditá slova *sleď* a *následovač* pro označení followerů, máme k dispozici pouze původní anglické termíny. Jejich skloňování může v češtině působit

problémy, které by zavedení českých termínů spolehlivě vyřešilo. Je však otázkou, zda v češtině vůbec nalezneme slova, která by jednoduše a přesně vystihovala jejich původní význam. Pro badge se jistě nabízí český *odznak*, k dalším slovům však budeme překlad hledat jen obtížně. Významu slova venue pravděpodobně nejvíce odpovídá české *místo*, autor se přesto domnívá, že ani toto slova zcela nevystihuje původní význam anglického originálu. Stejně tak *přihlášení* bychom mohli považovat za shodné s významem původního slova check-in, těžko bychom však již rozlišovali, zda *přihlášením* máme na mysli skutečné přihlášení (login) uživatele do sítě, či check-in na konkrétním místě. Pro termíny související s Twitterem bude nalezení jednoslovných ekvivalentů nejspíše nemožné; autor právě jednoduchost českých termínů považuje za klíčovou, mají-li uspět v praxi, nemá totiž smysl zavádět složité české názvosloví místo jednoduchého anglického.

Identifikace prostorových dat

Základním předpokladem k získání dat souvisejících s prostorem je možnost jejich odlišení od dat, která prostorovou informaci neobsahují. V prostředí internetu je geografické umístění nejčastěji reprezentováno buď GPS souřadnicemi, nebo textovým řetězcem (např. Kotlářská 2, Brno). Pro zobrazení geografické informace reprezentované textovým řetězcem může být v závislosti na zvolené technologii vizualizace vyžadován její převod do číselného formátu GPS souřadnic. K tomuto úkolu může být využit portál GeoNames (http://geonames.org). Pro vyšší úroveň podrobnosti (ulice, domovní čísla) lze využít Google Geocoding API (2013), je však třeba respektovat podmínky užití služby.

Můžeme předpokládat, že GPS souřadnice jsou k datům většinou připojovány automaticky, a nemohou tak být ovlivněny zásahem člověka. V mezilidské komunikaci je přirozenější užití textové reprezentace místa, ta však může obsahovat chyby (např. překlepy - Kotlářská × Ktlářská). Geokódování takového řetězce končí neúspěchem, pakliže není v nástroji implementován některý z fuzzy string vyhledávacích algoritmů (více viz Hall a Dowling, 1980).

Pro *automatizovanou identifikaci* prostorových dat je tedy určení vztažného bodu pomocí GPS souřadnic vhodnější.

Geokódování názvů prostřednictvím databáze GeoNames

GPS souřadnice však mnohdy ke zkoumanému textu připojeny být nemusí, v takovém případě je tedy nutné využít geokódování k zjištění polohy místa, k němuž se text vztahuje. Geokódováním rozumíme převod zadaného řetězce na dvojici zeměpisných souřadnic (mapy.cz, 2013). Tuto službu nabízí například portály Mapy.cz či Google Maps, dle licenčních podmínek však není povoleno její automatizované využití.

Alternativou k online službám tak může být již zmíněná databáze GeoNames, která obsahuje více než deset milionů geografických názvů z celého světa. Ty jsou členěny do kategorií popsaných v dokumentaci (geonames.org, 2013) a bodově lokalizovány. Pro účely této práce jsou nejdůležitější položky, které mají ve sloupci `feature class` vyplněno „P“ - jedná se o obydlená místa.

Data jsou k dispozici pro jednotlivé státy v podobě textového souboru s hodnotami oddělenými tabulátorem. Zdrojovou sadu pro české názvy tvoří projekt ČVUT FreeGeodataCZ (grass.fsv.cvut.cz, 2012). Stažený textový soubor lze jednoduše importovat do připravené PostgreSQL tabulky (Příloha č. 2)[5]:

Kód 1: Import databáze Geonames do PostgreSQL.

```
1 COPY geonames FROM '/absolutní_cesta/CZ.csv' CSV DELIMITER ' ';
```

Pokud bychom chtěli v průběhu práce využívat prostorové dotazy nebo zobrazovat data v desktopovém GISu, můžeme do tabulky přidat sloupec s geometrií a naplnit ho daty ze sloupců `longitude` a `latitude` (kód 2).

Kód 2: Vytvoření geometrie a prostorového indexu v tabulce geonames.

```
1 -- Vytvoří sloupec geom s datovým typem geometry
2 ALTER TABLE geonames ADD COLUMN geom geometry(Point,4326);
3 -- Sloupec geom naplní geometrií získanou ze sloupců longitude a latitude
4 UPDATE geonames SET geom = ST_SetSRID(ST_MakePoint(longitude, latitude),4326);
5 -- Nad tabulkou vytvoří prostorový index
6 CREATE INDEX idx_geonames_geom ON geonames USING GIST(geom);
```

V předchozí části textu jsme identifikovali dva základní způsoby lokalizace dat. Čtení GPS souřadnic můžeme považovat za bezproblémové, geokódování by za nás v ideálním případě obstarala popsaná databázová tabulka v kombinaci se správně pokládanými dotazy. Ještě před jejím použitím je však třeba poukázat na problém, který se autorovi nepodařilo během práce uspokojivě vyřešit.

Vzhledem k tomu, že čeština je flexivní[6] jazyk, dostaneme se při hledání shod velmi záhy do nesnází. Vše si můžeme ukázat na jednoduché větě „*topologicky čistá data vytvořená státní správou aby člověk pohledal. #gis*“.

Na první pohled nesouvisí s nějakým konkrétním místem. Přesto — v závislosti na nastavení fulltextového vyhledávání — můžeme i v této větě najít jedno české sídlo. Při použití českého slovníku[7] totiž budeme využívat právě flexe češtiny, a tak z lexému *pohled* přidáním přípony -y dostaneme název Pohledy. Takové chování jistě není žádoucí, můžeme tedy změnit jazykové nastavení vyhledávání a využít pro hledání shod angličtinu. Databáze tak sice nevrátí název obce Pohledy, stejně tak nám však nevrátí ani texty, v nichž je název sídla použit v jiném než 1. pádu. Ani tyto výsledky tak nemůžeme považovat za uspokojivé.

Autor za možné řešení považuje manuální definici předpon a přípon v `affix` souboru, potažmo definování tvarů toponym v jednotlivých pádech. Obě tyto činnosti však přesahují rámec práce.

Zavedení českého slovníku a oba způsoby nastavení popisuje Příloha 3.

[5]Znak tabulátoru lze též zapsat jako E'\t'. Při pokusu o import operace pravděpodobně selže na položce s `geonameid` 3076084. Stačí nahradit Gru"gau za Grügau a příkaz spustit znovu.

[6]Flexivní jazyk vyjadřuje gramatické funkce pomocí skloňování, časování, předpon a přípon.

[7]Dostupný na http://www.pgsql.cz/data/czech.tar.gz.

Získání prostorových dat

Automatizované získání prostorových dat může probíhat buď zasíláním požadavků prostřednictvím API, nebo web scrapingem[8]. Výhodou API je standardizovaný přístup k informacím poskytovaným sociální sítí, který však může být omezen licenčními podmínkami nebo počtem dotazů. Web scraping se sice umí vyhnout časovému omezení počtu dotazů, jeho soulad s licenčními podmínkami však může být zpochybněn. Vzhledem k povaze obou přístupů lze za jednoznačně vhodnější považovat využití API.

Na základě předchozích pozitivních zkušeností byl jako jazyk vhodný k využívání API různých sociálních sítí zvolen Python 2.7. Python je dynamický interpretovaný jazyk vhodný jak pro objektově orientované, tak procedurální programování (python.org, 2013). Autor za jeho výhody pokládá především strmou křivku učení, díky které lze být v jazyce rychle produktivní, a čistou syntaxi.

K vizualizaci dat pak byly zvoleny jak desktopový GIS software, tak jazyky vhodné k tvorbě interaktivních map, tedy především JavaScript (open source mapová knihovna Leaflet), HTML a CSS. Autor však využil také méně známé technologie, například programovací jazyk Processing sloužící k automatickému generování vizualizací, či JavaScriptovou platformu node.js vhodnou ke streamování dat v reálném čase.

Během práce nebyl využit žádný licencovaný či proprietární software. Celá práce vznikla na operačním systému elementaryOS založeném na linuxové distribuci Ubuntu, což lze vzhledem k licenčním podmínkám a ceně rovněž považovat za výhodu. Tím pádem však autor neměl možnost testovat v textu popsané postupy na operačním systému Windows.

Není-li uvedeno jinak, je zdrojem podkladových dat ve vizualizacích projekt OpenStreetMap.

Uložení prostorových dat

K uchování prostorových dat lze použít databázi nebo některý z výměnných formátů.[9] S ohledem na možnou potřebu využívat prostorové operace byla vybrána databáze PostgreSQL s rozšířením PostGIS. Pro menší objemy dat či jejich dočasné uložení byla zvolena databáze SQLite, respektive její prostorové rozšíření SpatiaLite. Pokud není uvedeno jinak, jsou surová data k dispozici spolu s výslednými vizualizacemi na přiloženém CD.

Formulace hypotéz

Lze z informací jednotlivců usuzovat na obecné principy prostorového chování? Vypovídají něco o individuálním způsobu života a postojích? Dají se z nich získat další, na první pohled třeba skryté informace?

[8] Automatický sběr dat v síti WWW. Více viz Hartley (2011).

[9] JSON, GeoJSON, XML.

Naším cílem není samoúčelné získání prostorových dat ze sociálních sítí. Ostatně že je tato úloha proveditelná, dokázaly již texty zmíněné v oddílu Východiska práce. Proto byly stanoveny následující hypotézy, o jejichž prokázání či vyvrácení se autor v práci pokusí.

1. Počet vazeb mezi uživateli klesá s rostoucí vzdáleností.
    - Platí také na sociálních sítích, že lidé častěji navazují kontakty s těmi, kteří jsou jim v prostoru blíže?
2. Uživatelé mohou být na základě svých příspěvků sledováni takřka v reálném čase.
    - Uvědomují si toto bezpečnostní riziko, nebo jej naopak nevnímají?
3. Sociální síť může sloužit jako zdroj informací o složení obyvatelstva na určitém území.
    - Lze analýzou jazykového nastavení uživatele vymezit území, na němž se například v České republice vyskytuje ruská menšina?
4. Data lze využít jako podpůrnou síť informací o jevech, které se dotýkají společnosti na určitém místě.
    - Dá se z dat získat informace o pohybu osob v městském prostředí? Mohla by být data využita jako zdroj informací pro sociální geografy? Lze data prakticky využít?

Cíle práce

Předchozí text dokazuje, že sociální sítě jsou jedinečným zdrojem prostorových informací, které jsou výjimečné především svým množstvím a autenticitou.

Testování hypotéz předchází výběr sítí vhodných k vytěžování dat na základě obecných postupů zmíněných v oddílu Metodika. S ohledem na sítí poskytované možnosti sběru dat jsou zvoleny nejvhodnější způsoby jejich automatizovaného sběru. Poslední fází tohoto řetězce je prezentace dat, při níž je kladen důraz především na návrh uživatelsky přitažlivých vizualizací, která dají dobře vyniknout zvláštnostem získaných dat.

Práce se rovněž zabývá právními aspekty souvisejícími se získáním dat, které jsou reprezentovány především podmínkami použití jednotlivých sítí. Ty by mohly v extrémních případech jakékoliv nakládání s uživatelskými daty zcela zakazovat, což je při výběru sítí nutno zohlednit.

Kapitola 1

Sociální sítě: vznik, vlastnosti, principy

Sociální sítě jsou bezesporu nejrychleji rostoucí oblastí internetu. Přitom existovaly dávno před příchodem internetu, to jen dnes můžeme nabývat mylného dojmu, že se jedná o horkou novinku posledních let. Chápeme-li sociální síť jako sadu spojení mezi lidmi, organizacemi či národy (Kadushin 2012, s. 3), je okamžitě jasné, že taková síť vzniká s narozením člověka či vznikem státu, a jako taková je tedy nezávislá na technologii umožňující propojení jejích členů - tím může být stejně tak kontakt tváří v tvář, jako e-mailová komunikace. Lidé jsou tedy členy sociálních sítí již po tisíciletí.

Internet pouze změnil povahu vztahů v těchto sítích. Jestliže byly velmi dlouho založeny především na příbuzenských a komunitních vztazích uvnitř prostorově blízkých skupin, s vynálezem telefonu docházelo k jejich postupnému oslabování na úkor vztahů vzdálenějších, a členové sítě tak nebyli limitováni vzdálenostmi. Internetové sociální sítě pak byly jen posledním článkem vývoje směřujícího od malých lokálních sítí k sítím globálním. Logicky se tedy nabízí otázka, jak (a zda vůbec) možnost internetové komunikace ovlivnila množství přímé komunikace mezi lidmi. Klesá se zvyšujícím se množstvím e-mailů či IM zpráv počet osobních kontaktů? Podle studie, kterou publikovali Boase a Horrigan (2006) a jíž se zúčastnilo přes dva tisíce respondentů, internet nejenže neoslabuje intenzitu osobních kontaktů, ba naopak přispívá k jejich dalšímu rozvoji. Autoři doslova uvádějí:

> Our evidence calls into question fears that social relationships — and community — are fading away in America. Instead of disappearing, people's communities are transforming: The traditional human orientation to neighborhood- and village-based groups is moving towards communities that are oriented around geographically dispersed social networks. [...] The internet and email play an important role in maintaining these dispersed social networks. Rather than conflicting with people's community ties, we find that the internet fits seamlessly with in-person and phone encounters. With the help of the internet, people are able to maintain active contact with sizable social networks, even though many of the people in those networks do not live nearby. (Boase a Horrigan, 2006)

Zájem o sociální sítě dokazuje zpráva společnosti Nielsen Holdings N. V. (2012),

která se zabývá situací ve Spojených státech amerických. Z naměřených dat vyplývají některá velmi zajímavá fakta:

- Uživatelů přistupujících na sociální sítě přes mobilní web či mobilní aplikaci mezi lety 2011 a 2012 přibylo o 82, resp. 85 %, a jejich počet dosahuje sta milionů.
- Uživatelé tráví používáním mobilních aplikací pro přístup k sociálním sítím o 76 % času více než v roce 2011.
- Třetinu času stráveného prohlížením internetu na mobilním telefonu lidé věnují sociálním sítím.
- Největší sociální sítí byl v roce 2012 Facebook s více než 1,5 miliardou uživatelů. Nejrychleji rostoucí sítí se stal Pinterest, když se počet registrovaných uživatelů ve sledovaném období zvýšil o 1 047 % na 272 milionů.
- Sociální sítě ovlivňují také vztah jednotlivců a firem - každý třetí člověk dává přednost kontaktu pomocí sociální sítě před telefonním hovorem.

Přestože fenomén sociálních sítí je předmětem vědeckého výzkumu již dlouho, internet přinesl nové aspekty, které dříve přítomné nebyly. Jedním z nich je *betweenness paradox* (Kadushin 2012, s. 8): uživatel má sice na první pohled nad svou sociální sítí úplnou kontrolu, to však pouze tehdy, kdy je síť funkční. V případě výpadku serveru můžeme mluvit o krátkodobém zániku uživatelovy sítě. V běžných mezilidských vztazích podobná situace nastat nemůže. Druhým důsledkem tohoto jevu je fakt, že odstraněním jednoho uzlu v síti může dojít k rozpojení na více částí. Typickým případem může být například zablokování uživatele, který tak ztratí možnost přistupovat k dalším spřáteleným uživatelům. I tato situace je v reálném světě velmi nepravděpodobná.

1.1 Jak vznikají sociální sítě

Vznik mezilidských kontaktů, potažmo sociálních sítí, má dvě hlavní příčiny (Kadushin 2012, s. 18):

- blízkost *(propinquity)* v prostoru, která hraje roli i na internetu. Čím blíže se lidé nalézají, tím větší je pravděpodobnost, že mezi nimi vznikne vazba. Pokusit se kvantifikovat a případně i kvalitativně charakterizovat vliv blízkosti na internetových sociálních sítí je jedním z cílů práce.
- podobnost *(homophily)*, s jejíž rostoucí mírou roste i pravděpodobnost vzniku spojení mezi lidmi. Můžeme rovněž konstatovat, že pokud mezi dvěma lidmi existuje vztah, lze předpokládat, že sdílejí společné vlastnosti, postoje či hodnoty. Podobnost dvou lidí může být buď výsledkem vrozených vlastností (věk, pohlaví), nebo získaných postojů. V anglické literatuře se setkáme s pojmy *status-homophily*, resp. *value-homophily*. Sociologové se však dodnes zcela neshodují, zda kontakt dvou lidí je důsledkem jejich společných zájmů, či zda společné zájmy jsou důsledkem kontaktu.

Obě zmíněné příčiny postupně vedou k tomu, že stejné typy lidí mezi sebou navazují kontakty, následně se začínají navzájem ovlivňovat a postupem času se mohou začít vyskytovat na stejných místech, která mají rovněž vliv na jejich chování a dále je k sobě přibližují.

Podle Simmela (Wolff, 1969) skutečná sociální síť vzniká přidáním třetího uzlu k již existující dvojici. Na adresu dyadických vztahů uvádí:

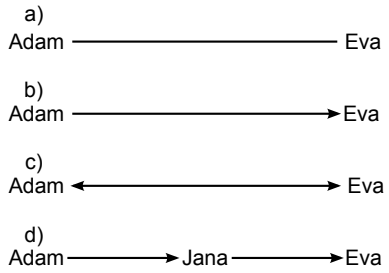
> ...the dyadic element is much more frequently confronted with All or Nothing than is the member of the larger group. ... No matter how close a triad may be, there is always the occasion on which two of the three members regard the third as an intruder. (Wolff 1969, s. 135)

Přidání třetího člena sítě rovněž rozšiřuje paletu možných vztahů, které vysvětluje následující oddíl.

Spolu s blízkostí a podobností hrají důležitou roli při vstupu člověka do sociální sítě také protichůdné psychické pohnutky. Na jedné straně je motivace cítit se bezpečně, čehož se dá dosáhnout začleněním do sítě, která jedinci poskytuje podporu a dodává mu právě pocit bezpečí. Proti ní působí snaha navazovat nové kontakty, která může pocit bezpečí oslabovat. Kromě těchto pohnutek pak také samotná síť vytváří podněty, které ovlivňují chování jejího člena. Jedná se především o potřebu vyrovnat se ostatním členům, která úzce souvisí s postavením jednotlivce v síti; lidé mají tendenci srovnávat se spíše s těmi, jejichž úroveň je pro ně samotné dosažitelná.[10] (Kadushin 2012, 56).

1.2 Typy vztahů mezi uzly sítě

Jak již bylo dříve uvedeno, sociální síť sestává z navzájem propojených uzlů. V předchozím textu byl rovněž vysvětlen způsob vzniku těchto spojení na úrovni jednotlivců. Zbývá tedy definovat, jaké vztahy mezi uzly sítě vznikají. Jejich přehled podává obr. 2. Kadushin (2012) jednotlivé typy charakterizuje takto:



Obr. 2: Vztahy mezi uzly sítě (upraveno podle Wolff, 1969).

a) nejjednodušší typ vztahu, který odpovídá situaci, kdy v místnosti stojí dva lidé. Pokud Adam stojí ve stejné místnosti jako Eva, potom Eva stojí ve stejné místnosti jako Adam. Vztah mezi uzly nemá směr *(not directional)*.

[10] Jinými slovy má větší smysl srovnávat svůj příjem se sousedem než s hollywoodskou filmovou hvězdou.

b) pokud má Adam rád Evu, neznamená to, že Eva má ráda Adama. Vztah mezi uzly má směr *(directional)*.

c) Adam má rád Evu a Eva má ráda Adama. Vztah se podobá a), ovšem v tomto případě je možné najít rozdíl v síle obou směrů. Jedná se o symetrický *(symmetric)* vztah, ačkoliv většina dyadických vztahů bývá asymetrická (nadřízený × podřízený, otec × syn).

d) Adam je s Evou v kontaktu prostřednictvím Jany. V případě triád hovoříme buď o vztazích tranzitivních (pokud má Adam rád Janu, Jana má ráda Evu), nebo netranzitivních. Tranzitivní vztahy se často vyskytují v oficiálních hierarchických sítích (Adam předá zprávu Janě, která ji předá Evě). Anglická literatura takovýto vztah označuje jako *relationship through an intermediary*.

Schematické vztahy na obr. 2 označujeme jako sociogramy; jedná se o jeden z nejpoužívanějších nástrojů v oblasti výzkumu sociálních sítí.

1.3 Typy sítí

Kadushin (2012, s. 17) hovoří o třech typech sociální sítě, které jsou předmětem sociologického výzkumu. Jsou to:

- egocentrická síť, která je tvořena z kontaktů okolo jednoho uzlu (např. síť přátel).
- sociocentrická síť, která je typická svou uzavřeností. Může se jednat například o žáky ve třídě nebo pracovníky firmy.
- otevřená síť, jejíž hranice jsou neznámé, a v níž se uplatňuje například princip *six degrees of separation*, popsaný v oddílu Východiska práce.

Na internetu můžeme najít jak sítě egocentrické, tak sítě sociocentrické, které jsou nedílnou součástí některých služeb (např. skupiny na Facebooku), ale také otevřené sítě, které mohou být reprezentovány kupříkladu členy fanouškovských stránek na Facebooku.

1.4 Objektivní charakteristiky sítí

Jak již bylo zmíněno, jedním ze silných sociologických výzkumných nástrojů jsou sociogramy. Jejich čitelnost je však omezena počtem uzlů a hran, které zobrazují. Pokud nejsou zobrazovány v interaktivním prostředí, stávají se při větším počtu objektů nepřehlednými, a neumožňují jednoduchou extrakci informací či zajímavých vzorů. Z toho důvodu nejsou příliš vhodné ani ke vzájemnému porovnání sítí. K tomu lze však využít charakteristiky, jejichž přehled je uveden níže.

1.4.1 Hustota sítě

Je definována jako počet všech existujících přímých spojení ku celkovému možnému počtu spojení v síti(Kadushin 2012, s. 29). Vysoká hustota je typická pro sítě s malým počtem uzlů, v nichž je jednodušší udržovat vztahy. Hustými sítěmi jsou například venkovské společnosti, jejichž členové mezi sebou většinou udržují více druhů vztahů (spolupracovníci, příbuzní) než jejich protějšci ve městech.

Hustota sítě rovněž pozitivně ovlivňuje přenos informací, stejně tak však může přispět k rychlému šíření chorob.

1.4.2 Síla vazby

S myšlenkou existence slabých vazeb *(weak ties)* a jejich důležitosti pro přenos informací přišel M. S. Granovetter (1973). Síla vazby je podle něj lineární kombinací trvání, emocionality a intimity daného vztahu. Na základě těchto ukazatelů lze vazby dělit na silné, slabé a neexistující, z nichž pouze slabé vazby mají reálnou možnost fungovat jako mosty pro přenos informací.

Jinými slovy řečeno, pokud člověk sdělí informaci lidem, se kterými ho pojí silné vazby a kteří učiní to samé, je velmi pravděpodobné, že každý z nich tuto informaci uslyší několikrát znovu. Naopak lidé, s nimiž je původce zprávy spojen slabou vazbou, mají potenciál tuto zprávu předat k těm, ke kterým by se jinak téměř jistě nedostala (Granovetter, 1973). Tento předpoklad ostatně empiricky dokazuje také již zmiňovaná Milgramova studie (1967).

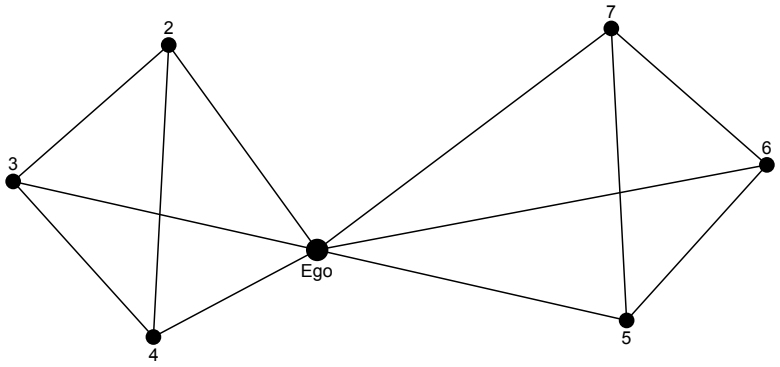
Pro sociální sítě složené ze slabých vazeb je typická nízká hustota spojení, čímž se liší od sítí, jejichž členové jsou spojeni silnými vazbami (Kadushin 2012, s. 30). Granovetter byl ve své době průkopníkem v oblasti výzkumu slabých vazeb, ani po čtyřiceti letech od publikování stěžejního článku tohoto vědeckého odvětví však neexistuje definice slabé vazby založená na objektivních faktorech.

1.4.3 Structural holes

Již zmíněný koncept hustoty vychází z předpokladu, že v síti existují vztahy, a popisuje jejich množství. Koncept *structural holes* však popisuje přesný opak - chybějící spojení. Burt (1992, s. 65) definuje *structural hole* jako neredundantní spojení mezi dvěma uzly. Toto spojení má výsadní postavení v rámci sítě, neboť přináší přidanou hodnotu, aniž by dublovalo již existující vztahy.

Na obr. 3 vidíme schéma jednoduché sítě, v níž jsou na první pohled patrné dva shluky: jeden tvoří uzly 2, 3 a 4, druhý potom uzly 5, 6 a 7. Členové obou shluků jsou mezi sebou vzájemně propojeni, avšak jediná cesta ke členovi jiného shluku vede přes uzel Ego, který v této situaci vystupuje jako *structural hole*. Jde o jediný uzel, který je klíčový pro propojení sítě, a jehož odebráním by se síť rozpadla na dvě podsítě. Význam *structural holes* se naplno projevuje při získávání nových informací. Optimalizované sítě se proto řídí dvěma základními principy (Burt, 1992, s. 67):

- princip výkonnosti spočívá ve snaze maximalizovat počet neredundantních spojení v síti. Zásadním krokem k výkonné síti je volba primárního kontaktu, jehož pomocí jsou realizovány vazby na další kontakty.



Obr. 3: Síť složená ze dvou shluků spojených prostřednictvím *structural hole*.

- princip účinnosti odpovídá snaze s co nejmenším úsilím spravovat co největší síť neredundantních kontaktů, které fungují jako brány k dalším kontaktům.

Burtův koncept navazuje na Granovetterovu teorii (1973), v některých aspektech se od ní však liší. Burt (1992, s. 73) především tvrdí, že příčinou přenosu informace není síla vazby, nýbrž trhlina, kterou *structural hole* reprezentuje. Síla vazby je podle něj pouze doprovodným jevem. Burt rovněž tvrdí, že přílišný důraz na sílu vazby překrývá další z významných schopností *structural hole*, kterou je kontrola nad děním v síti.

1.4.4 Popularita (centralita) uzlů

Ne všechny uzly v síti lze považovat z hlediska jejich vlivu za rovnocenné. Pro uzly v reciproké síti (např. Facebook: pokud jsi ty mým přítelem, jsem i já tvým přítelem) můžeme popularitu měřit počtem spojení, ve kterých daný uzel figuruje. V nereciprokých sítích tato spojení dělíme na příchozí *(indegree)* a odchozí *(outdegree)*.

Platí, že popularita uzlu nesouvisí jen s počtem spojení, ale také s jejich kvalitou. Spojení přicházející od populárních uzlů mají větší váhu než ta, která přicházejí od uzlů s malým množstvím příchozích spojení.

1.4.5 Vzdálenost mezi uzly

Hovoříme-li o vzdálenosti v sociální síti, máme na mysli nejkratší možnou cestu mezi dvěma uzly vedoucí buď přes další uzly, nebo přímo od uzlu k uzlu. Efektivita spojení nejkratší vzdáleností je v síti nejvyšší, zároveň však nijak neoslabuje roli delších spojení. Ta jsou pro fungování sítě rovněž důležitá, neboť právě redundantní spojení umožňují rychlý přenos informací mezi uzly. (Kadushin 2012, s. 35) V běžném životě se redundantní vztahy projevují v situacích, kdy jednu a tu samou informaci člověk obdrží z více různých zdrojů.

Pomocí vzdálenosti lze rovněž odhadovat pravdivost zpráv, která se k nám od uzlu dostává. Jako důvěryhodnou můžeme označit tu, kterou nám předal člověk, jehož známe osobně, méně důvěryhodné mohou být zprávy, které putují mezi několika lidmi, až se nakonec dostanou i k nám.

Uzly, které jsou přímo spojeny s daným uzlem, tvoří jeho tzv. *first-order zone*. Počet takových uzlů se může u jednotlivých lidí velmi lišit. Obecně platí, že čím větší počet přímo spojených uzlů, tím větší počet uzlů, které jsou vzdáleny o dva, tři a více kroků.

1.5 Milgramův „malý svět“

Přestože byly výsledky Milgramova experimentu (1967) již několikrát zmíněny, fenomén „malého světa“ je ve výzkumu sociálních sítí natolik stěžejní myšlenkou, že je vhodné se u něj znovu zastavit. Milgram ostatně nebyl první, kdo se tímto sociálním jevem zabýval, de Sola Pool a Kochen (1978 [1958]) jeho existenci empiricky ověřovali už na konci 50. let, své poznatky však uveřejnili až o dvacet let později, neboť jejich práce přinesla více otázek než odpovědí. Položili však základní kameny pro své následovníky v podobě dvou tezí: prokázali, že celkový počet kontaktů má vliv na počet kroků nutných ke spojení dvou lidí a že pokud jsou dva lidé členy stejného sociálního kruhu, nemohou se z něj vymanit a získat nové kontakty (de Sola Pool a Kochen, 1978 [1958]).

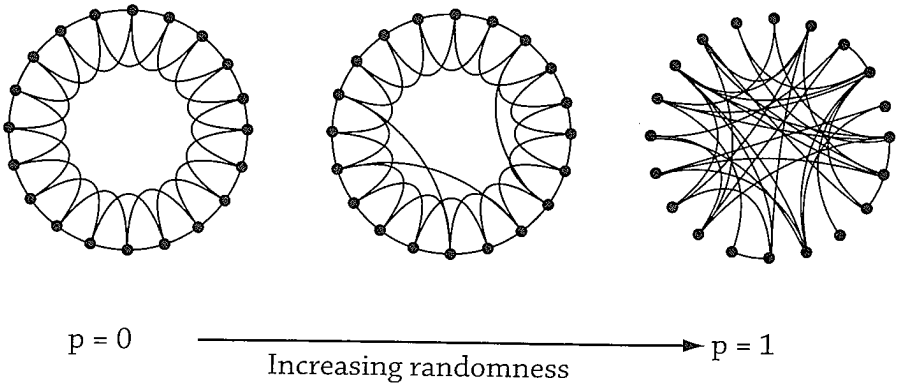
Milgram ze svého experimentu získal další empirická data a dále teorii „malého světa“ rozpracoval. Kadushin (2012, s. 108) dnes zdůvodňuje existenci „malého světa“ takto:

1. Lidé znají dostatečné množství jiných lidí, aby mezi jednotlivými okruhy přátel docházelo k průnikům. To potvrzují jak de Sola Pool a Kochen (1978 [1958]), tak Zheng, Salganik a Gelman (2006), z jejichž výzkumu vyplývá, že průměrný Američan zná asi 650 lidí, průměrná Američanka potom 590 lidí. Naměřená data vykazují vysokou variabilitu, což nás přivádí k dalšímu jevu, který má vliv na vznik „malého světa“.

2. Rozdělení množství kontaktů v populaci je silně zešikmené a platí, že velmi málo lidí má velmi mnoho kontaktů a naopak. Pro lidi s velkým množstvím kontaktů je snazší získávat další, čímž se tento jev ještě zvýrazňuje. Tito populární lidé (ve smyslu části 1.4.4) zajišťují soudržnost sítě a právě kontakty do jejích různých částí umožňují vznik „malého světa“. K tomu podle Kadushina (2012) dochází buď vzájemným spojováním populárních lidí, nebo naopak spojováním populárních lidí s těmi málo populárními. Dosud není známo, v jakých sítích se ten či onen model uplatňuje.

3. Množství kontaktů každého člena sítě je dostatečné na to, aby mohl kontaktovat jakéhokoliv dalšího člena sítě během konečného (a poměrně malého) počtu kroků, přestože je tento počet vždy vyšší než očekávaný na základě matematického výpočtu.

4. Větší než očekávaná vzdálenost mezi lidmi je způsobena jejich členstvím v sociálních kruzích *(social circle)*, které stojí za vznikem „malého světa“. Zatímco uvnitř kruhu jsou vzdálenosti kratší, mezi kruhy se vzdálenost zvětšuje a je obtížnější proniknout ke členovi jiného kruhu.



Obr. 4: Princip *rewiringu* (převzato Kadushin, 2012).

Kadushin (2012, s. 125) je charakterizuje jako hustší oblasti sítě, které nemají lídra ani jasné hranice. Jedná se tedy o neformální struktury. Vzdálenosti mezi jejich členy jsou kratší, přestože hustota sítě dramaticky neroste.[11]

5. Shluky vznikají kvůli společenské struktuře a na principu podobnosti (viz část 1.1).

6. Shluky se navzájem překrývají, čímž dochází k *rewiringu*, který je hlavní příčinou existence „malého světa“ a jehož princip popisuje obr. 4. Pokud by členové sítě byli spojeni pouze se svým nejbližším okolím (situace na obr. 4 vlevo), byl by jejich svět „obrovský“ a cesta ke členovi na protější straně sítě by byla velmi dlouhá.

   Rewiring reprezentuje stav, kdy existují kontakty i mezi vzdálenějšími členy sítě (obr. 4 vpravo), které umožňují podstatně zkrátit cestu mezi jakýmikoliv dalšími dvěma členy.[12]

7. Průniky shluků vytvářejí vertikální hierarchii.

Autor si mohl ověřit koncept „malého světa“ a *structural holes* na konci roku 2013, kdy v blízkosti svého bydliště nalezl fotoaparát. Prostřednictvím svého profilu na síti Twitter, kterou blíže popisuje oddíl 3.1, odeslal žádost o pomoc při hledání majitele s připojenou fotografií. Během necelých tří dnů tímto způsobem potenciálně oslovil více než 35 000 uživatelů této sítě a majitele se mu podařilo kontaktovat. Podrobnosti týkající se této události uvádí Příloha 11 na CD.

[11] Lidé se nejčastěji znají díky někomu dalšímu; jde tedy o přátele přátel.
[12] Vystupují v roli *structural hole*, viz část 1.4.3

Kapitola 2

Online sociální sítě

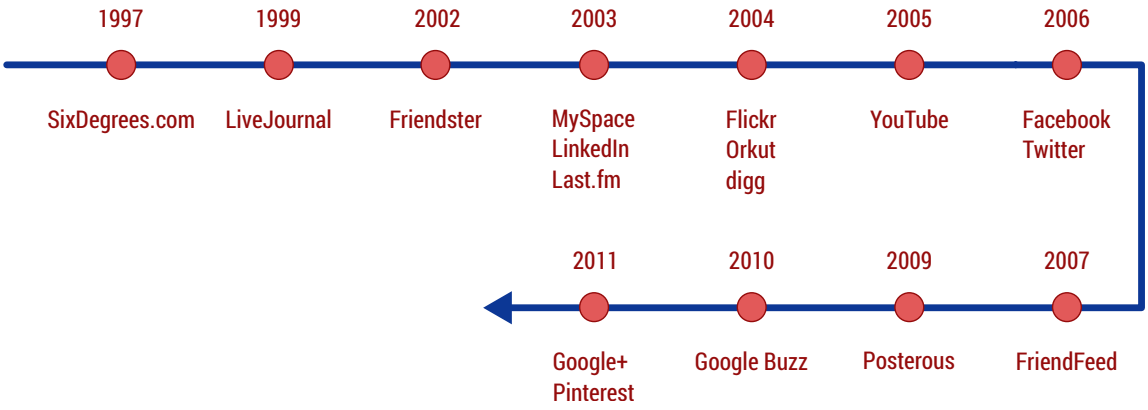
Boyd a Ellison (2007) definují online sociální sítě jako služby, které jsou založeny na třech jednoduchých principech:

- uživatel si v rámci systému nejdříve založí profil,
- poté se spojí s dalšími uživateli systému,
- nakonec si může prohlížet informace o ostatních uživatelích systému.

Opravdu razantní nástup sociálních sítí přichází po roce 2003 (viz obr. 5) a souvisí s příchodem technologií a postupů souhrnně označovaných jako Web 2.0. V tomto období vznikla většina dnes úspěšných sítí (např. MySpace, Last.FM, Orkut, později také Facebook a Twitter).

2.1 Web 2.0

Přestože Web 2.0 dodnes není zcela jednoznačně definován a bývá označován jako „buzzword“, znamenal významnou změnu v chápání webu a možnostech jeho využití. Jak uvádí O'Reilly (2005), lze jej popsat spíše pomocí pojmů než ucelenou definicí.



Obr. 5: Historie sociálních sítí (upraveno podle Ritholtz, 2010).

Stejně tak weby vytvořené na základě tohoto paradigmatu se od sebe liší, některé prvky však mají společné.

Zásadní změnou oproti minulé éře je role uživatelů webu při tvorbě obsahu. Zatímco v minulosti byli pasivními konzumenty textů a obrázků, dnes se mohou významnou měrou podílet na tom, jak web vypadá. Se změnou chápání webu souvisí vznik blogů nebo crowd-based projektů (např. Wikipedie). Web 2.0 nereprezentuje jen změnu v myšlení, ale také změnu technologickou, která uživatelům přináší přívětivější prostředí.

2.2 Historie sociálních sítí

Boyd a Ellison (2007) za první online sociální síť považují dnes již neexistující portál SixDegrees.com, který v roce 1997 skloubil do té doby separované funkce jiných služeb: umožnil uživatelům založit si svůj profil, spřátelit se s ostatními uživateli a prohlížet si jejich profily. Jedním z důvodů neúspěchu portálu byl nedostatek uživatelů; připojením k internetu sice lidé disponovali, mnoho jejich přátel však nikoliv, a tak nemohli smysluplně využít možnosti služby.

Další významnou sítí se stal o pět let později Friendster, původně navržený jako místo k seznámení pro přátele přátel. Jeho raketový vzestup však zastavily potíže související s nedostatečným hardwarovým vybavením a rovněž množství nově vznikajících falešných účtů, jejichž mazání bylo impulzem k odchodu mnoha uživatelů, kteří přestali sociální síti důvěřovat.

V roce 2003 vznikla síť MySpace, která těžila právě z odlivu uživatelů Friendsteru a rovněž ze zájmu o profily hudebních kapel, které začaly rychlým tempem přibývat. Za dva roky byla síť prodána společnosti News Corporation za 580 milionů dolarů, což poprvé vzbudilo větší zájem médií.

Tou dobou už mohli studenti Harvardu používat svoji vlastní privátní síť - Facebook. Nedlouho po nich dostali tuto možnost také studenti středních škol ve Spojených státech amerických, a v roce 2006 bylo omezení registrace na Facebooku zrušeno úplně. V témže roce dosáhl počet uživatelů 12 milionů, o tři roky později byl jejich počet třicetinásobný, za další rok se znovu téměř zdvojnásobil a v prosinci 2010 překročil hranici 608 milionů. V roce 2012 Facebook používala miliarda lidí. (newsroom.fb.com, 2013)

Dva roky po spuštění Facebooku spatřil světlo světa Twitter, sociální síť založená na sdílení tweetů, která jako jedna z mála využívá nereciprokých vztahů mezi uživateli. Podle Statistics Brain (2013) ji v současné době používá více než 550 milionů uživatelů, kteří denně odešlou téměř 60 milionů tweetů.

2.3 Uživatelé sociálních sítí

Podle závěrů šetření společnosti Pew Research Center (2013) provedeného na konci roku 2012 ve Spojených státech amerických na vzorku 1 802 respondentů lze konstatovat, že sociální sítě nejvíce využívají mladí dospělí, mezi nimi nejvíce ženy, Afroameričané a obyvatelé Latinské Ameriky.

Twitter používá 16 % dotázaných, nejvíce lidé žijící ve městech. Mitchell a Page (2013) dále uvádějí, že téměř polovina uživatelů Twitteru má 18-28 let, dosáhla vysokoškolského vzdělání a síť používá také na mobilním telefonu. Fotografie a videa na Instagramu sdílí 13 % respondentů. Tato síť je nejvíce populární mezi lidmi mladšími padesáti let, ženami, Afroameričany, Hispánci a městským obyvatelstvem.

Aktuální statistiky týkající se demografie uživatelů Foursquare se autorovi bohužel nalézt nepodařilo, výzkum publikovaný na serveru Alive Wired (2010) však ukazuje, že ženy síť používají více než muži, je velmi populární u uživatelů ve věku 13-17 a 35-49 let. Velké oblibě se síť těší mezi Asiaty a vysokoškolsky vzdělanými uživateli internetu.

Kapitola 3

Sociální sítě vhodné k získávání prostorových dat

Autor na základě svých zkušeností s užíváním sociálních sítí, jakož i studiem jejich podstaty a zaměření, identifikoval ty, které mohou být zdrojem zajímavých prostorových dat a informací. Jejich přehled přináší následující text. Zdaleka se nejedná o všechny využitelné sítě, v této práci se však budeme věnovat pouze jim.

3.1 Twitter

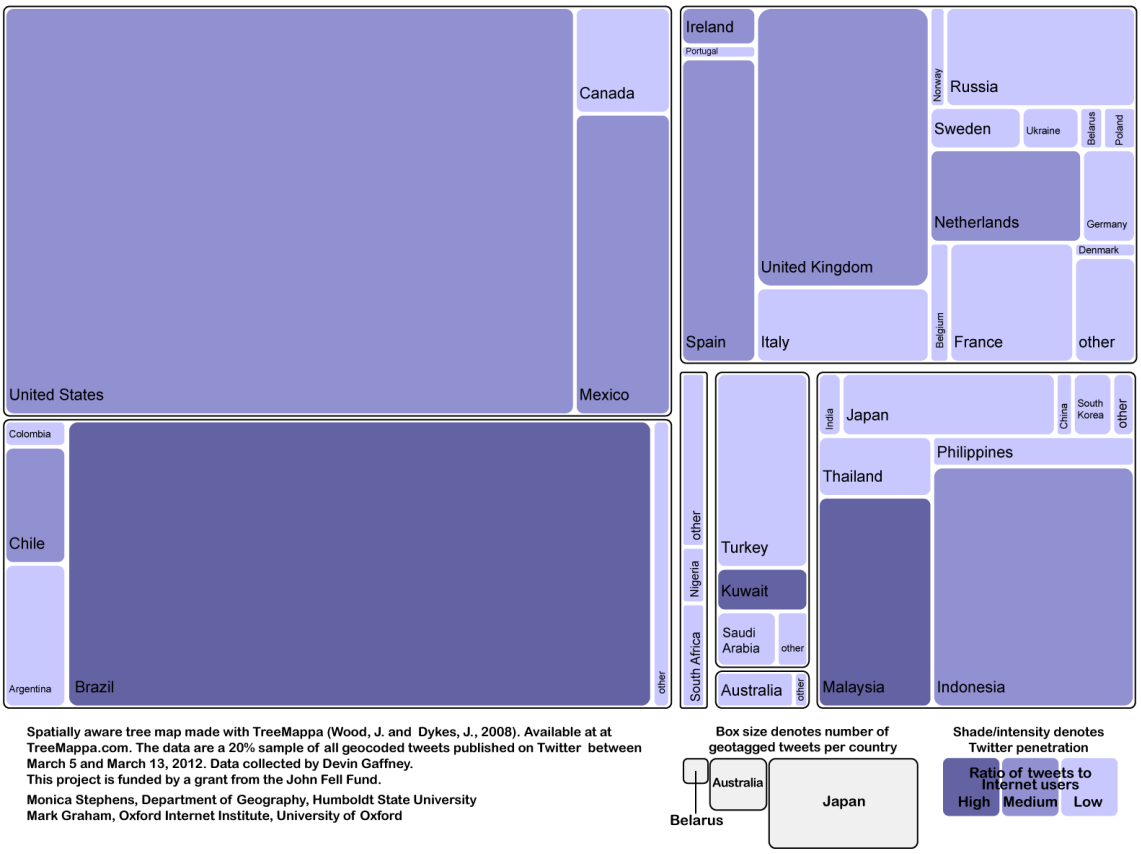
- Adresa služby: http://twitter.com
- Počet uživatelů ve světě: 555 000 000 (statisticbrain.com, 2013)
- Počet uživatelů v ČR: 190 000 (klaboseni.cz, 2014)

Vůbec první zpráva se na Twitteru objevila v březnu 2006 a jejím autorem byl Jack Dorsey, jeden ze zakladatelů sítě. Dva měsíce po spuštění bylo na Twitteru registrováno teprve 5 000 uživatelů (Carlson, 2011), instalace obrazovek ukazujících aktuální tweety z South By Southwest Interactive festivalu však vedla ke ztrojnásobení aktivity uživatelů.

Záhy se ukázalo, že největší síla Twitteru je především v rychlosti, jakou se na něm šíří informace. Ať už šlo o nouzové přistání letounu US Airways na řece Hudson, íránské volby v roce 2009, cestu roveru Curiosity k Marsu nebo prezidentskou kampaň Baracka Obamy, Twitter přinášel aktuální informace, ze kterých mnohdy vycházely zpravodajské servery (Griggs, 2013).

Velmi zajímavá je rovněž geografie Twitteru, kterou přibližuje obr. 6. Bylo by naprosto legitimní očekávat, že nejvíce uživatelů bude mít tato síť v technologicky vyspělých zemích, tedy především ve státech severní polokoule. Opak je ale pravdou, kromě Spojených států amerických a Velké Británie přední pozice ve výzkumu provedeném na Oxford Internet Institute (oii.ox.ac.uk, 2012) okupují Indonésie, Malajsie nebo Brazílie.

Síť je založena na pilířích, které byly zmíněny na začátku druhé kapitoly. Po registraci, vázané na e-mailovou adresu, si uživatel může začít budovat síť lidí, které



Obr. 6: Geografie Twitteru (převzato http://www.oii.ox.ac.uk/vis/?id=4fe09570, 2012).

chce sledovat. Domovská stránka portálu mu nabízí několik možností, jak s účtem pracovat:

- na záložce *Home* se zobrazují všechny tweety z účtů, které uživatel sleduje. Má možnost na ně odpovědět, přidat je mezi oblíbené nebo je retweetnout, tedy přeposlat všem uživatelům, kteří jej sledují.
- záložka *Connect* obsahuje dva seznamy:
  - *Mentions* obsahuje všechny tweety, v nichž někdo odkazuje na uživatelův účet, k čemuž na Twitteru slouží znak „@“.[13] Takto uvozené zprávy z účtů, které uživatel sám sleduje, se objeví rovněž na záložce *Home*. Ostatním uživatelům se zpráva zobrazí pouze tehdy, pokud sledují jak odesílatele, tak adresáta.
  - *Interactions* zobrazuje stejné položky jako předchozí seznam, a navíc rovněž informuje o tweetech, které si uživatelé přidali mezi oblíbené, a o nových uživatelích, kteří začali daný účet sledovat.

[13]Např. „@zimmicz jak se máš?“ se objeví v záložce *Mentions* autorova účtu.

- záložka *Discover* nabízí následující možnosti:
  - v seznamu *Tweets* se načítají zprávy, které Twitter vybírá na základě informací o uživateli
  - pod odkazem *Activity* lze najít informace o činnosti účtů, které uživatel sleduje
  - zbývající odkazy (*Who to follow*, *Find friends*, *Popular accounts*) navrhují uživateli účty, jejichž obsah by ho mohl zajímat
- pod položkou *Me* se na kartě *Tweets* nachází seznam posledních tweetů uživatele, *Following* obsahuje seznam lidí, které uživatel sleduje, *Followers* jsou naopak účty, které sledují uživatele, *Favorites* vypíše veškeré tweety, které uživatel označil jako oblíbené, karta *Requests* skrývá žádosti nových uživatelů (viz další text) a v nabídce *Lists* uživatel nalezne seznamy, které vytvořil a které lze používat například pro agregaci zpráv z podobně zaměřených účtů.

V nastavení účtu uživatel najde možnosti známé z dalších webových služeb (jméno, heslo, časové pásmo, zasílání novinek, jazykové nastavení), kromě nich se zde však nalézají také zajímavější položky. Jednou z nich je výběr země, v níž uživatel Twitter používá. Na základě tohoto nastavení mu pak mohou být zobrazovány reklamy v podobě *Promoted Tweets* či poskytován jiný obsah.

Uživatel svůj účet může chránit a zpřístupnit ho pouze vybraným lidem, jejichž žádost musí sám schválit (na zmiňované kartě *Requests*). Stejně tak může ke svým tweetům přidávat informaci o místě, ze kterého byly poslány.

3.1.1 Přístupná data

Od roku 2012 poskytuje Twitter uživatelům možnost stáhnout si veškerý obsah, který na server poslali (blog.twitter.com, 2012). Žádost o jeho vygenerování je po přihlášení k dispozici na stránce *Settings* → *Account*. Na registrovaný e-mail je následně zaslán komprimovaný archiv obsahující nástroj k zobrazení odeslaných tweetů po měsících, ale hlavně data samotná, a to ve formátech CSV a JSON.

Dalším způsobem jak přistupovat k datům publikovaným na Twitteru je využití veřejně dostupného API.

Jak vidíme, server nabízí přístup k individuálním i kolektivním datům prostřednictvím standardizovaných, strojově zpracovatelných formátů. Oba typy dat a jejich možné zpracování dále rozvádí následující část.

3.1.2 Osobní archiv uživatele

Jak již bylo uvedeno, data konkrétnímu uživateli Twitter nabízí ve formátech CSV a JSON. Jaký je mezi nimi rozdíl a jak je lze využít?

CSV

Kompletní sadu tweetů uživatel po stažení a rozbalení vygenerovaného archivu nalezne v souboru `tweets.csv`. CSV soubory lze dnes bez problémů importovat do

kancelářských tabulkových procesorů, k efektivnějšímu zpracování však může mnohem lépe posloužit systém řízení báze dat, v našem případě PostgreSQL 9.1.

Před importem dat do připravené tabulky (viz Příloha č. 1) je třeba nahradit zdvojené uvozovky apostrofy, což lze jednoduše učinit v jakémkoliv pokročilém textovém editoru (viz kód 3).

Kód 3: Úprava dat z archivu uživatele a import do databáze.

```
<a href=""http://seesmic.com/"" rel=""nofollow"">Seesmic</a>
<a href='http://seesmic.com/' rel='nofollow'>Seesmic</a>

COPY tweets FROM 'cesta_k/tweets.csv' WITH DELIMITER ',' CSV ESCAPE '\';
```

Nyní již můžeme využít nástroje PostgreSQL pro získávání informací z archivu. Autor, který účet na Twitteru využívá od roku 2007, zaslal do září roku 2013 celkem 6 258 tweetů.

Během práce s exportovaným CSV souborem autor narazil na dva významné problémy:

- K tweetům nejsou připojeny souřadnice. Ty jsou však uvedeny v JSON verzi exportu.
- U většiny tweetů je uveden nesprávný údaj o čase jejich vytvoření. Zatímco datum je v pořádku, u více než dvou třetin příspěvků je jako hodina vzniku uveden čas 00:00:00, což znemožňuje časovou analýzu těchto dat.

Prostorová analýza dat

Obsah archivu je specifický především tím, že jeho majitel může mnoho jevů očekávat. Autor práce tak může s jistotou tvrdit, že nejvíce tweetů bude souviset s Olomoucí a Brnem, tedy místy, kde se vyskytuje nejčastěji. Lze nejen tato místa identifikovat automaticky bez ohledu na jejich předchozí znalost?

Chybějící GPS souřadnice nás nutí využít postup popsaný v metodické části, který získává informace o místě z databáze GeoNames. K efektivnímu vyhledávání v obsahu tweetů však musíme poněkud pozměnit existující tabulku `tweets`. Kód 4 do ní přidá sloupec s datovým typem `tsvector` a naplní ho daty.[14] Hodnoty v tomto sloupci jsou reprezentovány jako abecedně seřazený seznam jedinečných lexémů (postgresql.org, 2013) a umožní nám fulltextově vyhledávat.

Kód 4: Přidání sloupce typu tsvector do tabulky tweets.

```
-- Do tabulky přidá sloupec s datovým typem tsvector
ALTER TABLE tweets ADD COLUMN textsearchable_index_col tsvector;
-- Sloupec naplní daty pomocí funkce to_tsvector()
UPDATE tweets SET textsearchable_index_col = to_tsvector('czech', text);
-- Nad sloupcem vytvoří index
CREATE INDEX tweets_textsearchable_index_col_gin_idx
    ON tweets USING GIN(textsearchable_index_col);
```

[14] V našem případě už nebudeme do tabulky vkládat další data, nemusíme proto implementovat trigger.

Jak již bylo zmíněno, místa v databázi GeoNames jsou rozdělena do kategorií, z nichž pro nás tou nejzajímavější jsou osídlená místa. Pomocí SQL dotazu uvedeného v kódu 5 vybereme tweety, v nichž se nachází některý z názvů obsažených v tabulce `geonames`, a uložíme je do tabulky `tweets_select`.

Kód 5: Výběr tweetů vhodných ke geokódování.

```
1 SELECT DISTINCT ON (tweets.text, geonames.name) tweets.text,
2        timestamp::timestamp, geonames.name, geonames.geom
3     INTO tweets_select
4     FROM tweets, geonames
5     WHERE textsearchable_index_col @@ plainto_tsquery(geonames.name)
6     -- Omezí hledání pouze na osídlená místa
7     AND geonames.feature_class = 'P'
8     ORDER BY geonames.name;
```

Zběžná kontrola výsledků dotazu ukázala, že do tabulky `tweets_select` bylo uloženo velké množství chybných záznamů. Tyto chyby můžeme rozdělit do několika kategorií:

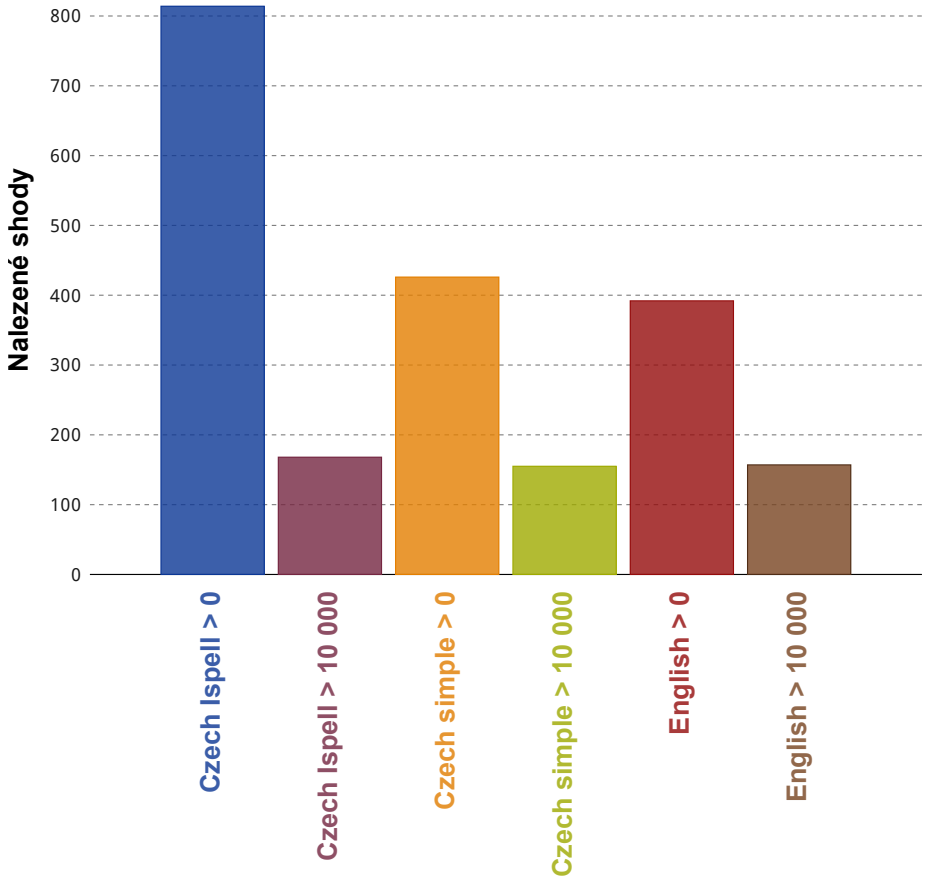
1. falešné shody pramenící z vlastností češtiny popsaných v metodické části práce,
2. falešné shody pramenící z obecnosti názvů (Dlouhé, Krásné),
3. falešné shody, které jsou výsledkem shodných názvů více míst v České republice (Vítkovice),
4. falešné shody, které jsou výsledkem shodných názvů míst v České republice a zahraničí (Mnichov).

Přesnost a kvalitu výsledků se autor pokusil zlepšit dvěma způsoby. Tím prvním je nastavení spodní hranice počtu obyvatel prohledávaných míst, která byla stanovena na 10 000. Obce s nižším počtem obyvatel nejsou brány v potaz, což z velké části eliminuje především místa s obecnými názvy (2. bod předchozího seznamu). Toho lze jednoduše dosáhnout přidáním restrikce `geonames.population > 10 000` k výše uvedenému dotazu.

Druhou cestou ke zpřesnění výsledků je změna jazykového nastavení vyhledávání, jehož možnosti jsou teoreticky popsány v metodice práce, prakticky je uvádí Příloha 3. Porovnání počtu výsledků při různých nastaveních ukazuje obr. 7. Čísla u popisu sloupců reprezentují minimální počet obyvatel sídla použitý při položení konkrétního SQL dotazu.

Jako nejpřesnější se jeví výsledky vyhledávání s použitím simple slovníku pro obce s více než 10 000 obyvateli (viz obr. 8). Je třeba zmínit, že i v nich se vyskytují falešné shody, které jsou však způsobeny nejednoznačností dat a jako takové je nelze rozpoznat, případně ošetřit v procesu geokódování. Srovnání rozdílů ve výsledcích přináší Příloha 4.

Četnost výskytu názvu ve výsledcích je na sérii map reprezentována intenzitou barvy. Z obr. 8 je patrné, že se potvrdila předpokládaná koncentrace tweetů v okolí Olomouce a Brna. Mapa vznikla v programu QGIS, který umožňuje načítat mapové



Obr. 7: Výsledky fulltextového hledání při různém jazykovém nastavení a nejmenším počtu obyvatel sídla.

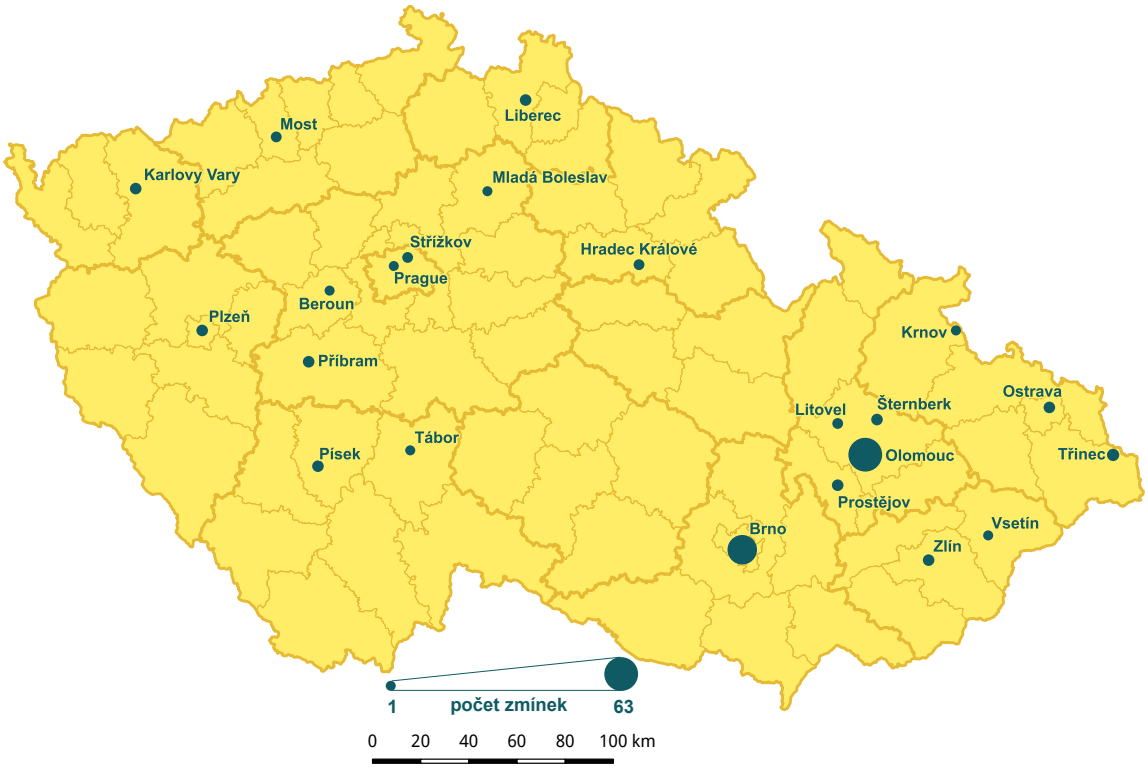
vrstvy z tabulek PostGIS databází. Jelikož byla intenzita barvy využita v prvé řadě k tvorbě animace (Příloha 5 na CD), musíme ji chápat pouze jako relativní ukazatel. Výsledná mapa vznikla postupným skládáním časových oken vrstvy sídel s definovanou průhledností. Jak dochází během animace k překládání časových intervalů přes sebe, postupně také roste intenzita barvy těch sídel, která jsou v datech zastoupena nejčastěji.

Velikost symbolů na mapě na obr. 8 reflektuje počet zmínek ve zdrojových datech. Umožňuje srovnání a je z ní ještě lépe patrná dominance Olomouce (63 zmínek) a Brna (47 zmínek).

Nedokonalost vyhledávacího mechanismu můžeme demonstrovat na sídlech Písek a Tábor, která jsou typickým příkladem falešných shod v důsledku obecnosti názvu. Jelikož vyhledávání nerozlišuje mezi velikostí písmen[15], nelze tomuto chování předejít. Přítomnost Mostu na mapě je rovněž výsledkem falešné identifikace slova ve zdrojových datech, kde se sice vyskytlo, nicméně v jiném kontextu (jednalo se o stejně psané anglické slovo most).

Další chyby ve výsledcích si čtenář může prohlédnout v již zmiňované animaci

[15]V prostředí sociálních sítí by toto chování bylo nežádoucí, neboť názvy jsou často uváděny s malými písmeny.



Obr. 8: Nalezené shody s použitím simple slovníku nad obcemi s více než 10 000 obyvateli.

v Příloze 5, která byla rovněž vytvořena v QGISu za použití pluginu TimeManager.

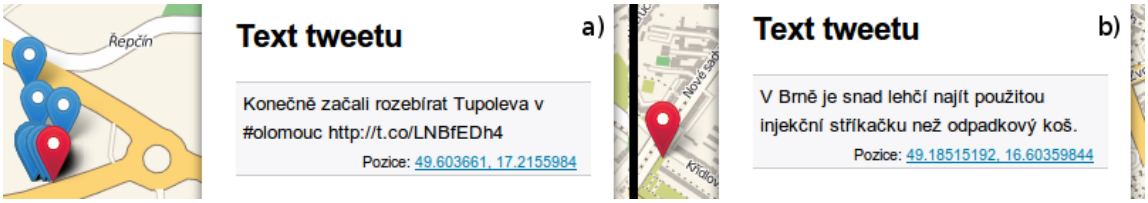
Z původních 6 258 tweetů bylo ve variantě s nejlepší shodou geokódováno pouze 155, tedy asi 2,5 % zdrojových dat, z nichž sedm bylo identifikováno jako falešné shody. Můžeme tedy tvrdit, že geokódování za daných podmínek vykázalo úspěšnost více než 95 %. Kromě již zmíněného Tábora (3×), Písku (2×) a Mostu byl špatně identifikován také Beroun, jehož zmínka v tweetu odkazovala k Moravskému Berounu, nikoliv k Berounu ve Středočeském kraji.

JSON

JSON je „odlehčený“[16] formát určený pro výměnu dat na internetu. Je jednoduše strojově zpracovatelný a dobře čitelný také pro člověka. Je nezávislý na jakémkoliv programovacím jazyku. Pro více informací o syntaxi formátu lze odkázat na server http://json.org/.

Jak již bylo uvedeno, tweety exportované ve formátu JSON obsahují rovněž prostorové souřadnice, které vizualizaci dat značně ulehčují. Po stažení osobního archivu v adresáři `data/js/tweets` nalezneme soubory s příponou .js, z nichž každý obsahuje vždy data za jeden měsíc daného období. Hledat v těchto souborech tweety s geotagy by samozřejmě byla zdlouhavá a neefektivní práce, kterou si však můžeme ulehčit jednoduchým skriptem napsaným v Pythonu (kód 6). Ten projde všechny

[16] Například v porovnání s XML.



Obr. 9: Zajímavé skutečnosti vyplývající z mapy: a) odchylka oproti poloze v čase odeslání; b) popis konkrétního místa. Vysvětlivky v následujícím textu.

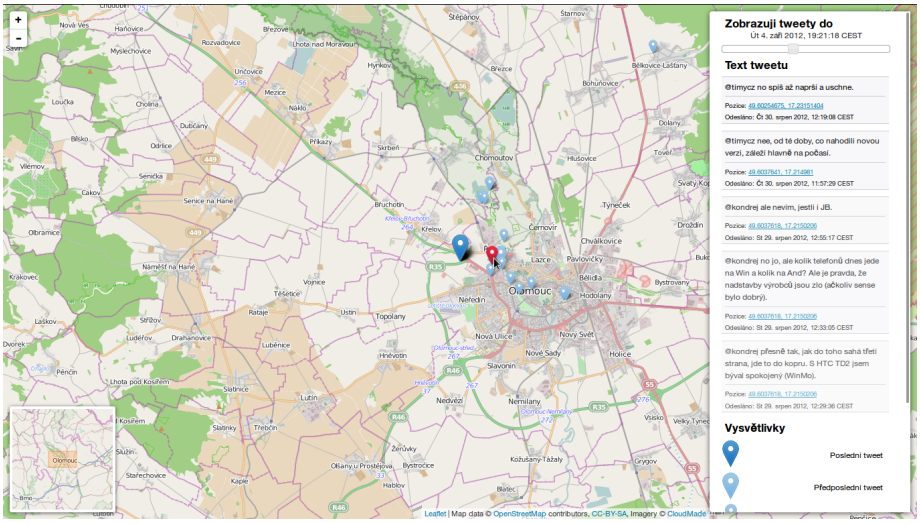
soubory v adresáři, prohledá jejich obsah, a pokud nalezne v JSON řetězci klíč `geo` reprezentující prostorové umístění tweetu, zapíše tento záznam do proměnné `tweets` výstupního souboru `tweets.js`.

Tímto způsobem tedy velmi jednoduše získáme všechny prostorově lokalizované texty a zbavíme se zbytečných záznamů.

Kód 6: Získání geotagovaných tweetů z osobního archivu uživatele.

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

import simplejson as json
import glob

data = []

# Využijeme názvy souborů a projdeme je od nejnovějšího k nejstaršímu
for file in sorted(glob.glob('*.js'), reverse=True):
        with open(file) as opener:
            # Vynecháme první řádek, který není validní částí JSONu
            lines = ''.join(opener.readlines()[1:])
            json_data = json.loads(lines)
            for d in json_data:
                if len(d['geo']) > 0:
                    data.append(json.dumps(d))
print 'Počet záznamů: ' + str(len(data))

if len(data) > 0:
    tweets = open('tweets.js', 'w')
    # Vytvoříme z JSON řetězce JavaScriptovou proměnnou tweets
    tweets.write('var tweets = [')
    tweets.write(','.join(data))
    tweets.write(']')
    tweets.close()
```

Z autorova osobního archivu bylo takto získáno 95 tweetů, které měly přiřazeny GPS souřadnice. K jejich vizualizaci byla zvolena JavaScriptová knihovna Leaflet (http://leafletjs.com) s doplňky L.Control.Sidebar (https://github.com/Turbo87/leaflet-sidebar), Leaflet-MiniMap (https://github.com/Norkart/Leaflet-MiniMap) a LeafletSlider (https://github.com/dwilhelm89/LeafletSlider). Jedná se o dynamicky se rozvíjející open source knihovnu určenou pro tvorbu interaktivních map,



Obr. 10: Náhled časové mapy tweetů (Příloha 6 na CD).

která není svázána s žádným poskytovatelem mapových podkladů.

Mapa je k dispozici v Příloze 6 na CD, náhled aplikace vidíme na obr. 10. Při její tvorbě byla pozornost věnována především časovému aspektu, který je u tohoto typu vizualizace velmi důležitý. Domníváme se, že zvolenou metodou se nám podařilo vytvořit mapu poskytující nejen informaci o prostorovém rozmístění tweetů, ale také o jejich časové souslednosti. Zobrazované texty byly z časového hlediska rozděleny do čtyř skupin:

- Poslední tweet je zobrazený sytě modrou barvou, aby čtenáře upoutal na první pohled.
- Předposlední tweet je zobrazený stejnou velikostí značky, ale světlejší barvou. Má sloužit pro rychlou informaci o tom, zda se autorova poloha od odeslání posledního tweetu výrazně změnila.
- Tweety odeslané před méně než týdnem (vztaženo k času odeslání posledního tweetu zobrazeného na mapě) jsou reprezentovány menšími značkami ve stejné barvě jako předposlední tweet.
- Tweety odeslané před více než týdnem (vztaženo k času odeslání posledního tweetu zobrazeného na mapě) představují nejmenší značky ve stejné barvě jako předposlední tweet.

Hranice jednoho týdne byla zvolena arbitrárně a v případě zobrazení jiných dat je možné ji změnit. Mapa je interaktivní, po kliknutí na některou ze značek je aktualizován výpis souvisejících tweetů v pravém panelu. Pro lepší orientaci byla do mapové kompozice přidána ještě přehledová mapa.

Z 6 258 tweetů bylo GPS souřadnicemi opatřeno pouze 95, jejichž rozmístění ještě lépe dokumentuje místa, kde se autor vyskytuje nejčastěji. Ze zobrazených informací stojí za zmínku především dvě skutečnosti.

- Většina tweetů odeslaných z autorova bydliště je oproti skutečnému umístění posunuta asi o jeden kilometr (obr. 9a). Jedná se o shluk bodů v severozápadní části Olomouce. Jejich přesnost byla pravděpodobně ovlivněna polohou mobilního telefonu v době odeslání; většina těchto zpráv totiž byla odesílána z místnosti, což samozřejmě negativně poznamenalo příjem signálu ze satelitů.

- V okolí Poříčí a brněnského hlavního nádraží autor třikrát odeslal tweet týkající se zanesení daného místa odpadky (obr. 9b). Tato vyjádření jistě nemůžeme považovat za reprezentativní popis místa, bezesporu však signalizují možné budoucí využití Twitteru například pro oznamování problémů ve městech. Co by se stalo, kdyby takové tweety nebyly tři, ale byly jich stovky od desítek uživatelů? Zvýšila by se jejich důvěryhodnost? Bylo by možné problémy v dané lokalitě přehlížet?

Jak vyplývá z předchozího textu, i určení polohy pomocí GPS souřadnic může vykazovat chyby. Ty jsou způsobeny nepřesným zaměřením zařízení v době odeslání tweetu. Odchylka je však nesrovnatelná s chybami, které jsme pozorovali při geokódování textových řetězců.

Shrnutí

V předchozím textu jsme zevrubně prozkoumali data, která má každý registrovaný uživatel na požádání k dispozici. Poukázali jsme na rozdíly mezi výstupními formáty, navrhli možnosti extrakce prostorových dat a způsob jejich vizualizace a identifikovali problémy, které s jejich zpracováním souvisí. Znalosti získané o dostupných datech využijeme v následující části práce, která popisuje API pro přístup k veřejně dostupnému obsahu ze serverů služby.

V souvislosti s možností stáhnout si veškerá data odeslaná na Twitter vyvstává otázka, jakým způsobem jsou tato data chráněna před zneužitím třetí stranou a jaký je vůbec postoj Twitteru k jejich správě. Možnost dohledat data od doby, kdy byl účet aktivován, může pro některé uživatele představovat potenciální bezpečnostní riziko. Jaký je tedy právní pohled Twitteru na vztah s uživatelem?

Základním dokumentem upravujícím tento vztah jsou podmínky užití služby (twitter.com/tos, 2012). V části 2 tohoto dokumentu, potažmo v dokumentu o ochraně osobních údajů (twitter.com/privacy, 2013), stojí:

> When using any of our Services you consent to the collection, transfer, manipulation, storage, disclosure and other uses of your information as described in this Privacy Policy. Irrespective of which country you reside in or supply information from, you authorize Twitter to use your information in the United States and any other country where Twitter operates. (twitter.com/privacy)

Jinými slovy tedy uživatel souhlasí s jakýmkoliv užitím obsahu, který na server odešle. V části 5 pojednávající o právech uživatele následuje:

> You retain your rights to any Content you submit, post or display on or through the Services. By submitting, posting or displaying Content

> on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). (twitter.com/tos)

Uživatel tak sice odesláním obsahu nepřichází o svá práva (např. autorské právo), zároveň však dává Twitteru svůj souhlas s jakýmkoliv dalším využitím informací. Založení uživatelského účtu je podmíněno souhlasem s těmito podmínkami, a je tedy na posouzení každého uživatele, na kolik například výše zmíněné části vnímá jako rizikové.

3.1.3 API: cesta k veřejně dostupnému obsahu

V předchozí části jsme pracovali s daty jednoho konkrétního uživatele. Jejich získání bylo podmíněno zasláním žádosti na server, který jako výsledek vrátil komprimovaný archiv. Takový postup je při větším počtu sledovaných uživatelů nepoužitelný, existuje však velmi jednoduchý způsob, jak z Twitteru získat data týkající se různých uživatelů, událostí nebo míst. Celá paleta metod pro přístup k těmto datům se skrývá pod zkratkou API.

Twitter nyní poskytuje tři různá API, na která se podíváme podrobněji. Je na místě rovněž zmínit, že metoda web scrapingu, popsaná v metodické části práce, není v souladu s podmínkami použití služby. Jediným způsobem pro přístup k obsahu je právě API, což Twitter explicitně uvádí v podmínkách použití, konkrétně v kapitole 8 *Restrictions on Content and Use of the Services*:

> (iii) access or search or attempt to access or search the Services by any means (automated or otherwise) other than through our currently available, published interfaces that are provided by Twitter (and only pursuant to those terms and conditions), unless you have been specifically allowed to do so in a separate agreement with Twitter (NOTE: crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, scraping the Services without the prior consent of Twitter is expressly prohibited); (twitter.com/tos)

REST API v1.1

REST je styl architektury používaný v distribuovaných systémech, který poprvé navrhnul Fielding (2000). Jeho definice přesahuje rozsah práce, pro detailní popis architektury proto odkazujeme na zmiňovaný text (především kapitolu 5). Pro účely této práce vystačíme s konstatováním, že RESTful API Twitteru nám umožňuje přistupovat ke zdrojům, respektive jejich reprezentacím ve formátu JSON, definovaným v dokumentaci (dev.twitter.com, 2013). Můžeme tak činit pomocí HTTP metod GET a POST, z nichž první jmenovaná slouží k získání informací ze zdroje, zatímco druhá do něj uloží odeslaný požadavek (RFC 2616, 1999).

Patrně nejvýraznější novinkou představenou ve verzi 1.1 je povinnost autentizovat jakýkoliv dotaz na server prostřednictvím protokolu OAuth (RFC 6749, 2012); již

tedy není možné zasílat anonymní požadavky. Pokud chceme přistupovat k veřejně dostupným datům, je nezbytné si na Twitteru založit uživatelský účet, k němuž si následně můžeme vygenerovat tokeny potřebné pro spojení se serverem. Více informací o využívání OAuth autentizace uvádí dokumentace Twitter API (dev.twitter.com, 2013).

Povinná autentizace serverových dotazů nebyla jedinou novinkou spojenou s nasazením API v1.1. Další důležitou změnou bylo zmenšení časového okna, z něhož se počítá limit dotazů. Zatímco v předchozí verzi mohl uživatel s jedním tokenem za 60 minut poslat 350 GET dotazů, nyní může za 15 minut poslat 15, respektive 180 dotazů, a to v závislosti na jejich povaze. Přehled omezení opět uvádí dokumentace (dev.twitter.com, 2013). Důležité je si uvědomit, že každý typ dotazu má nyní svůj nezávislý „rozpočet“: pokud například vyčerpáme všechny dotazy na zdroj `account/settings`, neovlivní to nijak počet dotazů, které můžeme zaslat na ostatní zdroje.

Streaming API

Z hlediska přístupu ke zdrojům je také Streaming API RESTful, od výše zmíněného API je však v dalších ohledech velmi odlišné. Zmiňme především dva nejdůležitější aspekty:

1. Streaming API je určeno pro získávání tweetů od okamžiku spojení dále do budoucnosti, tzn. že jeho prostřednictvím nemůžeme získat tweety odeslané dříve, než bylo spojení navázáno.

2. Streaming API vyžaduje neustále otevřené spojení na server. Zatímco REST API přijme HTTP požadavek, zpracuje ho a zašle odpověď, ke Streaming API se prostřednictvím klienta připojíme a zachytáváme tok dat.

Twitter samozřejmě prostřednictvím Streaming API neposkytuje kompletní seznam odeslaných tweetů, ale pouze jejich vzorek. Ten je, bez ohledu na token použitý ke spojení, pro všechny přistupující klienty vždy stejný. Toto API nabízí tři různé streamy, z nichž pro nás nejzajímavější je Public stream, který nabízí filtrovací nástroje (dev.twitter.com, 2013). Ty mohou sloužit k omezení streamovaných tweetů na základě jejich umístění či slov, která obsahují. Pokud výsledky dotazu tvoří více než 1 % provozu na serveru, server zašle zprávu informující o počtu tweetů, které překročily tuto hranici a nebyly streamem doručeny. Přístup ke Streaming API je rovněž podmíněn OAuth autentizací.

Search API

Na rozdíl od Streaming API umožňuje Search API přístup k tweetům odeslaným před zasláním požadavku na server. Dle dokumentace (dev.twitter.com, 2013) však klade důraz spíše na relevanci výsledků než na jejich úplnost, a může se tedy stát, že v odpovědi na požadavek budou některé tweety chybět. Rovněž toto API vyžaduje OAuth autentizaci a umožňuje zaslat maximálně 180 dotazů během 15 minut.

Pro pohodlnější práci s API vzniklo velké množství knihoven pro různé programovací jazyky. V této práci budeme využívat knihovnu tweepy (pythonhosted.org, 2013) určenou pro Python, respektive její fork dostupný na https://github.com/nirg/tweepy, který kromě metod určených pro přístup k API implementuje rovněž monitorování limitů časových oken.

3.1.4 Sledování uživatelů

Jedna z hypotéz formulovaných na začátku práce tvrdí, že prostřednictvím sociálních sítí je možné jejich uživatele sledovat téměř v reálném čase. V následujícím textu se pokusíme tuto hypotézu potvrdit.

Vycházejme z předpokladu, že sledovat uživatele je možné tehdy, když o sobě dává nějakým způsobem vědět - v našem případě tedy odesílá tweety, které obsahují informaci o místě odeslání ve formě GPS souřadnic. Jak takového uživatele na Twitteru nalézt? API takovou možnost nenabízí, a tak nám zbývají v zásadě dvě možnosti:

1. najít uživatele manuálně, tedy metodou pokus–omyl,
2. využít Search API či Streaming API a hledat geotagované tweety, následně prohledat další tweety jejich autorů a zjistit, zda informace o poloze odesílají pravidelně.

Autor v tomto případě zvolil první, přímočaré řešení, a možnosti sledování uživatelů se rozhodl demonstrovat na účtu Romana Kreuzigera, českého cyklisty působícího v současné době ve stáji Saxo–Tinkoff. Kreuziger je na Twitteru poměrně aktivní, pravidelně k tweetům připojuje GPS souřadnice a vzhledem ke svému povolání často mění místa pobytu.

Prvním krokem k ověření hypotézy je získání tweetů uživatele. Kód napsaný v Pythonu, který zajistí stažení a uložení tweetů ve formátu JSON, uvádí Příloha 7. Některé zajímavé části si představíme podrobněji. Při založení aplikace na Twitteru uživatel obdrží čtyři alfanumerické řetězce sloužící pro ověření práv pomocí protokolu OAuth. Po úspěšném ověření přístupu máme v proměnné `api` uložen objekt tweepy.API (viz kód 7), který implementuje všechny metody API popsané v dokumentaci (dev.twitter.com, 2013).

Kód 7: Ověření aplikace pomocí knihovny Tweepy.

```
consumer_key = 'xxxxxxxxxxxxxxxxxxxx'
consumer_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth, monitor_rate_limit=True, wait_on_rate_limit=True)
```

Nyní tedy pošleme požadavek na server a jako odpověď budeme požadovat tweety uživatele s identifikátorem *Roman86_K* (viz kód 8). Projdeme všechny stránky odpovědi a tweety, které obsahují informaci o poloze, si uložíme do pole spolu s časem jejich odeslání v sekundách. V Příloze 7 je k dispozici kompletní kód, který se následně postará o uložení získaných tweetů do souboru v chronologickém pořadí.

Z účtu Romana Kreuzigera bylo získáno 135 tweetů, z nichž nejstarší pochází ze 4. ledna 2013 a poslední z 27. prosince téhož roku. Výsledek běhu skriptu máme uložen v souboru `tweets.json` a můžeme přikročit k jeho vizualizaci.

Velký časový i prostorový rozsah dat nás nutí zohlednit oba tyto rozměry ve výsledné vizualizaci. Za tímto účelem jsme se rozhodli použít časoprostorovou kostku jako vhodný způsob zobrazení dat.

Kód 8: Získání geotagovaných tweetů z REST API.

```
for page in tweepy.Cursor(api.user_timeline, id="Roman86_K").pages():
    for tweet in page:
        if tweet.geo is not None:
            ...
            t = json.loads(tweet.json)
            t['timestamp'] = string_to_timestamp(t['created_at'])
            tweets.append(json.dumps(t))
```

Časoprostorová kostka je dobře popsaný a používaný způsob vizualizace (Kraak, 1995; Andrienko, Andrienko a Gatalsky, 2003; Li a Kraak, 2005) časoprostorových dat. Osy X a Y reprezentují prostorové rozložení jevu, osa Z potom jeho vývoj v čase.

Pro konstrukci kostky byl použit programovací jazyk Processing (processing.org, 2013), respektive jeho port do jazyka Python - processing.py (github.com, 2013). Jedná se o open source programovací jazyk a vývojové prostředí určené pro tvorbu 2D a 3D vizualizací, které bylo původně postaveno nad jazykem Java a následně přepsáno do dalších programovacích jazyků. Spuštění vizualizace provedeme voláním dávkového souboru z příkazového řádku, kdy jako parametr předáme jméno souboru, v němž je uložen skript.

Interaktivní vizualizace se v prostředí Processing definují dvěma funkcemi:

- `setup()` proběhne pouze jednou při spuštění programu a slouží k inicializaci vizualizace,
- `draw()` běží ve smyčce (pokud programu neřekneme jinak) a obstarává animaci.

Kompletní kód použitý k tvorbě kostky uvádí Příloha 8, některé jeho části si opět rozebereme podrobněji. Při startu programu načteme data (viz kód 9) z připraveného souboru `tweets.json` a využijeme připravených časových otisků, které jsme vytvořili při ukládání tweetů (viz Příloha 7).

Prostorovou část dat reprezentuje podkladová mapa zobrazená v podstavě kostky. Rozměry podstavy v pixelech tedy musí odpovídat rozměrům podkladové mapy. Zároveň je nutné znát GPS souřadnice minimálního ohraničujícího obdélníku zobrazeného území, které potřebujeme k namapování zeměpisných souřadnic do souřadnic vizualizace. K převodu souřadnic mezi systémy můžeme využít funkci `map()`,

které jako parametry předáme mapovanou hodnotu, rozsah původního souřadného systému a rozsah cílového souřadného systému (viz inicializace proměnných `x` a `y` v kódu 10). Na osu X vynášíme zeměpisnou délku, na osu Y zeměpisnou šířku.

Kód 9: Načtení dat k vizualizaci v časoprostorové kostce.

```
1 data = loadJSONArray('./tweets.json')
2 last = data.getJSONObject(data.size()-1).getFloat('timestamp')
3 first = data.getJSONObject(0).getFloat('timestamp')
```

Časový aspekt dat vynášíme na osu Z a proměnnou `z` inicializujeme podobně jako předchozí proměnné, místo zeměpisných souřadnic však mapujeme časový otisk. Takto si připravíme pole asociativních polí s klíči `x`, `y`, `z` (kód 10), které následně vykreslíme v kostce.

Vykreslení bodů a linií provádíme v metodě `draw()` pomocí funkcí `point()` a `line()`, kterým jako parametry předáváme namapované souřadnice. Počátek souřadného systému v prostředí Processing leží v levém horním rohu, proto musíme souřadnice na ose Y odečítat od výšky kostky, abychom body v prostoru umístili správně.

Kód 10: Vytvoření bodů se souřadnicemi `x`, `y`, `z`.

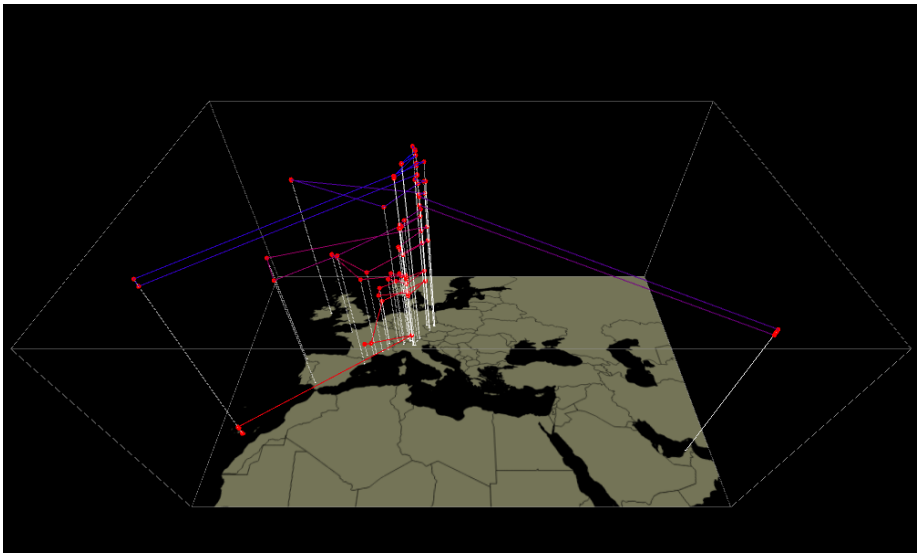
```
1  for i in range(0, count):
2      lon = data.getJSONObject(i).getJSONObject('coordinates')
3          .getJSONArray('coordinates').getFloat(0)
4      lat = data.getJSONObject(i).getJSONObject('coordinates')
5          .getJSONArray('coordinates').getFloat(1)
6      time = data.getJSONObject(i).getFloat('timestamp')
7
8      x = map(lon, -19.68624620368202116, 58.92453879754536672, 0, width)
9      y = map(lat, 16.59971950210866964, 63.68835804244784526, 0, height)
10     z = map(time, first, last, 0, 500)
```

Body jsou spojeny linií znázorňující časový průběh, která mění barvu od červené (nejstarší tweety) po modrou (nejnovější tweety). Ke každému bodu je navíc v rovině podstavy zobrazen jeho kolmý průmět, který je s původním bodem spojen bílou linií (viz obr. 11). V prostředí Processing je možné kostkou libovolně rotovat, případně měnit přiblížení kamery. Video vytvořené rotací kostky je k dispozici v Příloze 9 na CD.

V časoprostorové kostce můžeme jednoduše identifikovat místa, kde se osoba vyskytuje často - na těchto místech se zobrazuje více červených bodů nad sebou. V případě Kreuzigera se v České republice jedná o Plzeň a Prahu, v Itálii potom o okolí Lago di Garda, kde jsou pro cyklisty vhodné tréninkové podmínky. Dále z vizualizace vidíme, že Kreuziger část zimního období pravidelně tráví na Kanárských ostrovech, což zřejmě opět souvisí s jeho profesí. Můžeme tedy lehce rozpoznat *návyky*, které se odrážejí v prostorovém chování.

Tento způsob vizualizace nám umožňuje dobře zachytit relativní[17] vývoj jevu v čase, nejsme však schopni určit jeho absolutní polohu na ose Z, tedy přesný čas

[17]Umíme rozlišit, jaká je souslednost zobrazených dat.



Obr. 11: Časoprostorová kostka v prostředí Processing.

vzniku. Zároveň nám zakrývá samotný obsah jevu: kromě vývoje z vizualizace nejsme schopni zjistit, co se za jednotlivými body skrývá.

Shrnutí

Dokázali jsme, že sledovat konkrétní osobu prostřednictvím obsahu odesílaného na Twitter je poměrně jednoduché, a to v souladu s podmínkami použití služby. Předpokladem k úspěšné vizualizaci dat je však možnost získat zeměpisné souřadnice určující místo, ze kterého byl tweet odeslán. Původní hypotézu tedy můžeme, při splnění zmíněných podmínek, označit za platnou. Rovněž jsme zkonstruovali časoprostorovou kostku jako zajímavý způsob zobrazení tohoto typu dat.

Zároveň je zřejmé, že podobnou vizualizaci můžeme využít nejen při sledování pohybu jednoho uživatele na více místech, ale také při sledování pohybu více uživatelů na jednom místě.

3.1.5 Menšiny v České republice

V úvodu práce jsme nastínili možnost využít sociální sítě pro identifikaci různých etnik na vybraném území. Na základě znalostí Twitter API popsaných v části 3.1.3 jsme se rozhodli za tímto účelem využít Streaming API, z něhož můžeme v reálném čase získávat tweety pro požadované území. V tuto chvíli se pro nás výchozím bodem stává zdroj `statuses/filter`, kterému je možné předat parametr `locations` definující obdélník prostorově omezující tweety odesílané serverem.

Streaming API na rozdíl od REST API vyžaduje neustále otevřené spojení se serverem, což ovlivňuje také volbu nástrojů použitých k získání dat. Knihovnu tweepy[18] určenou pro Python jsme proto vyměnili za node.js (nodejs.org, 2013), což je platforma postavená na JavaScriptovém enginu V8, který používá prohlížeč Google

[18]Ta rovněž podporuje Streaming API, pro náš cíl se ale více hodí jiné nástroje.

Chrome. Ve spojení s knihovnou socket.io (https://socket.io) je node.js vhodnou volbou pro streamování dat.

Pro připojení ke Streaming API jsme využili knihovnu ntwitter (https://github.com/AvianFlu/ntwitter). Celý kód je k dispozici v Příloze 10 na CD. Skript lze spustit příkazem `nodejs app.js`[19]. Ve složce `public` je k dispozici také statický soubor `index.html` obsahující konečnou vizualizaci, který k běhu nepotřebuje node.js. Aplikace spuštěná prostřednictvím node.js běží na adrese http://localhost:3000.

Po spuštění začne program stahovat tweety, které spadají do definovaného obdélníku (viz proměnná `cz` v následujícím kódu), a ukládá je do SQLite databáze (kód 11). Kromě číselného identifikátoru tweetu, jejich autora a souřadnic nás především zajímá vlastnost `lang`, která obsahuje dvojpísmenný kód jazyka tweetu dle standardu ISO 639-1. Hodnotu vlastnosti stanovují algoritmy pro automatickou detekci jazyka (dev.twitter.com, 2013), které však nejsou stoprocentně spolehlivé.

Kód 11: Získání a uložení dat z Twitter Streaming API.

```
1  var cz = '12.09,48.55,18.87,51.06';
2
3  t.stream('statuses/filter', { 'locations': cz }, function(stream) {
4
5  // Jsme připojeni k serveru a zachytáváme proud dat
6   stream.on('data', function(tweet) {
7    if (tweet.coordinates) {
8     var lang = tweet.user.lang ? tweet.user.lang : 'und';
9     db.serialize(function() {
10     db.run("INSERT INTO " + tablename + " VALUES(?,?,?,?,?,?,?,?)", tweet.id,
11     tweet.user.name, tweet.user.id, tweet.text, tweet.created_at,
12     tweet.coordinates.coordinates[0], tweet.coordinates.coordinates[1],
13     tweet.lang);
14    })
15   }
16  });
17 });
```

Vlastnost `lang` nalezneme rovněž v entitě reprezentující uživatele, která je také součástí odpovědí API na zasílané požadavky; zde reprezentuje jazykové nastavení účtu uživatele na serveru. Pokud se hodnoty obou vlastností shodují, můžeme z nich s velkou pravděpodobností usuzovat na národnost uživatele. Rozdílné hodnoty většinou obdržíme v důsledku následujících událostí:

- Serverový algoritmus nesprávně určí jazyk tweetu. To je běžné především u krátkých tweetů.
- Jazyk tweetu se liší od jazykového nastavení uživatele. V tomto případě dochází ke zkreslení dat v důsledku dvou jevů:
  - Uživatel má na serveru nastavený svůj rodný jazyk, ale tweety odesílá v jiném v jazyce.

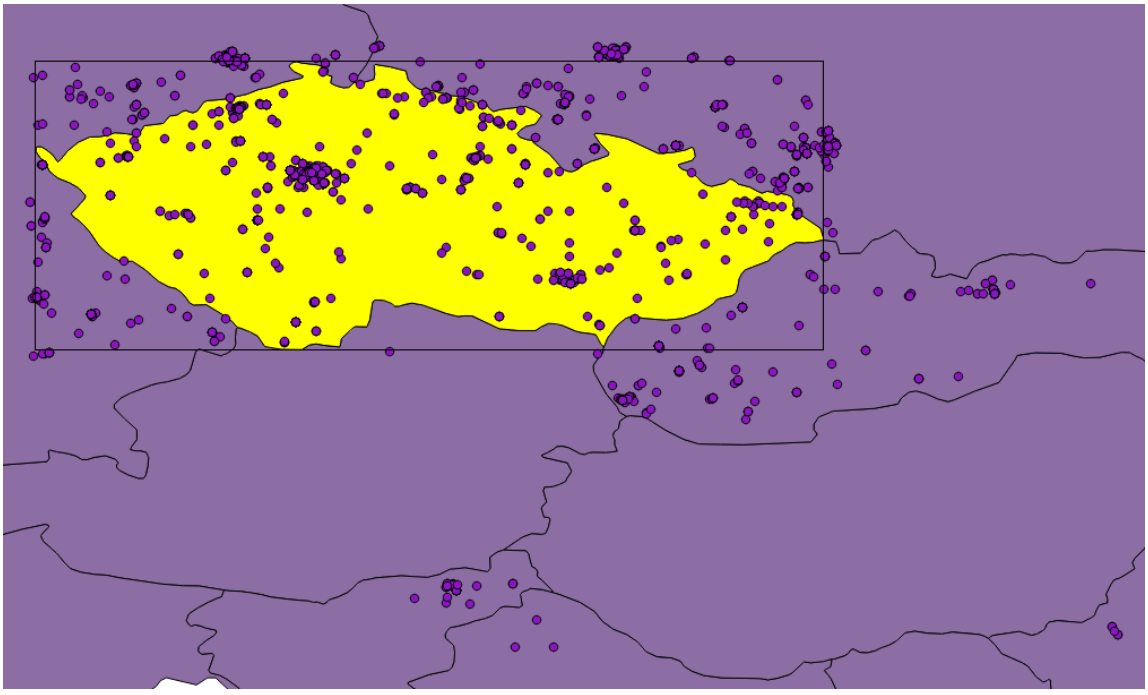
[19] Za předpokladu, že cesta ke spustitelnému souboru nodejs je definována v systémové proměnné a uživatel má k dispozici klíče požadované k ověření aplikace.

  - Uživatel nemá na serveru nastavený svůj rodný jazyk, ale tweety v něm odesílá.

  Při detekci jazyka upřednostňujeme nastavení konkrétního tweetu před serverovým nastavením, neboť vycházíme z předpokladu, že uživatel bude nejčastěji odesílat tweety psané v rodném jazyce. Typickým příkladem mohou být čeští uživatelé, kteří si na serveru jako výchozí jazyk nastavili angličtinu, ale tweety odesílají v češtině.

- Tweet je vygenerován automaticky jinou sociální sítí. V takovém případě je odesílán výhradně v angličtině a uživatel většinou nemá možnost či zájem text změnit.

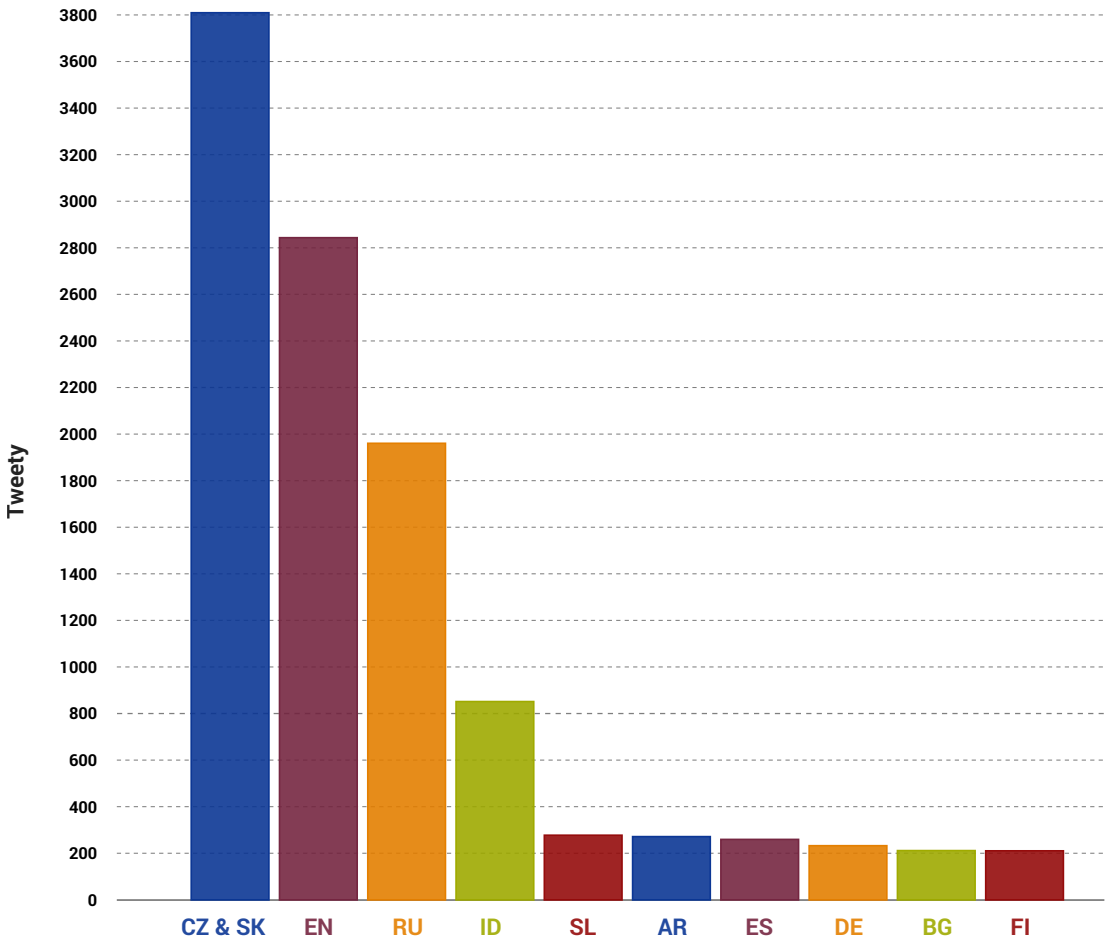
Nesprávně detekovaný jazyk tweetu není jediným problémem, který musíme při zpracování dat řešit. Twitter totiž z autorovi neznámých důvodů do streamu dat zahrnuje také tweety, které očividně nesplňují podmínku v požadavku odeslaném na server. Na obr. 12 vidíme, že ohraničující obdélník dat odpovídá rozsahu České republiky, přesto jsou k nám doručeny také tweety z území Slovenska, Rakouska či dokonce Slovinska, jejichž území do ohraničujícího obdélníku vůbec nezasahuje.



Obr. 12: Získané tweety při zadaném bounding boxu 12.09, 48.55, 18.87, 51.06.

Nežádoucí tweety byly z nasbíraných dat odstraněny a pro vizualizaci bylo použito 12 737 tweetů nasbíraných na území České republiky mezi 2. a 7. lednem roku 2014. Je na místě zmínit, že 5. ledna došlo k chybě v běhu aplikace, která znemožnila sběr dat mezi půlnocí a osmou hodinou ranní téhož dne. Data nebyla jinak upravena, reprezentují tedy stav, v jakém byla získána ze serverů Twitteru.

Data byla po zpracování zobrazena v interaktivní mapě za použití knihovny Leaflet (obr. 14). Autor do vizualizace zahrnul pouze deset nejpočetnějších jazyků



Obr. 13: Národy vybrané pro vizualizaci. Dvojpísmenné kódy vycházejí ze standardu ISO 639-1.

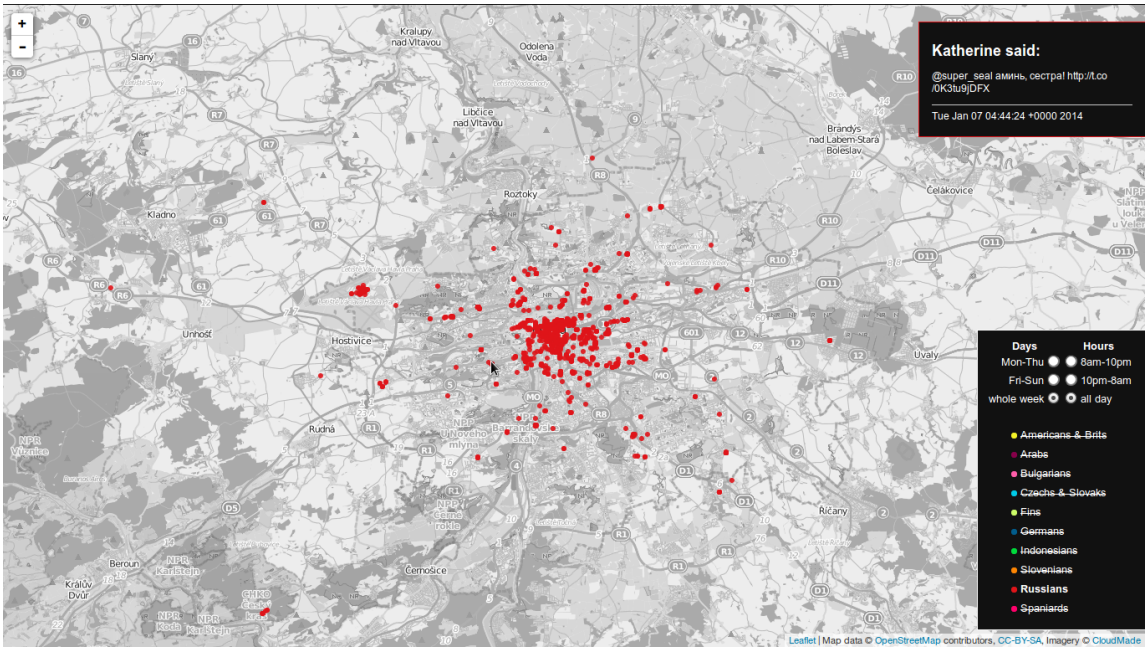
(viz obr. 13), kompletní soubor je však k dispozici v Příloze 10 na CD. Mapa umožňuje uživateli filtrovat zobrazené tweety jak podle jazyků, tak podle času. Lze tedy jednoduše porovnávat chování uživatelů různých národností v průběhu času. Autor k zobrazení dat zvolil tečkovou metodu a pomocí různých barev rozlišil jednotlivé národy.

Vzhledem k vysokému počtu zobrazených bodů je nutné počítat s pomalejší odezvou aplikace. Toto chování lze odstranit shlukováním bodů, to by však značně ztížilo čtení mapy a prostorové vzory by nám mohly zůstat skryty, proto jsme k jeho implementaci nepřistoupili.

Z užívání mapy a zkoumání dat můžeme konstatovat několik závěrů:

- Algoritmy pro automatickou detekci jazyka neumí spolehlivě identifikovat všechny jazyky. Z toho důvodu kupříkladu nelze rozlišit české a slovenské tweety. Také drtivá většina slovinských tweetů je výsledkem špatného odhadu algoritmů.

  Můžeme však tvrdit, že algoritmy velmi spolehlivě detekují pro nás „exotické“ jazyky. Jejich úspěšnost je vysoká především u jazyků, které nepoužívají latinku.



Obr. 14: Ukázka vizualizace tweetů (Příloha 10 na CD): Rusové na území Prahy.

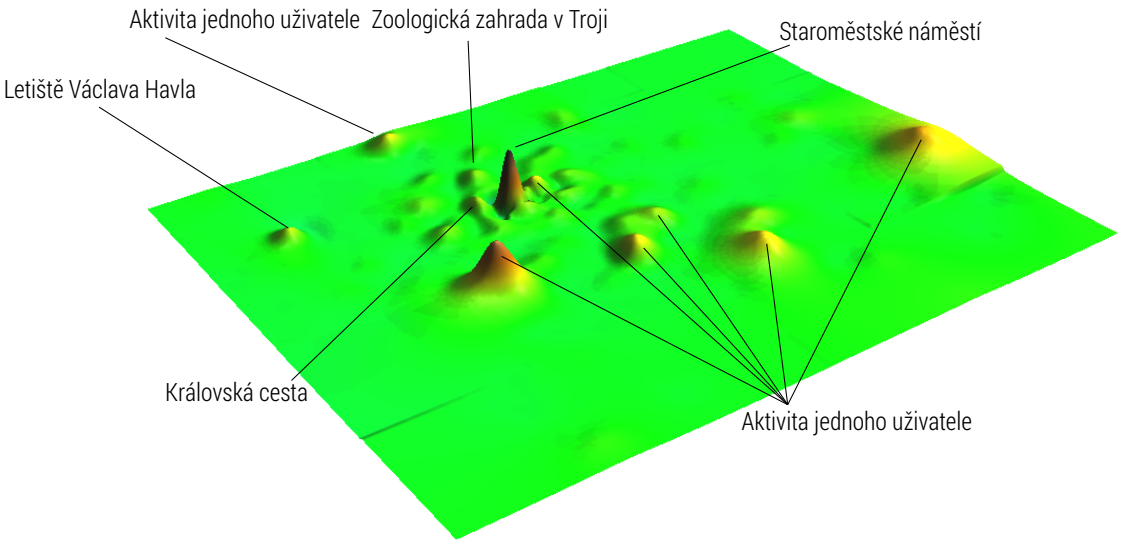
- Praha je místem s největším počtem cizojazyčných tweetů bez rozdílu národnosti jejich autorů. Mapa v Příloze 10 na CD pak dále odhaluje přítomnost ruské menšiny v Karlových Varech či Mariánských Lázních, tweetující Araby v Teplicích, v Olomouci pak nalezneme skupinu Indonésanů[20].

  Za vstupní bránu do České republiky bychom mohli označit Letiště Václava Havla, které je, stejně jako Praha, bez rozdílu národnosti významným bodem na mapě takřka pro všechny zobrazené skupiny. Jeho význam je dobře patrný z obr. 14, kde jej představuje shluk bodů západně od Prahy.

Tweety uložené v SQLite databázi je možné exportovat do formátu CSV, který lze načíst do QGISu. Zde můžeme provést potřebné úpravy, data exportovat do GeoJSON souboru a ten následně použít pro načtení bodů do mapy.

Kromě interaktivní webové mapy se autor pokusil zachytit rozložení tweetů pomocí 3D vizualizace (obr. 15). Za tímto účelem bylo zájmové území, vymezené konvexním obalem pražské zástavby získané z projektu OpenStreetMap, pokryto pravidelnou hexagonovou sítí, v níž každý hexagon představuje území o rozloze asi 8,5 km^2. Tato síť byla použita pro získání počtu tweetů v každém z jejích prvků. Ty byly poté nahrazeny svými centroidy nesoucími informaci o počtu tweetů spadajících do příslušného hexagonu. V GRASS GIS z nich byl pomocí splinové funkce vypočten model povrchu, kdy jako výška bodů byl použit právě počet tweetů, které reprezentují. Vizualizaci modelu v prostředí NVIZ vidíme na obr. 15. Získali jsme tak model „sociální“ Prahy, tedy města tvořeného aktivitou na Twitteru, který je od své skutečné předlohy velmi odlišný a umožňuje nám identifikovat místa s nejvyšší koncentrací tweetů.

[20]Zde můžeme konstatovat, že se jedná s velkou pravděpodobností o studenty, neboť většina tweetů je odesílána z prostoru vysokoškolských kolejí.



Obr. 15: „Sociální“ Praha: 3D model města založený na množství tweetů. Pohled od jihozápadu.

Tato ohniska můžeme rozdělit na dva druhy:

- Jedna jsou výsledkem aktivity jednoho či několika málo jedinců a jako taková mnoho nevypovídají o prostorovém chování populace Twitteru na daném území, nejsou pro nás tedy příliš zajímavá.
- Druhá jsou výsledkem aktivity mnoha uživatelů a dávají nám obraz o tom, která místa na území Prahy přitahují pozornost mnoha lidí. Jedná se o následující lokality.
  - Letiště Václava Havla, které jsme zmiňovali již dříve, je pro mnoho návštěvníků Prahy prvním a rovněž posledním místem, které během svého pobytu navštíví,
  - Zoologická zahrada v Troji, která byla v roce 2007 časopisem Forbes označena jako sedmá nejlepší zoologická zahrada na světě,
  - Královská cesta,
  - Staroměstské náměstí, které koncentrací tweetů zdaleka přesahuje ostatní lokality.

Shrnutí

Pro území České republiky můžeme hypotézu zmíněnou na začátku části 3.1.5 potvrdit, neboť jsme během sběru dat filtrováním nepřekročili hranici 1 % provozu na

serveru, a byly nám doručeny všechny tweety. V opačném případě by data mohla být zkreslena, aniž by bylo možné zjistit skutečný poměr zastoupení jazyků.

Přestože pracujeme se všemi tweety, které byly v průběhu sledovaného období na území České republiky odeslány, složení populace sociální sítě nemusí nutně odrážet složení populace reálné. Z toho důvodu je na místě opatrnost při vynášení závěrů založených na získaných datech: to, že někde identifikujeme etnikum tweetující v jiném jazyce než češtině, neznamená, že se totéž etnikum nenalézá rovněž na dalších místech, kde o sobě prostřednictvím sociálních sítí akorát nedává vědět.

Jak jsme si ukázali, algoritmy pro automatickou detekci jazyka mohou produkovat poměrně velké množství chyb, které bychom před zevrubnou analýzou museli ručně odstranit. Vzhledem k objemu nasbíraných dat autor práce k tomuto kroku nepřistoupil a vizualizoval surová data.

Pomocí 3D modelu jsme zkonstruovali „sociální“ obraz města, který nám pomohl identifikovat místa, na nichž dochází k větší aktivitě uživatelů Twitteru.

3.1.6 Shrnutí

V této kapitole jsme se podrobně seznámili s některými způsoby získávání prostorových dat ze sociální sítě Twitter. Ukázali jsme si, jaká data mají k dispozici všichni uživatelé ve formě svých osobních archivů a k jakým datům se můžeme dostat prostřednictvím veřejného API. Popis funkcí API zdaleka není vyčerpávající, představili jsme si však principy REST a Streaming API, jejichž pochopení je nezbytné pro získání dat a následnou tvorbu vizualizací.

Výstupem této části práce jsou časová mapa tweetů autora (Příloha 6 na CD), časoprostorová kostka tweetů konkrétního uživatele vytvořená v prostředí Processing (Příloha 8 na CD) a interaktivní mapa rozlišující tweety podle jazyka, ve kterém byly napsány (Příloha 10 na CD). Podařilo se nám rovněž zkonstruovat 3D model hlavního města Prahy, kde místo výšek byly použity četnosti tweetů. Ten nám velmi názorně ukazuje, na jakých místech se aktivita uživatelů na Twitteru zvyšuje.

3.2 Foursquare

- Adresa služby: http://foursquare.com
- Počet uživatelů ve světě: více než 45 000 000
- Počet uživatelů v ČR: 48 000 (odhad gosquare.cz, 2013)

Dennis Crowley a Naveen Selvadurai spustili první verzi Foursquare v březnu 2009. Jedná se o sociální síť a zároveň mobilní aplikaci určenou ke sdílení informací o místech, která uživatelé navštívili. Foursquare funguje na principu gamifikace, uživatel totiž za každý check-in dostává body, které slouží k odemknutí badges[21]. Uživatelé mezi sebou soutěží a ten s největším počtem check-inů na daném místě získává titul mayor.

[21] Za deset check-inů z různých míst tak uživatel získá například badge Explorer.

Foursquare je zároveň location-based service, která uživateli v závislosti na jeho poloze a předchozím chování nabízí seznam zajímavých míst v jeho okolí. Ta jsou rozdělena do několika kategorií (jídlo, pití, nákupy...) a velmi často hodnocena dalšími uživateli, je tedy velmi snadné vybrat si místo na základě zkušenosti ostatních. Stejně tak je možné filtrovat místa, která již navštívili přátelé uživatele, která mají právě otevřeno nebo nabízejí nějakou specialitu vázanou na check-in (může se jednat například o nápoj zdarma).

Na základě výše uvedeného můžeme Foursquare považovat za online sociální hru, jejíž hráči jsou zajímavým cílem obchodníků. Názory a hodnocení uživatelů jsou cenným zdrojem informací, sociální síť lze navíc využít pro vytvoření pouta mezi podnikem a zákazníkem.

3.2.1 Přístupná data

Foursquare na rozdíl od Twitteru neumožňuje registrovaným uživatelům stáhnout data, která na server odeslali. Jedinou cestou k datům je tak API, podmínky použití služby jsou však přece jen benevolentnější a umožňují omezené využití web scrapingu (developer.foursquare.com, 2014), stejně tak u všech zdrojů nevyžadují bezpodmínečnou autentizaci prostřednictvím protokolu OAuth.

3.2.2 API: cesta k veřejně dostupnému obsahu

Foursquare API můžeme, podobně jako API Twitteru, rozdělit na dvě části: první z nich je RESTful API fungující tak, jak bylo popsáno v části 3.1.3, druhou je Real-Time API, které můžeme přirovnat k Twitter Streaming API. Real-Time API nám však umožňuje získat pouze data uživatelů, kteří naši aplikaci autorizují. Z tohoto důvodu se v následujícím textu budeme věnovat výhradně RESTful API. Bližší informace o Real-Time API jsou k dispozici v dokumentaci (developer.foursquare.com, 2014).

Zatímco Twitter API nám umožnilo po autentizaci přistupovat k jakémukoliv zdroji, Foursquare je v poskytování přístupu k informacím o uživatelích opatrnější. Ani po autentizaci aplikace tak není možné získat informace o check-inech konkrétního uživatele, pokud si tvůrce aplikace nepřidal do seznamu svých přátel.

Foursquare API je rozděleno na jednotlivé zdroje, na něž jsou vázány pohledy a akce, které s nimi lze provádět. Jako příklad můžeme uvést zdroj `users`, z nějž pro uživatele se zadaným identifikátorem získáme seznam jeho badges (pohled) a takového uživatele prostřednictvím API například odebereme ze seznamu přátel (akce).

Jediným zdrojem, který lze bez překážek využít ke zkoumání pohybu osob, respektive významu lokalit v rámci online komunity, je zdroj `venues`, kterým lze mimo jiné prozkoumávat místa v okolí zadaných souřadnic.

Stejně jako Twitter i Foursquare uplatňuje hodinové limity na počet dotazů, které se vždy vážou ke konkrétnímu zdroji. V případě `venues` se jedná o 5 000 dotazů za hodinu (developer.foursquare.com, 2014). V podmínkách použití (foursquare.com, 2014) se dále odráží choulostivá povaha dat ve vztahu k bezpečnosti uživatele, v části IV tvůrce aplikací souhlasí s tím, že:

> You may not track a user's check-in history or retain any data derived from a user's check-in history without first making the desired use clear to the user and obtaining affirmative consent to that use from that user. This includes tracking users via "here now"or top visitors of a venue.
> (https://foursquare.com/legal/api/platformpolicy)

Vzhledem k tomu, že mnoho tweetů zkoumaných v části 3.1.5 bylo automaticky odesláno právě z Foursquare, je možné tuto podmínku velmi jednoduše obejít právě stahováním tweetů namísto původních check-inů. Je otázkou, zda si uživatelé Foursquare vůbec uvědomují nebezpečí, která se mohou skrývat ve sdílení informací mezi sociálními sítěmi, jejichž jsou členy.

3.2.3 Neviditelné město a jeho obyvatelé

Autor již v části 3.1.5 prostřednictvím obr. 15 naznačil, že prostředí sociálních sítí může deformovat prostor, který běžně vnímáme. Zajímavé ukázky těchto rozdílů přináší projekt Flowing City (http://flowingcity.com/).

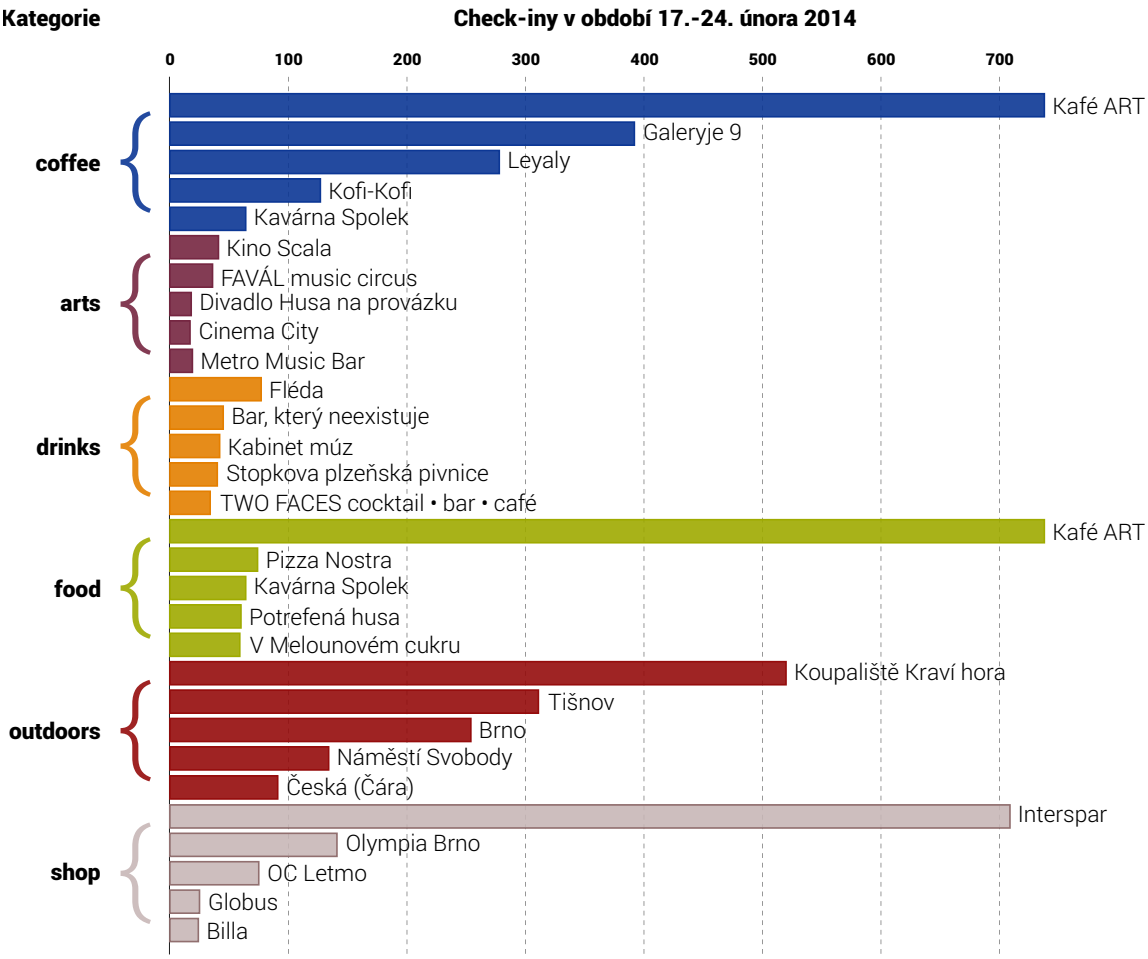
Aktivita uživatelů sociální sítě může být zdrojem informací o prostorovém chování populace na určitém místě, může rovněž odhalit, která místa lidi přitahují a kterým se naopak vyhýbají. Stejně tak z ní můžeme odvodit rytmus míst, jak o něm píše Lavická (2012).

Pro získání dat z Foursquare API znovu využijeme Python a knihovnu foursquare (https://pypi.python.org/pypi/foursquare), která nám práci se zdroji značně ulehčí. Kvůli již zmiňovaným omezením v přístupu k některým datům se autor rozhodl podrobněji prozkoumat venues na území města Brna, která lze získat ze stejnojmenného zdroje pomocí metody `explore`. Kompletní kód uvádí Příloha 12, zde zmíníme znovu pouze nejdůležitější části.

Kód 12: Získání seznamu venues prostřednictvím REST API.

```
...
categories = ['food', 'drinks', 'coffee', 'shops', 'arts', 'outdoors']
...
for cat in categories:
    response = client.venues.explore(params={
        'll': latlon,
        'section': cat,
        'radius': radius,
        'v': '20131016',
        'intent': 'browse',
        'limit': '50'
    })
...
```

Pro práci s API se nejdříve musíme autentizovat pomocí protokolu OAuth. Ke stahování dat využijeme dělení venues do kategorií definovaných na Foursquare (proměnná `categories`). Pro každou kategorii pak budeme v okolí bodu se souřadnicemi `latlon` hledat v okruhu o velikosti `radius` konkrétní venues (kód 12). Takto můžeme



Obr. 16: Check-iny uživatelů dle kategorií míst v období 17.-24. února 2014.

získat vždy maximálně 50 položek, přičemž se autorovi nepodařilo zjistit, jakým způsobem Foursquare rozhoduje o tom, které položky budou v odpovědi vráceny. Drtivá většina venues byla v průběhu stahování dat obsažena ve všech odpovědích serveru. Dokumentace bohužel nezmiňuje, zda například při delší neaktivitě uživatelů na daném místě může být místo vyřazeno z odpovědi vrácené dotazovaným zdrojem.

Odpověď následně projdeme a zapíšeme do CSV souboru s názvem určeným výrazem `%Y_%m_%d-%H_%M_cat.csv`[22]. Autor data sbíral v období 17.-24. února 2014 vždy v osm hodin ráno, v poledne, ve čtyři hodiny odpoledne, v osm hodin večer a v jedenáct hodin večer. Každý z uložených souborů obsahuje hlavičku s názvy sloupců a poté odpovídající záznamy venues. Sloupce reprezentují:

- `name`: název místa
- `lon`: zeměpisná délka místa
- `lat`: zeměpisná šířka místa

[22]Soubor 2014_02_17-12_00_drinks.csv byl tedy stažen 17. února 2014 ve 12 hodin a obsahuje údaje o kategorii drinks.

- `checkins`: celkový počet check-inů na místě
- `herenow`: počet uživatelů, kteří se na místě právě nacházejí
- `likes`: počet uživatelů, kteří na místě provedli akci like
- `tips`: počet tipů týkajících se místa
- `users`: celkový počet uživatelů, kteří na místě provedli check-in

Autor ze stažených dat získal celkový počet check-inů ve sledovaném období a pro každou kategorii stanovil pět nejnavštěvovanějších míst. Z grafu na obr. 16 je dobře patrné, že mezi místa, kde uživatelé nejčastěji provedou check-in, patří kavárny. Z nich je v tomto ohledu vůbec nejúspěšnější Kafé ART. Zajímavé je pořadí v kategorii obchodů, kde více než 700 check-inů za sledované období zaznamenal Interspar v Galerii Vaňkovka, za ním pak s velkým odstupem figurují Olympia Brno a OC Letmo.

Koupaliště Kraví hora zaujímá první místo v kategorii outdoors s více než 500 check-iny. Prudký nárůst z původních asi 300 check-inů na více než 600 během sledovaného období se odehrál v Tišnově.

Z grafu dobře vidíme, že lidé Foursquare nejčastěji využívají v kavárnách. Můžeme tedy usuzovat, že častěji provedou check-in tam, kde tráví více času.

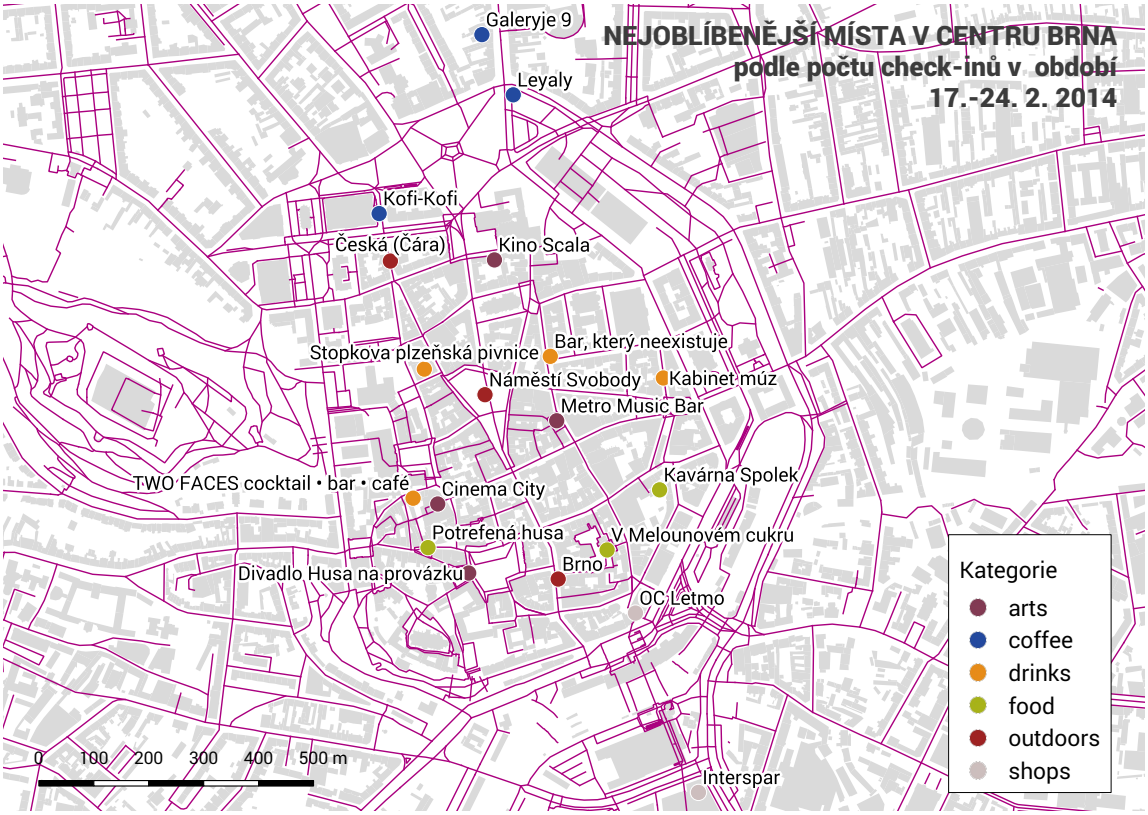
Obr. 17 ukazuje distribuci těchto míst v městském prostoru. Devatenáct z celkových třiceti nejpopulárnějších míst se nachází v centru Brna a lze je ohraničit polygonem o ploše menší než 0,5 km^2.

Dalším zajímavým údajem týkajícím se Foursquare venues je počet návštěvníků, kteří se na daném místě v okamžiku získání dat nacházejí. Autor vycházel z předpokladu, že na základě počtu check-inů by bylo možné odhalit závislost pohybu uživatelů sítě na čase (např. ráno převládá návštěvnost obchodů, v pozdních hodinách bude naopak více uživatelů v barech a restauracích). Tuto myšlenku však nelze pro Brno ze získaných dat potvrdit, neboť neposkytují dostatečně velký vzorek uživatelů. Přesto se můžeme domnívat, že pro statisticky významnější počet uživatelů[23] bychom byli schopni podobné zákonitosti odhalit.

Kromě údajů o počtu check-inů ve sledovaném období máme ze stažených dat k dispozici také informace o celkovém počtu check-inů od doby, kdy bylo místo na Foursquare přidáno. Z tabulky 2 vidíme, že pouze necelá polovina nejnavštěvovanějších míst z období 17.-24. února 2014 je stejně oblíbená také při časově neomezeném srovnání. Pomocí počtu check-inů na den můžeme srovnat oblíbenost jednotlivých míst.

Pomineme-li venue s názvem Brno, která je téměř se 7 check-iny denně jasně nejčastěji navštěvovaným místem, patří mezi venues s největším počtem check-inů na den náměstí Svobody, OC Olympia, East Village Bar & Diner a Bar, který neexistuje. Poslední dva můžeme považovat za příklady venues, která jsou na Foursquare teprve krátce (v době psaní práce 645, respektive 340 dní), což je paradoxně může ve sledovaném poměru check-inů na den zvýhodňovat, vezmeme-li v úvahu jednak

[23]Maximem bylo devět uživatelů vyskytujících se současně na jednom místě.



Obr. 17: Místa s největším počtem check-inů v období 17.-24. února 2014.

zvyšující se počet mobilních zařízení s připojením k internetu, jednak rostoucí oblibu Foursquare.

Velkou výhodu v tomto srovnání získávají rovněž venues, jejichž provoz není výrazně omezen časem. Například v obou brněnských divadlech uvedených v tabulce 2 bude možné provést check-in jen v době představení, tedy během poměrně krátkého času, kdežto bary, restaurace nebo nákupní centra jsou přístupná nepoměrně delší dobu.

Z tabulky 2 jsou patrné také rozdíly mezi kategoriemi. Nejméně check-inů denně zaznamenáváme u venues z kategorie Arts a Coffee, nejvíce naopak v kategorii Shop a Food. Hodnoty kategorie Outdoors nemůžeme kvůli zmiňovanému venue s označením Brno považovat za reprezentativní.

Uvedli jsme některé možné příčiny rozdílů v návštěvnosti venues. Zůstává otázkou, do jaké míry má na počet check-inů vliv sociodemografická charakteristika návštěvníků, respektive jaké podněty vedou k tomu, že někde uživatelé provádějí více check-inů než jinde, přestože jsou si venues podobná. Proč je například počet check-inů v Baru, který neexistuje, dvakrát větší než v Bistru Franz? Na podobné otázky bohužel bez znalosti návštěvníků nemůžeme odpovědět.

Vrátíme-li se ještě k nejnavštěvovanějším venues na obr. 17, bude nám jistě nápadná jejich koncentrace na malém prostoru. Tento očekávatelný jev dle autora souvisí s prostorovou syntaxí, jak ji definovali Hillier et al. (1976). Jedná se o teorii vysvětlující vzájemné působení prostoru a člověka, jehož výsledkem je jednak

určité uspořádání prostoru, jednak určité chování lidí v něm. Získané poznatky nám umožňují předpovídat výskyt lidí na místech, která mají v rámci sítě nejvyšší konektivitu. Ačkoliv jsme v rámci práce nezkoumali topologii uliční sítě Brna, lze předpokládat, že právě centrum města bude se zbytkem sítě velmi dobře propojeno. Výsledkem tohoto propojení je nutně větší počet lidí vyskytující se v oblasti, který zcela jistě přitahuje provozovatele barů a restaurací.

3.2.4 Shrnutí

Ačkoliv je sociální síť Foursquare na rozdíl od Twitteru zaměřena primárně na sdílení prostorových informací, je získání těchto dat paradoxně obtížnější. Data sdílená na Foursquare lze totiž považovat za choulostivější a potenciálně nebezpečnější než ta, která uživatelé sdílejí na Twitteru, a z tohoto důvodu API nenabízí zdaleka tolik možností k dolování dat jako API Twitteru.

Přístup ke zdrojům souvisejícím s aktivitou konkrétních uživatelů je podmíněn jejich autorizací, bez níž lze získat pouze „bezpečná“ data o venues, tedy místech, která uživatelé navštěvují. Jak jsme se pokusili dokázat, mohla by i tato být cenným zdrojem informací o pohybu uživatelů sítě jako celku, v Brně však síť nemá dostatek uživatelů na to, abychom z jejich chování mohli vyvodit reprezentativní závěry.

Avšak i pokud bychom získali dostatečné množství dat (těžením na území města s větším počtem uživatelů), dokázali bychom popsat pouze masové chování populace sítě. Mohli bychom tedy například konstatovat, že „*v nočních hodinách se větší množství uživatelů vyskytuje v podnicích určitého typu*“, už bychom však nebyli schopni zjistit informace o jejich věku či vzdělání, neboť i ty jsou pro aplikace, které uživatelé neautorizovali, nepřístupné.

Tab. 2: Brněnské venues s největším počtem check-inů k 24. únoru 2014.
Kurzívou jsou uvedeny nejnavštěvovanější venues v období 17.-24. února 2014, viz obr. 17.

Kategorie	**Venue**	**Dny**	**Check-iny**	**Check-iny/Dny**
Arts	*Cinema City (Velký Špalíček)*	1 466	*1 426*	1,0
Arts	Cinema City (OC Olympia)	1 539	1 027	0,7
Arts	Metro Music Bar	1 480	907	0,6
Arts	Janáčkovo divadlo	1 460	500	0,3
Arts	Městské divadlo Brno	1 259	456	0,4
Coffee	*Kavárna Spolek*	1 481	*1 043*	0,7
Coffee	Coffee Fusion	963	905	0,9
Coffee	Cafe Podnebi	1 520	744	0,5
Coffee	*Kofi-Kofi*	915	*678*	0,7
Coffee	Café Mezzanine	1 542	630	0,4
Drinks	*Bar, který neexistuje*	645	*1 531*	2,0
Drinks	*Fléda*	1 536	*1 314*	0,8
Drinks	*Stopkova plzeňská pivnice*	1 071	*1 243*	1,4
Drinks	Pivovar Pegas	1 500	1 187	0,8
Drinks	Mamut Pub	1 502	793	0,5
Food	*Potrefená husa*	855	*1 176*	1,4
Food	*Kavárna Spolek*	1 481	*1 043*	0,7
Food	Bistro Franz	825	759	0,9
Food	Kavárna Kunštátská Trojka	1 447	744	0,5
Food	East Village Bar & Diner	340	739	2,2
Outdoors	*Brno*	695	*4 776*	6,9
Outdoors	*Náměstí Svobody*	1 520	*3 630*	2,4
Outdoors	Zelný trh	1 464	1 552	1,1
Outdoors	*Česká (Čára)*	1 272	*1 351*	1,1
Outdoors	Moravské náměstí	1 460	1 323	0,9
Shop	*Olympia Brno*	1 478	*3 895*	2,6
Shop	IKEA	1 518	2 291	1,5
Shop	Avion Shopping Park	1 326	1 750	1,3
Shop	NC Královo Pole	1 574	1 404	0,9
Shop	Campus Square	1 437	1 266	0,9

3.3 Instagram

- Adresa služby: http://instagram.com
- Počet uživatelů ve světě: 150 000 000 (instagram.com, 2014)
- Počet uživatelů v ČR: asi 14 000 (Dočekal, 2013)

Instagram je sociální síť využívaná ke sdílení fotografií a videí, jejíž první verzi spustili v roce 2010 Kevin Systrom a Mike Krieger. Určena je pro uživatele mobilních zařízení, kteří pomocí stejnojmenné aplikace mohou mediální záznamy pořizovat, upravovat a následně sdílet s dalšími uživateli sítě.

Během prvního dne své existence se do sítě zaregistrovalo 25 000 uživatelů. Za další tři měsíce síť sloužila 1 000 000 uživatelů, ani ne po roce se jejich počet zdesetinásobil[24](Cutler, 2012). V té době přitom existovala pouze aplikace pro telefony iPhone s operačním systémem iOS, uživatelé operačního systému Android či Windows Phone neměli možnost síť využívat.

Právě uvolnění aplikace pro operační systém Android, kterou si během prvního dne stáhl více než 1 000 000 uživatelů, pravděpodobně vedlo Facebook k jednání o odkoupení Instagramu. 9. dubna 2012 se tak rychle rostoucí Instagram stal za 300 milionů dolarů a 23 milionů akcií Facebooku splatných po vstupu na burzu součástí světově nejpoužívanější sociální sítě (Protalinski, 2012).

Instagram vznikl později než dříve zkoumané sítě, a mohl se u nich tak v některých ohledech inspirovat. Využívá hashtagy, které poprvé představil Twitter. Stejně jako Foursquare nabízí uživatelům možnost připojit k příspěvku geotagy a lokalizovat tak pořízenou fotografii či video. Zatímco Twitter lze plnohodnotně využívat jak prostřednictvím webového rozhraní, tak z mobilního zařízení, Foursquare a Instagram jsou určeny k použití na mobilních zařízeních, zatímco webové rozhraní slouží jako prohlížečka již existujících dat.

S rostoucím počtem chytrých mobilních zařízení vybavených fotoaparáty lze očekávat také růst uživatelské základny Instagramu.

3.3.1 Přístupná data

Instagram v současné době neumožňuje uživatelům stažení dat nahraných prostřednictvím aplikace do sítě. Podobně jako Twitter a Foursquare však provozuje Realtime a REST API, které lze využít k získávání dat.

3.3.2 API: cesta k veřejně dostupnému obsahu

V porovnání s dříve zmíněnými sítěmi je dokumentace (instagram.com, 2014) Instagram API neúplná, jeho samotná implementace nám po dřívějších zkušenostech může připadat přinejmenším zvláštní a založení nové aplikace je dle názoru autora zbytečně komplikované.

[24]Registrace 10 000 000 uživatelů na Foursquare trvala více než dva roky.

Jelikož ověření přístupu k API obstarává protokol OAuth, je potřeba nejprve zaregistrovat aplikaci, kterou budeme využívat ke získávání dat. Její registrace je dokončena teprve po zadání bezpečnostního kódu doručeného na zadané telefonní číslo SMS zprávou. Autorovi práce není známo, co vede Instagram ke zvýšené opatrnosti při zakládání aplikací, domnívá se však, že objektivní důvody pro to ve srovnání s Twitterem či Foursquare neexistují.

Jak jsme zmínili, také Instagram poskytuje Realtime API, jehož implementaci na serverech Twitteru jsme si představili v části 3.1.5. Bohužel i zde Instagram postupuje z hlediska našich předchozích zkušeností poněkud nestandardně, a to hned ve dvou ohledech.

Tvůrce aplikace musí při její registraci zadat adresu callbacku, tedy funkce volané po autorizaci aplikace uživatelem, která musí být na internetu veřejně dostupná. V předchozích případech takové omezení neplatilo, a bylo tedy možné použít adresu lokálního serveru. Toto omezení je nicméně vzhledem k implementaci Realtime API na Instagramu nezbytné.

Celý mechanismus funguje zhruba následujícím způsobem:

1. uživatel zaregistruje aplikaci,
2. vytvoří skript schopný zpracovávat GET a POST požadavky, který umístí na adresu callbacku,
3. prostřednictvím API zašle požadavek na subscription, tedy vytvoření objektu určeného ke sledování v reálném čase, přičemž tím může být:
   - objekt User
   - objekt Tag
   - objekt Location
   - objekt Geography

   Příklad založení subscription uvádí kód 13.

   Kód 13: Vytvoření objektu Geography (převzato instagram.com, 2014).

```
curl -F 'client_id=CLIENT-ID' \
    -F 'client_secret=CLIENT-SECRET' \
    -F 'object=geography' \
    -F 'aspect=media' \
    -F 'lat=35.657872' \
    -F 'lng=139.70232' \
    -F 'radius=1000' \
    -F 'callback_url=http://YOUR-CALLBACK/URL' \
    https://api.instagram.com/v1/subscriptions/
```

4. Instagram od této chvíle začne na adresu callbacku doručovat informace o dostupnosti nových dat.

Na rozdíl od Twitteru však nezasílá data samotná. Je tak už na autorovi aplikace, aby je ze serverů získal, a to pravděpodobně pomocí REST API s definovanými parametry stránkování.

Povinnost učinit adresu callbacku veřejně dostupnou nám, s ohledem na služby tuzemských hostingů, velmi zužuje výběr technologií. Alternativou ke standardnímu hostingu může být Heroku (http://heroku.com), které nabízí širokou paletu služeb v cloudu, případně aplikace ngrok (http://ngrok.com) umožňující zpřístupnit porty lokálního serveru prakticky odkudkoliv.

Autor se na základě těchto zjištění a po prozkoumání možností REST API rozhodl nepřistoupit k tvorbě aplikace využívající Realtime API.

Data tak byla získávána prostřednictvím dvou zdrojů REST API, `media` a `tags`. V obou případech byly v období 17.-24. března 2014 nové fotografie každých deset minut ukládány do databáze Spatialite. Autor se nemůže zaručit za úplnost stažených dat, neboť dokumentace neuvádí, zda, a případně jak, jsou odpovědi z těchto zdrojů omezeny maximálním počtem vrácených položek.

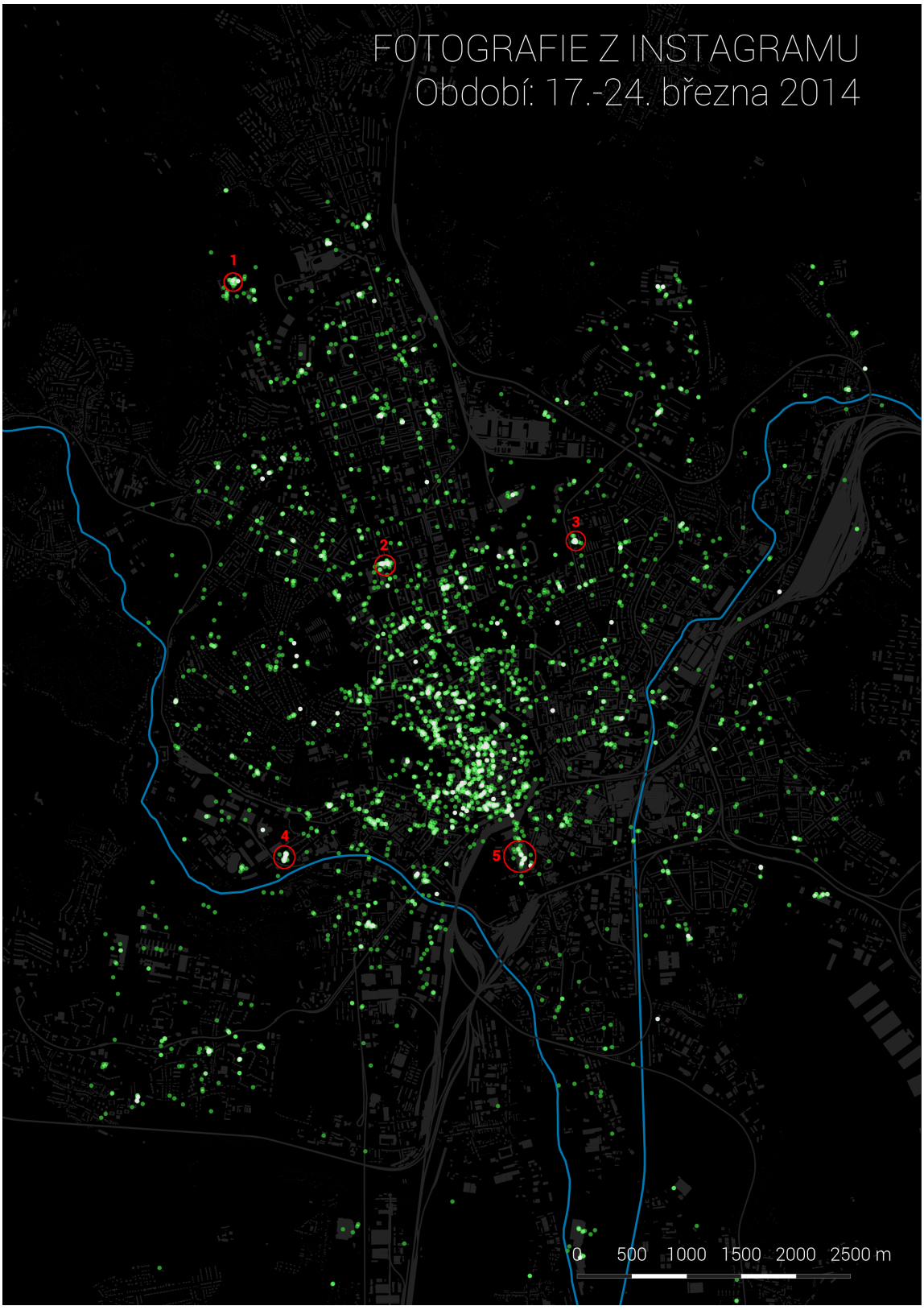
Ze zdroje `media` byly stahovány fotografie v dosahu pěti kilometrů od souřadnic $49,19479\,^\circ$ s. š. a $16,60888\,^\circ$ v. d. V uvedeném období bylo získáno 3 478 fotografií od 1 407 různých uživatelů. Ke všem záznamům jsou připojeny GPS souřadnice místa jejich pořízení. Při práci s tímto zdrojem autor využil knihovnu python-instagram (https://github.com/Instagram/python-instagram), kterou však kvůli neúplné dokumentaci pro případné budoucí nasazení nedoporučuje. Kód skriptu je k dispozici v příloze 13 na CD.

Rozmístění odeslaných fotografií ukazuje obr. 18, kde nejsvětlejší odstíny reprezentují nejvyšší hustotu fotografií. Je patrné, že tyto shluky se nacházejí především v centru města (náměstí Svobody, prostor ulic Česká a Masarykova). Opět se potvrzuje výrazná dominance malého území centra, na němž dochází k vyšší aktivitě uživatelů na sociálních sítích. Se stejným jevem jsme se setkali také v části věnující se síti Foursquare.

Autor kromě zmíněných lokalit identifikoval další místa ležící dále od centra města, která se množstvím odeslaných fotografií vymykají svému okolí (na obr. 18 vyznačena červeně). Mezi ně mimo jiné patří:

1. Koleje pod Palackého vrchem
2. Koleje Kounicova 50
3. třída Generála Píky č. p. 2036
4. hotel Voroněž
5. Galerie Vaňkovka

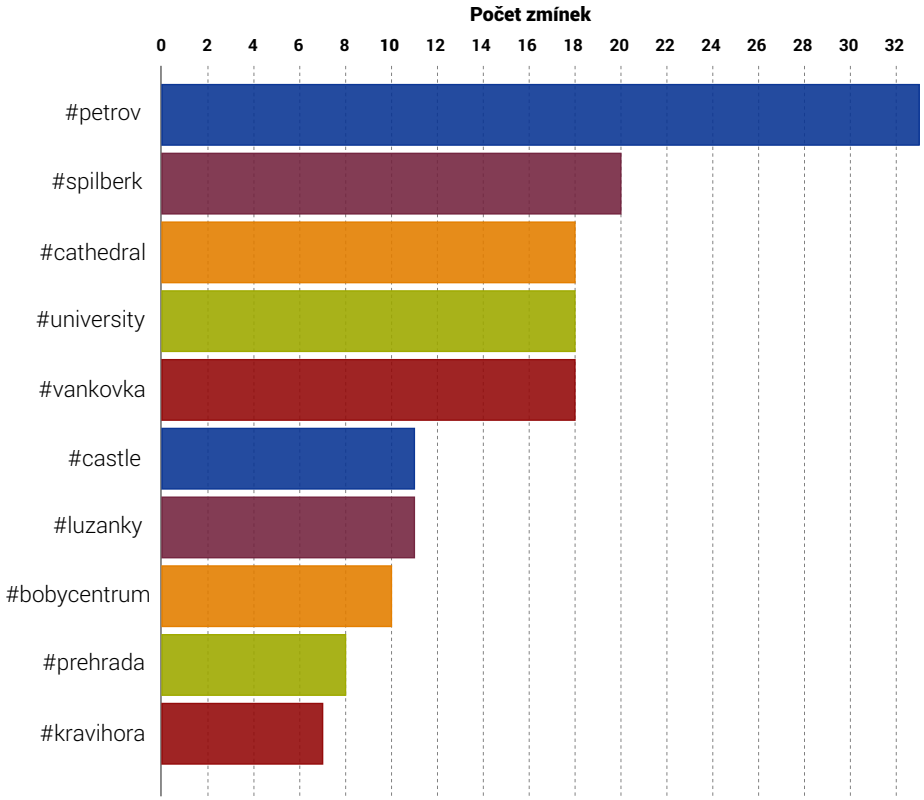
Je třeba zdůraznit, že souřadnice připojené k fotografii označují místo, ze kterého byla odeslána aplikací na server, nikoliv místo jejího pořízení. Není v silách autora zkontrolovat, na kolika ze získaných bodů byly fotografie ve skutečnosti pořízeny, a na kolika pouze došlo k jejich odeslání. URL adresy fotografií jsou nicméně obsaženy ve stažených datech, a je možné je k takové analýze využít.



Obr. 18: Vizualizace fotografií odeslaných z Brna v období 17.-24. března 2014 (vysvětlivky v textu).

Autor zkoumáním získaných dat nabyl dojmu, že Instagram je, podobně jako Twitter, velmi vhodný ke sledování pohybu konkrétních osob. Některé shluky bodů na obr. 18 jsou tvořeny fotografiemi pouze jednoho uživatele (např. shluk č. 3), a dávají nám tak představu o jeho pohybu ve městě. Takto aktivní uživatelé bohužel snižují informační hodnotu získaných dat, neboť zkreslují celkovou aktivitu na síti. Při sběru dat v delším časovém období by tyto extrémy pravděpodobně zanikly, v týdenním řezu jsou však stále patrné.

Ze zdroje `tags` jsme ve stejném období a ve stejných intervalech získávali fotografie označené hashtagem #brno, které byly opět ukládány do databáze Spatialite. Fotografie stažené z tohoto zdroje nemusejí, na rozdíl od fotografií ze zdroje `media`, obsahovat informace o poloze odeslání. Na druhou stranu však lze předpokládat, že fotografie označené tímto hashtagem byly v Brně nejen odeslány, ale také pořízeny. U záznamů s připojenými souřadnicemi se můžeme domnívat, že tyto souřadnice reprezentují, i vzhledem k velikosti Brna, skutečné místo pořízení snímku.



Obr. 19: Místa v Brně s nejvíce hashtagy v období 17.-24. března 2014.

Autor provedl analýzu hashtagů, z nichž vybral ty, které reprezentují konkrétní místa na území města Brna. Jejich přehled přináší obr. 19. Je patrné, že mezi deseti nejzmiňovanějšími místy převažují dominanty města. Jedná se především o rozlehlejší prostory (Špilberk, Kraví hora), které jsou veřejnosti bez omezení přístupné. Lze předpokládat, že množství fotografií pořízených v různých lokalitách města bude souviset také s ročním obdobím či denní dobou. Domníváme se, že množství vznika-

jících dat bude rovněž velmi citlivé na významné události odehrávající se na území města. Z tohoto hlediska by dle autora velmi zajímavým obdobím byl první květnový týden, v němž probíhá majáles a s ním související akce.

Ke stažení dat ze zdroje `tags` autor nepoužil žádnou knihovnu obalující API Instagramu, kód je k dispozici v příloze 13 na CD. Jelikož zdroj nevyžaduje autentizaci uživatele, neznamenalo by použití knihovny zásadní výhodu.

3.3.3 Shrnutí

Dostupností dat se Instagram spíše než Foursquare podobá Twitteru. Přestože samotné založení aplikace je poněkud komplikované, námi vytěžené zdroje nepožadují autentizaci a přístup k datům nijak neomezují. Na rozdíl od Foursquare se tak můžeme dostat k fotografiím konkrétních uživatelů, aniž bychom je museli žádat o souhlas.

Vizualizace nám jednoznačně ukazuje stejný trend jako u sítě Foursquare: uživatelé jsou nejaktivnější v centru města. Autor se domnívá, že při dlouhodobějším sběru by data byla vhodná také ke zkoumání chování uživatelů v čase. Vzhledem k omezenému počtu záznamů autor k této analýze nepřistoupil.

Diskuse

Sociální sítě zkoumané v této práci představují jen nepatrný zlomek podobných služeb, přesto jsou, ať už počtem uživatelů, či finanční hodnotou, významnými hráči na trhu. Data, která jsou v nich shromažďována, jsou mnohdy úzce spjata s prostorem, z něhož je uživatelé sítí odesílají.

Jak jsme během práce dokázali, jejich získávání a využití se sítě snaží vyjít vstříc implementováním standardizovaných rozhraní, z nichž lze bezpečným způsobem přistupovat k informacím, které uživatelé sdílejí. Ověření aplikace serverem je u všech představených sítí prováděno protokolem OAuth. Stejně tak tyto sítě nabízejí dva typy veřejně dostupných API:

- REST API
- Streaming/Realtime API

Formát JSON se stal dnešním de facto standardem pro poskytování a výměnu dat prostřednictvím API, což potvrzují rovněž zkoumané sítě. Každý typ API je navržen za jiným účelem, záleží tedy na autorovi aplikace, který z nich zvolí jako vhodnější pro řešení konkrétního problému. Při získávání dat prostřednictvím Streaming/Realtime API je velmi důležité mít na paměti, že výsledkem dotazů na server nemusí být zdaleka všechna data, která tomuto dotazu vyhoví. Jak jsme se přesvědčili, sítě mohou omezovat množství vrácených dat, neomezený přístup je pak zpoplatněn.

Kromě těchto omezení, která se týkají především Streaming/Realtime API, spolehlivost výsledků snižuje nejistota týkající se také dat získaných prostřednictvím REST API. Odpověď zdroje na zaslaný požadavek bývá v některých případech omezena maximálním počtem vrácených záznamů, API už však nedefinuje způsob, kterým jsou vrácené záznamy vybrány.

Přesto jsou sociální sítě neocenitelným zdrojem informací, které lze automatizovaně těžit a analyzovat. Jen stěží bychom dnes nalezli jiné místo, ať už na internetu či mimo něj, z něhož můžeme získat podobné množství informací. Pomineme-li jejich kolísající kvalitu, dostává se nám do rukou nástroj, který lze využít v mnoha oblastech. Zmiňme například dnes čím dál používanější mobilní aplikace pro hlášení nejrůznějších negativních jevů (černých skládek, odpadků, ...), jejichž funkci by mohl velmi jednoduše zastoupit Twitter.

Autor se domnívá, že sociální sítě budou hrát zásadní roli v předpovídání událostí, případně jejich vyhodnocování v reálném čase. Jestliže jsme v úvodu práce zmiňovali sociální sítě jako prostředek k analýze událostí minulých, lze v budoucnu počítat s jejich využitím kupříkladu při preventivních opatřeních, neboť zvýšená aktivita

uživatelů na určitém místě může indikovat, že se v oblasti děje něco neobvyklého, potenciálně nebezpečného.

Na jedné straně můžeme sociální sítě vnímat jako studnu informací o chování davu, jsou však specifické právě tím, že v sobě skrývají také individuální složku popisující chování každého svého uživatele. Je otázkou, kolik z aktivních uživatelů si uvědomuje možná nebezpečí spojená s častým sdílením informací na internetu. Víme, že licenční podmínky zkoumaných sociálních sítí zapovídají jakékoliv zneužití získaných dat, potenciálního útočníka snažícího se nabourat soukromí uživatelů však tyto řádky jen sotva zastaví.

Minimálně pro uživatele Foursquare může být dobrou zprávou snaha sítě co nejvíce chránit jejich data, zůstává však především na uživatelích, jak si svou minimální hranici bezpečí nastaví, tedy zda se rozhodnou sdílet na internetu téměř vše o svém životě, či přece jen některé osobní informace nezveřejní. Na sklonku roku 2013 však byla na serveru DigiDay (McDermott, 2013) zveřejněna dle autorova názoru poměrně znepokojivá zpráva o tom, že Foursquare využívá technologii pasivního sledování uživatelů, díky níž jim může doporučit zajímavá místa v jejich blízkosti. I zde je znovu potřeba zastat se zkoumaných sociálních sítí, především Twitteru a Instagramu, jejichž mobilní aplikace při výchozím nastavení k odeslaným datům údaj o poloze uživatele neodesílají. Obecně lze na adresu všech tří sítí konstatovat, že v podmínkách použití velmi jasně deklarují, jaká data a za jakým účelem jsou o uživateli shromažďována. Je tedy na něm, zda vysloví s těmito podmínkami souhlas a službu bude využívat.

Závěr

V úvodu práce byly stanoveny čtyři hypotézy, jejichž platnost jsme se pokusili ověřit. Potvrdili jsme, že *uživatele je možné sledovat v reálném čase*, pokud o sobě dávají prostřednictvím sítě vědět. Je otázkou, zda si jsou vědomi tohoto rizika, či jen omylem aktivovali tuto možnost v nastavení služby a netuší, že s dalšími uživateli sdílejí informace o své poloze.

Stejně tak se nám podařilo prostřednictvím aktivity uživatelů a zkoumáním jimi používaného jazykového nastavení sítě *identifikovat občany jiných států na území České republiky*. Rovněž jsme však upozornili na problémy, které mohou z této analýzy vyplývat. Kvůli nim jsme se vyvarovali ukvapených závěrů, přesto se domníváme, že je možné data tímto způsobem využít. Jejich kompletně automatizované zpracování se nám po zkušenostech získaných při psaní práce jeví jako nepravděpodobné, při ověření spolehlivosti dat člověkem by však analýza mohla poskytnout přesné výsledky.

Domníváme se, že získaná *data by mohla být užitečná pro místní samosprávu*. Přestože uživatelé sítě často používají k publikaci čistě osobních, z obecného hlediska nezajímavých informací, mohla by je například kampaň vedená samosprávou právě třeba na sociálních sítí přimět k tomu, aby své účty využili jiným způsobem. V diskusi jsme zmínili možnou roli sociálních sítí při hlášení negativních jevů.

Autor se v práci pokusil data popsat a analyzovat, je si však vědom, že sociální geograf by z nich pravděpodobně dokázal získat více informací. Lze se proto domnívat, že například pro vědce zabývající se problematikou městského prostoru by tato data mohla být poměrně cenná.

Hypotézu o nepřímé úměře mezi vzdáleností uživatelů a počtem vazeb se nám ověřit nepodařilo. Z dat je v podstatě nemožné získat relevatní informace týkající se skutečného místa pobytu všech uživatelů - z toho důvodu by jakékoliv závěry byly jen velmi málo spolehlivé. Je na uživatelích sítí, zda do svého profilu zadají místo bydliště, a pokud tak neučiní, nemůžeme prostorové vztahy mezi nimi zkoumat. Na Twitteru bychom se k těmto informacím patrně dostali nejsnáze, časové limity API a množství dotazů, které bychom museli zaslat, však byly další faktory, které nám potvrzení této hypotézy znemožnily. Proto odkazujeme na práci Lengyela et al. (2013), která se zabývá vztahy mezi uživateli maďarské sociální sítě iWiW v závislosti na jejich vzdálenosti. Autoři na základě zkoumání více než 700 milionů vztahů konstatují, že také na internetu častěji vznikají vazby mezi geograficky blízkými uživateli.

Na vybraných sociálních sítích se nám podařilo ukázat možnosti vytěžování prostorových dat. K jejich vizualizaci byly zvoleny v kartografické tvorbě méně známé technologie, které dobře posloužily při zobrazení časového rozměru dat. V průběhu

zpracování veškerých dat byl zvláštní důraz kladen jednak na využití zdarma dostupných nástrojů, jednak na co nejvyšší míru automatizace procesů vedoucích od získání dat k jejich vizualizaci. V tomto ohledu se autorovi ve vztahu ke kartografické tvorbě nejslibněji jeví prostředí Processing použité k tvorbě časoprostorové kostky, které by bylo jistě vhodné k plně automatizované konstrukci tohoto typu vizualizace. Mezi slabiny našeho řešení patří absence dalších prvků, které by rozšířily množství informací v kostce zobrazených, jak je ukazují Kveladze a Kraak (2012).

Při vizualizaci tweetů v části Osobní archiv uživatele jsme nastínili možnosti explorace dat v prostředí webového prohlížeče. S využitím knihovny Leaflet jsme vytvořili aplikaci, která uživateli umožňuje procházet tweety na časové ose a sledovat tak pohyb jejich autora. Stejně tak může porovnávat obsah tweetů, případně se za pomoci panelu zobrazujícícho obsah tweetů přesouvat na místa jejich vzniku.

Rozvinutím této metody, kterou můžeme považovat za zjednodušený příklad *dynamic query*, jak jej popisuje Roberts (2005), vznikla aplikace prezentující data získaná v části Menšiny v České republice. Uživateli nabízí jednoduché filtrování dat podle času či vybraného atributu, v našem případě jazyka tweetu. Lze tak zkoumat nejen prostorové vzory, ale také jejich změny v závislosti na denní době či dnu v týdnu, navíc můžeme svou pozornost věnovat pouze uživatelům vybrané národnosti. Rozšířit funkcionalitu našeho řešení je možné například implementací *small multiples* (MacEachren et al., 2003), k níž jsme však již v práci nepřistoupili. Sada na sobě nezávislých mapových polí se synchronizovaným posunem a měřítkem by nám umožnila porovnat chování jednotlivých národů v čase - tento úkol je v současném řešení jen těžko proveditelný.

I přes zmíněné dílčí nedostatky některých vizualizací se domníváme, že se jedná o vhodný způsob prezentace dat ze sociálních sítí. Jelikož byly vytvořeny v open source nástrojích, není jejich budoucí rozpracování a vylepšení vyloučeno.

Literatura

Knihy, časopisy a sborníky

[1] ANDRIENKO, N., G. ANDRIENKO a P. GATALSKY. Visual Data Exploration using Space-Time Cube. In: *Proceedings of the 21st International Cartographic Conference*. Durban, South Africa, 2003, s. 1981-1983. Dostupné na WWW: http://geoanalytics.net/and/papers/ica03.pdf

[2] BOASE, J. a J. B. HORRIGAN. *The Strength of Internet Ties*. Washington, D.C.: Pew Internet & American Life Project, 2006. Dostupné na WWW: http://www.pewinternet.org/files/old-media/Files/Reports/2006/PIP_Internet_ties.pdf.pdf

[3] BOYD, D. m. a N. B. ELLISON. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*. 2007, roč. 13, č. 1, s. 210-230. ISSN 10836101. DOI: 10.1111/j.1083-6101.2007.00393.x. Dostupné na WWW: http://doi.wiley.com/10.1111/j.1083-6101.2007.00393.x

[4] DE SOLA POOL, I. a M. KOCHEN. Contacts and influence. *Social Networks*. 1978, vol. 1, issue 1, s. 5-51. DOI: 10.1016/0378-8733(78)90011-4. Dostupné na WWW: http://linkinghub.elsevier.com/retrieve/pii/0378873378900114

[5] FIELDING, R. T. *Architectural Styles and the Design of Network-based Software Architectures* [online]. Irvine, California, 2000 [cit. 2013-12-10]. Dostupné na WWW: http://www.ics.uci.edu/%7Efielding/pubs/dissertation/top.htm. Disertační práce. University of California.

[6] HALL, P. A. V. a G. R. DOWLING. Approximate String Matching. *Computing Surveys*. 1980, roč. 12, č. 4. Dostupné na WWW: http://biit.cs.ut.ee/%7Evilo/edu/2002-03/Tekstialgoritmid_I/Articles/Approximate/Hall_Dowling_Approximate_Matching_Review_1980_p381-hall.pdf

[7] HILLIER, B., A. LEAMAN, P. STANSALL a M. BEDFORD. Space syntax. *Environment and Planning B*. 1976, č. 3, s. 147-185. Dostupné na WWW: http://eprints.ucl.ac.uk/1062/

[8] KADUSHIN, Ch. *Understanding social networks: theories, concepts, and findings*. New York: Oxford University Press, 2012, xii, 252 s. ISBN 978-019-5379-471.

[9] KRAAK, M. J. The Space-Time Cube Revisited from a Geovisualization Perspective. In: *Proceedings of the 21st International Cartographic Conference*. 1995, s. 1988-1996. Dostupné na WWW: http://www.itc.nl/library/Papers_2003/art_proc/kraak.pdf

KVELADZE, I. a M. J. KRAAK. What do we know about the space - time cube from cartographic and usability perspective?. In: *Proceedings of AutoCarto 2012 : the international symposium on Automated Cartography*. Columbus, Ohio, USA: Ohio: Cartography and Geographic Information Society, 2012, 16 p. Dostupné na WWW: http://www.cartogis.org/docs/proceedings/2012/Kveladze_Kraak_AutoCarto2012.pdf

[10] LAVICKÁ, P. *Koncept chronotopu a jeho aplikace v geografickém výzkumu* [online]. 2012 [cit. 2014-02-22]. Bakalářská práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce Ondřej Mulíček. Dostupné na WWW: http://is.muni.cz/th/358003/prif_b

[11] LI, X.; KRAAK, M. J. New views on multivariable spatiotemporal data: the space time cube expanded. In: *International Symposium on Spatio-temporal Modelling, Spatial Reasoning, Analysis, Data Mining and Data Fusion*. 2005. p. 199-201. Dostupné na WWW: http://www.isprs.org/proceedings/XXXVI/2-W25/source/NEW_VIEWS_ON_MULTIVARIABLE_SPATIO-TEMPORAL_DATA_THE_SPACE_TIME_CUBE_EXPANDED.pdf

[12] MACEACHREN, A., D. XIPING, F. HARDISTY a G. LENGERICH. Exploring high-D spaces with multiform matrices and small multiples. *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No.03TH8714)*. IEEE, 2003, s. 31-38. DOI: 10.1109/INFVIS.2003.1249006. Dostupné na WWW: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1249006

[13] MILGRAM, S. The Small-World Problem. *Psychology Today*. 1967, roč. 1, č. 1, s. 61-67. Dostupné na WWW: http://measure.igpp.ucla.edu/GK12-SEE-LA/Lesson_Files_09/Tina_Wey/TW_social_networks_Milgram_1967_small_world_problem.pdf

[14] MISLOVE, A., M. MARCON, K. P. GUMMADI, P. DRUSCHEL a B. BHATTACHARJEE. Measurement and Analysis of Online Social Networks. In: *IMC'07: proceedings of the 2007 ACM SIGCOMM Internet Measurement Conference, San Diego, California, USA, October 24-26, 2007*. New York, N.Y.: Association for Computing Machinery, c2007, s. 29-42. ISBN 978-1-59593-908-1. DOI: 10.1145/1298306.1298311. Dostupné na WWW: https://mpi-sws.org/%7Egummadi/papers/imc2007-mislove.pdf

[15] ROBERTS, J. C. Exploratory Visualization with Multiple Linked Views. In: Jason DYKES, Alan M. MACEACHREN, Menno-Jan KRAAK, eds. *Exploring geovisualization*. Amsterdam: Elsevier. 2005. pp. 159-180. ISBN 0-804-4533-0.

[16] The Social Structure of Competition. BURT, Ronald S. *Structural holes: the social structure of competition*. 1. Harvard Univ. Press paperback ed. Cambridge, Mass.: Harvard University Press, 1992, s. 57-91. ISBN 067484372x.

[17] VIÉGAS, F. B. a J. DONATH. Social Network Visualization: Can We Go Beyond the Graph?. In: *Workshop on Social Networks for Design and Analysis: Using Network Information in CSCW*. 2004, s. 6-10. Dostupné na WWW: http://web.media.mit.edu/%7Efviegas/papers/viegas-cscw04.pdf

[18] WOLFF, K. H. *The sociology of Georg Simmel*. 1st ed. New York: Free press, 1969, lxiv, 445 s.

[19] ZHENG, T., M. J. SALGANIK a A. GELMAN. How Many People Do You Know in Prison?. *Journal of the American Statistical Association*. 2006, vol. 101, issue 474, s. 409-423. DOI: 10.1198/016214505000001168. Dostupné na WWW: http://www.tandfonline.com/doi/abs/10.1198/016214505000001168

Zákony, normy a standardy

[20] RFC 2616. *Hypertext Transfer Protocol – HTTP/1.1*. 1999. Dostupné na WWW: http://www.w3.org/Protocols/rfc2616/rfc2616.html

[21] RFC 6749. *The OAuth 2.0 Authorization Framework*. 2012. Dostupné na WWW: http://tools.ietf.org/html/rfc6749

Elektronické zdroje

[22] 100 Social Networking Statistics & Facts for 2012. *Visual.ly* [online]. 2013 [cit. 2013-01-18]. Dostupné na WWW: http://visual.ly/100-social-networking-statistics-facts-2012

[23] 23 key moments from Twitter history. GRIGGS, B. *CNN* [online]. 2013 [cit. 2013-09-23]. Dostupné na WWW: http://edition.cnn.com/2013/09/13/tech/social-media/twitter-key-moments/index.html

[24] A Geography of Twitter. GRAHAM, M. *Visualizing Data at the Oxford Internet Institute* [online]. 2012 [cit. 2013-11-27]. Dostupné na WWW: http://www.oii.ox.ac.uk/vis/?id=4fe09570

[25] About. *Python* [online]. 2013 [cit. 2013-10-05]. Dostupné na WWW: http://www.python.org/about/

[26] Beautiful Soup Documentation. *Beautiful Soup* [online]. 2013 [cit. 2013-10-05]. Dostupné na WWW: http://www.crummy.com/software/BeautifulSoup/bs4/doc/

[27] BRODY, H. I Don't Need No Stinking API: Web Scraping For Fun and Profit. *Hartley Brody* [online]. 2012 [cit. 2013-02-18]. Dostupné na WWW: http://blog.hartleybrody.com/web-scraping/

[28] CARLSON, N. The Real History Of Twitter. *Business Insider* [online]. 2011 [cit. 2013-09-23]. Dostupné na WWW: http://www.businessinsider.com/how-twitter-was-founded-2011-4?page=2

[29] CUTLER, K. From 0 To $1 Billion In Two Years: Instagram's Rose-Tinted Ride To Glory. *TechCrunch* [online]. 2012 [cit. 2014-03-15]. Dostupné na WWW: http://techcrunch.com/2012/04/09/instagram-story-facebook-acquisition/

[30] DOČEKAL, D. Český Instagram? Podle Obrazení přes 14 tisíc uživatelů a 1,5 milionu fotek. *Lupa.cz* [online]. 2013 [cit. 2014-03-15]. Dostupné na WWW: http://www.lupa.cz/clanky/cesky-instagram-podle-obrazeni-pres-14-tisic-uzivatelu-a-1-5-milionu-fotek/

[31] Foursquare API Platform Policy. *Foursquare* [online]. 2014 [cit. 2014-03-23]. Dostupné na WWW: https://foursquare.com/legal/api/platformpolicy

[32] Foursquare Demographics — Age, Gender, Education. ALIVE WIRED. *Alive Wired* [online]. 2010 [cit. 2014-03-22]. Dostupné na WWW: http://alivewired.com/foursquare-demographics-age-gender/

[33] Foursquare goes beyond the check-in with passive tracking. MCDERMOTT, J. *Digiday* [online]. 2013 [cit. 2014-03-23]. Dostupné na WWW: http://digiday.com/platforms/foursquare-longer-needs-check-ins-track-store-visits/

[34] FreeGeodataCZ. *GRASSwikiCZ* [online]. 2012 [cit. 2013-10-14]. Dostupné na WWW: http://grass.fsv.cvut.cz/wiki/index.php/FreeGeodataCZ

[35] Geokódování. *API Mapy.cz* [online]. 2013 [cit. 2013-10-14]. Dostupné na WWW: http://api4.mapy.cz/view?page=geocoding

[36] *GeoNames* [online]. 2013 [cit. 2013-02-18]. Dostupné na WWW: http://www.geonames.org/

[37] GET search/tweets. Twitter Developers [online]. 2013 [cit. 2013-12-27]. Dostupné na WWW: https://dev.twitter.com/docs/api/1.1/get/search/tweets

[38] *Instagram Developer Documentation* [online]. 2014 [cit. 2014-03-23]. Dostupné na WWW: http://instagram.com/developer/

[39] *Klábosení* [online]. 2014 [cit. 2014-04-23]. Dostupné z: http://www.klaboseni.cz/

[40] LENGYEL, B., A. VARGA, B. SÁGVÁRI a Á. JAKOBI. *Distance dead or alive: Online Social Networks from a geography perspective*. Budapest, 2013. Dostupné na WWW: http://web.ibs-b.hu/data/upload/file/wp1_13_iwiw_distance.pdf

[41] MITCHELL, A. a D. PAGE. *Twitter News Consumers: Young, Mobile and Educated*. 2013. Dostupné na WWW: http://www.journalism.org/files/2013/11/Twitter-IPO-release-with-cover-page-new2.pdf

[42] Most Visited Websites. *Internet Research, Anti-Phishing and PCI Security Services — Netcraft* [online]. 2013 [cit. 2013-01-18]. Dostupné na WWW: http://toolbar.netcraft.com/stats/topsites

[43] MySpace. *CrunchBase* [online]. 2013a [cit. 2013-01-18]. Dostupné na WWW: http://www.crunchbase.com/company/myspace

[44] NIELSEN HOLDINGS N. V. *State of the Media: The Social Media Report* [online]. 2012, 15 s. [cit. 2013-07-31]. Dostupné na WWW: http://www.nielsen.com/us/en/reports/2012/state-of-the-media-the-social-media-report-2012.html

[45] *Node.js* [online]. 2013 [cit. 2014-01-04]. Dostupné na WWW: http://nodejs.org/

[46] PEW RESEARCH CENTER. *The Demographics of Social Media Users – 2012*. 2013, 14 s. Dostupné na WWW: http://www.lateledipenelope.it/public/513cbff2daf54.pdf

[47] Pg_trgm. *PostgreSQL: Documentation* [online]. 2013 [cit. 2013-10-05]. Dostupné na WWW: http://www.postgresql.org/docs/9.1/static/pgtrgm.html

[48] POST statuses/filter. *Twitter Developers* [online]. 2013 [cit. 2013-12-27]. Dostupné na WWW: https://dev.twitter.com/docs/api/1.1/post/statuses/filter

[49] *PostgreSQL: Documentation: 9.1: PostgreSQL 9.1.10 Documentation* [online]. 2013 [cit. 2013-10-15]. Dostupné na WWW: http://www.postgresql.org/docs/9.1/static/datatype-textsearch.html

[50] *PostHistory* [online]. 2004 [cit. 2013-02-10]. Dostupné na WWW: http://alumni.media.mit.edu/%7Efviegas/posthistory/

[51] Press Page. *Instagram* [online]. 2014 [cit. 2014-03-15]. Dostupné na WWW: http://instagram.com/press/

[52] *Processing.org* [online]. 2013 [cit. 2013-12-28]. Dostupné na WWW: http://processing.org/

[53] *Processing.py* [online]. 2013 [cit. 2013-12-28]. Dostupné na WWW: https://github.com/jdf/processing.py

[54] PROTALINSKI, E. Facebook buying Instagram for $300 million, 23 million shares. *ZDNet* [online]. 2012 [cit. 2014-03-15]. Dostupné na WWW: http://www.zdnet.com/blog/facebook/facebook-buying-instagram-for-300-million-23-million-shares/12097

[55] Rate Limits. *Foursquare for Developers* [online]. 2014 [cit. 2014-03-23]. Dostupné na WWW: https://developer.foursquare.com/overview/ratelimits

[56] Real-Time API. *Foursquare for Developers* [online]. 2014 [cit. 2014-03-23]. Dostupné na WWW: https://developer.foursquare.com/overview/realtime

[57] REST API v1.1 Limits per window by resource. *Twitter Developers* [online]. 2013 [cit. 2013-12-27]. Dostupné na WWW: https://dev.twitter.com/docs/rate-limiting/1.1/limits

[58] REST API v1.1 Resources. *Twitter Developers* [online]. 2013 [cit. 2013-12-10]. Dostupné na WWW: https://dev.twitter.com/docs/api/1.1

[59] RITHOLTZ, B. History of Social Media. *The Big Picture* [online]. 2010 [cit. 2014-02-15]. Dostupné na WWW: http://www.ritholtz.com/blog/2010/12/history-of-social-media/

[60] Statistika Foursquare v ČR. *GoSquare* [online]. 2013 [cit. 2014-03-23]. Dostupné na WWW: http://www.gosquare.cz/4sq-stats-cz/

[61] The Google Geocoding API. *Google Developers* [online]. 2013 [cit. 2013-02-18]. Dostupné na WWW: https://developers.google.com/maps/documentation/geocoding/

[62] Timeline - Facebook's latest news, announcements and media resources. *Facebook* [online]. 2013 [cit. 2013-09-14]. Dostupné na WWW: http://newsroom.fb.com/Timeline

[63] *Tweepy v1.4 documentation* [online]. 2013 [cit. 2013-12-27]. Dostupné na WWW: http://pythonhosted.org/tweepy/html/index.html

[64] Tweets. *Twitter Developers* [online]. 2013 [cit. 2014-01-04]. Dostupné na WWW: https://dev.twitter.com/docs/platform-objects/tweets

[65] *Twitter* [online]. 2013 [cit. 2013-02-10]. Dostupné na WWW: https://twitter.com/

[66] Twitter Privacy Policy. *Twitter* [online]. 2012 [cit. 2013-12-08]. Dostupné na WWW: https://twitter.com/privacy

[67] Twitter Statistics. *Statistic Brain* [online]. 2013 [cit. 2013-09-14]. Dostupné na WWW: http://www.statisticbrain.com/twitter-statistics/

[68] Twitter Terms of Service. *Twitter* [online]. 2012 [cit. 2013-12-08]. Dostupné na WWW: https://twitter.com/tos

[69] Using OAuth. *Twitter Developers* [online]. 2013 [cit. 2013-12-27]. Dostupné na WWW: https://dev.twitter.com/docs/auth/using-oauth

[70] Venues Platform. *Foursquare for Developers* [online]. 2014 [cit. 2014-03-23]. Dostupné na WWW: https://developer.foursquare.com/overview/venues

[71] What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *O'Reilly Media* [online]. 2005 [cit. 2013-01-18]. Dostupné na WWW: http://oreilly.com/pub/a/web2/archive/what-is-web-20.html

[72] YARDI, S. a D. BOYD. *Tweeting from the Town Square: Measuring Geographic Local Networks* [online]. 2010 [cit. 2013-02-18]. Dostupné na WWW: http://www.danah.org/papers/2010/TweetingTownSquare.pdf

[73] Your Twitter Archive. *Twitter Blogs* [online]. 19. 9. 2012 [cit. 2013-09-30]. Dostupné na WWW: https://blog.twitter.com/2012/your-twitter-archive

Přílohy

- Příloha 1: Struktura PostgreSQL tabulky pro import osobního archivu ze sítě Twitter
- Příloha 2: Struktura PostgreSQL tabulky pro import záznamů Geonames
- Příloha 3: Jazykové nastavení fulltextového vyhledávání v PostgreSQL
- Příloha 4: Porovnání výsledků fulltextového vyhledávání v databázi GeoNames při různých jazykových nastaveních
- Příloha 5: Animovaná mapa geokódovaných tweetů autora práce (na CD)
- Příloha 6: Interaktivní mapa geotagovaných tweetů autora práce (na CD)
- Příloha 7: Získání tweetů konkrétního uživatele
- Příloha 8: Vytvoření časoprostorové kostky v prostředí Processing
- Příloha 9: Animace časoprostorové kostky v prostředí Processing (na CD)
- Příloha 10: Získání tweetů prostřednictvím Twitter Streaming API (na CD)
- Příloha 11: Twitter power: a tale of lost camera (na CD)
- Příloha 12: Stažení dat o *venues* z Foursquare API (na CD)
- Příloha 13: Stažení dat z Instagram API (na CD)

Příloha 1: Struktura PostgreSQL tabulky pro import osobního archivu ze sítě Twitter

```
CREATE TABLE tweets
(
  tweet_id varchar,
  in_reply_to_status_id varchar,
  in_reply_to_user_id varchar,
  timestamp varchar,
  source varchar,
  tweet text,
  retweeted_status_id varchar,
  retweeted_status_user_id varchar,
  retweeted_status_timestamp varchar,
  expanded_urls varchar
);
```

Příloha 2: Struktura PostgreSQL tabulky pro import záznamů Geonames

```
CREATE TABLE geonames (
  geonameid varchar(200),
  name varchar(200),
  asciiname varchar(200),
  alternatenames varchar(5000),
  latitude decimal,
  longitude decimal,
  feature_class char(1),
  feature_code varchar(10),
  country_code char(2),
  cc2 varchar(60),
  admin1_code varchar(20),
  admin2_code varchar(80),
  admin3_code varchar(20),
  admin4_code varchar(20),
  population bigint,
  elevation integer,
  dem integer,
  timezone varchar(40),
  modification date
);
```

Příloha 3: Jazykové nastavení fulltextového vyhledávání v PostgreSQL

Vytvoří český slovník ze souborů czech.dict, czech.affix, czech.stop umístěných v /usr/share/postgresql/9.1/tsearch_data.

```
CREATE TEXT SEARCH DICTIONARY czech_ispell (
    TEMPLATE = ispell,
    DictFile = czech,
    AffFile = czech,
    StopWords = czech
)
```

Vytvoří českou konfiguraci hledání s parametry přejatými z anglického nastavení.

```
CREATE TEXT SEARCH CONFIGURATION czech ( COPY = pg_catalog.english );
```

Nastaví mapování vybraných typů na český slovník.

```
ALTER TEXT SEARCH CONFIGURATION czech
    ALTER MAPPING FOR asciiword, asciihword, hword_asciipart,
                      word, hword, hword_part
    WITH czech_ispell;
```

Nastaví mapování vybraných typů na anglický slovník.

```
ALTER TEXT SEARCH CONFIGURATION czech
    ALTER MAPPING FOR asciiword, asciihword, hword_asciipart,
                      word, hword, hword_part
    WITH english;
```

Příloha 4: Porovnání výsledků fulltextového vyhledávání v databázi GeoNames při různých jazykových nastaveních

Příloha 7: Získání tweetů konkrétního uživatele

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

import locale
import simplejson as json
from time import mktime
from time import strptime
import tweepy


#http://willsimm.co.uk/saving-tweepy-output-to-mongodb/
# Makes json response accesible as <response>.json
@classmethod
def parse(cls, api, raw):
  status = cls.first_parse(api, raw)
  setattr(status, 'json', json.dumps(raw))
  return status

def string_to_timestamp(created_at):
  """Return the timestamp from created_at object."""
  locale.setlocale(locale.LC_TIME, 'en_US.utf8')
  created_at = created_at.split(' ')
  created_at[1] = str(strptime(created_at[1], '%b').tm_mon)
  timestamp = strptime(' '.join(created_at[i] for i in [1,2,3,5]), '%m %d %H:%M:%S %Y') # returns Month Day Time Year
  return mktime(timestamp)

tweepy.models.Status.first_parse = tweepy.models.Status.parse
tweepy.models.Status.parse = parse

consumer_key = 'xxxxxxxxxxxxxxxxxxxxx'
consumer_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth, monitor_rate_limit=True, wait_on_rate_limit=True)

tweets = []
count = 1

for page in tweepy.Cursor(api.user_timeline, id="Roman86_K").pages():
  for tweet in page:
    if tweet.geo is not None:
      print 'Found a tweet number ' + str(count) + ' created at ' + str(tweet.created_at) + '.'
      t = json.loads(tweet.json)
      t['timestamp'] = string_to_timestamp(t['created_at'])
      tweets.append(json.dumps(t))
      count += 1
```

```
51
52 if tweets:
53   print 'Saving tweets...'
54   with open('tweets.json', 'w') as f:
55     tweets.reverse()
56     f.write('[')
57     f.write(',\n'.join(tweets))
58     f.write(']')
59   print 'Tweets saved.'
```

Příloha 8: Vytvoření časoprostorové kostky v prostředí Processing

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

from peasy import PeasyCam
import json

basemap = None
tweets = []
angle = 0

def setup():
    global basemap
    global tweets

    size(1010, 605, P3D)

    data = loadJSONArray('/home/michal/Dropbox/DP/prilohy/twitter/tweets.json')
    count = data.size()

    last = data.getJSONObject(data.size()-1).getFloat('timestamp')
    first = data.getJSONObject(0).getFloat('timestamp')

    for i in range(0, count):
        lon = data.getJSONObject(i).getJSONObject('coordinates')
              .getJSONArray('coordinates')
                .getFloat(0)
        lat = data.getJSONObject(i).getJSONObject('coordinates')
              .getJSONArray('coordinates')
                .getFloat(1)
        time = data.getJSONObject(i).getFloat('timestamp')

        x = map(lat, 16.5997195021086964, 63.68835804244784526, 0, height)
        y = map(lon, -19.68624620368202116, 58.92453879754536672, 0, width)
        z = map(time, first, last, 0, 500)

        tweets.append({'x': x, 'y': y, 'z': z})

    basemap = loadImage('basemap.png')

    cam = PeasyCam(this,53,100,-25,700)
    cam.setMinimumDistance(1)
    cam.setMaximumDistance(1500)

def draw():
    global basemap
    global tweets
    global angle

    background(0)

```

```
51      # Uncomment to rotate the cube
52      """if angle < 360:
53          rotateY(radians(angle))
54          angle += 1
55      else:
56          angle = 360 - angle"""
57
58      # box definition
59      stroke(150,150,150)
60      strokeWeight(.5)
61      noFill()
62      box(1010,500,605)
63
64
65      # basemap definition
66      translate(-505,250,-302.5)
67      rotateX(HALF_PI)
68      image(basemap,0,0)
69
70      for i in range(0, len(tweets)):
71          strokeWeight(.5)
72          stroke(255,255,255)
73          line(tweets[i].get('y'), height-tweets[i].get('x'), tweets[i].get('z'), ↩
                tweets[i].get('y'), height-tweets[i].get('x'), 0)
74
75          strokeWeight(5)
76          stroke(255,0,0)
77          point(tweets[i].get('y'), height-tweets[i].get('x'), tweets[i].get('z'))
78
79          strokeWeight(2)
80          stroke(255,255,255)
81          point(tweets[i].get('y'), height-tweets[i].get('x'), 0)
82          lrp = map(i, 0, len(tweets), 0, 1)
83          frm = color(255,0,0)
84          to = color(0,0,255)
85          if i < len(tweets)-1:
86              strokeWeight(1)
87              stroke(lerpColor(frm,to,lrp))
88              line(tweets[i].get('y'), height-tweets[i].get('x'), ↩
                    tweets[i].get('z'), tweets[i+1].get('y'), ↩
                    height-tweets[i+1].get('x'), tweets[i+1].get('z'))
89
90      # Uncomment to capture the screens
91      """if frameCount > 360:
92          noLoop()
93      else:
94          saveFrame('screens/frame-####.png')"""
```

Příloha 12: Stažení dat o *venues* z Foursquare API (na CD)

```
#!/usr/bin/python
# -*- coding: utf-8 -*-

import datetime
import foursquare

client_id    = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
client_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
latlon     = '49.2,16.616667' # point to look for venues around
radius     = '25000' # biggest distance of venue from the latlon
categories   = ['food', 'drinks', 'coffee', 'shops', 'arts', 'outdoors']

client  = foursquare.Foursquare(client_id=client_id, client_secret=client_secret)

for cat in categories:

  response = client.venues.explore(params={
    'll': latlon,
    'section': cat,
    'radius': radius,
    'v': '20131016',
    'intent': 'browse',
    'limit': '50'
  })

  result  = []
  filename = datetime.datetime.now().strftime("%Y_%m_%d-%H_%M") # 2014_02_16-12_02
  count   = 0

  groups = response['groups']
  for g in groups:
    items = g.get('items')
    for i in items:
      temp = {}
      temp['name']    = i['venue']['name']
      temp['lat']     = i['venue']['location']['lat']
      temp['lon']     = i['venue']['location']['lng']
      temp['checkins'] = i['venue']['stats']['checkinsCount']
      temp['users']   = i['venue']['stats']['usersCount']
      temp['tips']    = i['venue']['stats']['tipCount']
      temp['likes']   = i['venue']['likes']['count']
      temp['herenow'] = i['venue']['hereNow']['count']
      result.append(temp)

  # did we get any results?
  if result:
    with open(filename + '_' + cat + '.csv', 'w') as f:
      for r in result:
        for key in r:
          if count == 0:
```

```
51            f.write(key + ';')
52          else:
53            if isinstance(r[key], unicode):
54              r[key] = r[key].encode('utf-8')
55
56            f.write( str( r[key] ) + ';' )
57        f.write('\n')
58        count += 1
```