

# Časopriestorová Analýza Dát zo Sociálnych Sietí vo Vzt'ahu k Verejnej Doprave

Autor: **Martin Zajac**

Školiteľ: **Jiří Horák**

VŠB-TU Ostrava

Katedra geoinformatiky

[martin.zajac@vsb.cz](mailto:martin.zajac@vsb.cz)

# Motivácia

- Denne publikované nespočetné množstvo príspevkov
- Okamžité reakcie na udalosti
- Tweets o doprave – analýza distribúcie tweetov v čase a priestore, témy spojené s dopravou
- Sentiment – tweetov o doprave
- Dáta zo sociálnych sietí – veľký zdroj voľne dostupných dát (Každá platforma – vlastná API)
- Pre ucelený obraz – problémy:
  - Reklamy, nadbytočný obsah
  - Obtiažné dohľadanie kontextu príspevku
  - Problémy s lokalizáciou príspevku

# Sociální sítě



**Martin** @koumesak · 26. 7.  
Proč nemůže být zase chladno. To cestování #mhdpraha je za trest v busech bez klima. @DPPOficialni kdy se to zlepší?

**Sab** @sabie\_sab  
Davové šílenství v pražské MHD, řidič už ani neotvírá dveře :D #kabáti #koncert #MHD #praha #tram25praskne @twipraha

**Víctor Pérez** @MartinAlteman · 1 h  
Stuck in the Met line once again. @TfL do you have any plan to fix it?

**Tomáš** @zasevtom · 19. 7.  
venku je 18 stupňů, klimatizace v buse jede na plný výkon a fouká studený vzduch... nedělají to řidiči schválně? #sameocasky #mhdpraha

**Lucie Sajko** @LucinkaDiv  
Opravdu dekuji #dpp, ze prijel alespon jeden autobus za pul hodiny. Neni nad pohodovou cestu do prace. A jeste k tomu v pondeli. #mhd #praha  
8:59 dop. · 1. 9. 2014 · Twitter for iPhone

**Thérèse** @TiggerTherese · 4 m  
Odpověď uživatelům @harriekd @TfL a @TfLBusAlerts  
I don't know what they think they're doing with the buses. Almost every one I've been on in the last week has had COLD air blasting out of the vents. :<

# Twitter v číslech

- Aktívny užívatelia v 2020:
  - 330 mil. svet (<https://www.websitehostingrating.com/twitter-statistics/>)
  - 400 tis. Česká republika (<https://www.lupa.cz/aktuality/tiktok-v-cesku-pouziva-1-1-milionu-aktivnich-uzivatelu-mesicne/>)
  - 16,45 mil. Spojené Královstvo (<https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>)
- Ľudia sa vyjadrujú pomocou tweetov - krátka správa s fixnou dĺžkou (280 znakov)

# Zber dát

- Zmienky dopravných podnikov Londýna (@TfL), Prahy (@DPPOficialni) a Ostravy (@DPOstrava)
- V období od 26.3.2020 do 31.1.2021
- Počty cestujúcich na staniciach – štatistiky od TfL



# Zber dát

- Twitter Standard API v.1 (vývojársky účet; 1% tweetov; 7-9 dní spätne)
- Knižnica *rtweet* pre R
- Riešenie *GetOldTweets* pre Python (simuluje prezeranie webu; možno stiahnuť tweety staršie ako 7 dní; nefungovalo spoľahlivo)
- Londýn → 545 295, Praha → 4951, Ostrava → 257





# Je 1% tweetov dost' ?

- Morstatter, 2013 – Dokáže Twitter Standard API vhodne reprezentovat' situáciu na Twitteri?
- Twitter Firehose (100% tweetov) vs Twitter Standard API (1% tweetov)
- Územie Sýrie, zoznam hashtagov (#syria, #assad, #hama, #damascus, #daraa)
- Standard API – nie je ovplyvnená užívateľmi s vysokou aktivitou
- Pri malých vzorkách môže Standard API zavádzať
- **Twitter Standard API – všetky lokalizované tweety (napriek tomu len ~1%)**

# Postup spracovania

- Začistenie tweetov (dplyr, tidytext – R)
- Testovanie možností lokalizácie
- Geokódovanie
- Sentiment analýza
- Semantic analýza
- LDA modelovanie tém (topicmodel – R)



# Začistenie tweetov

- Transformácia do ASCII a malých písmen
- Odstránenie zmienok účtov, heštegov, interpunkcie
- Odstránenie URL odkazov
- Z pôvodných 545 295 → 344 062 (-37%)

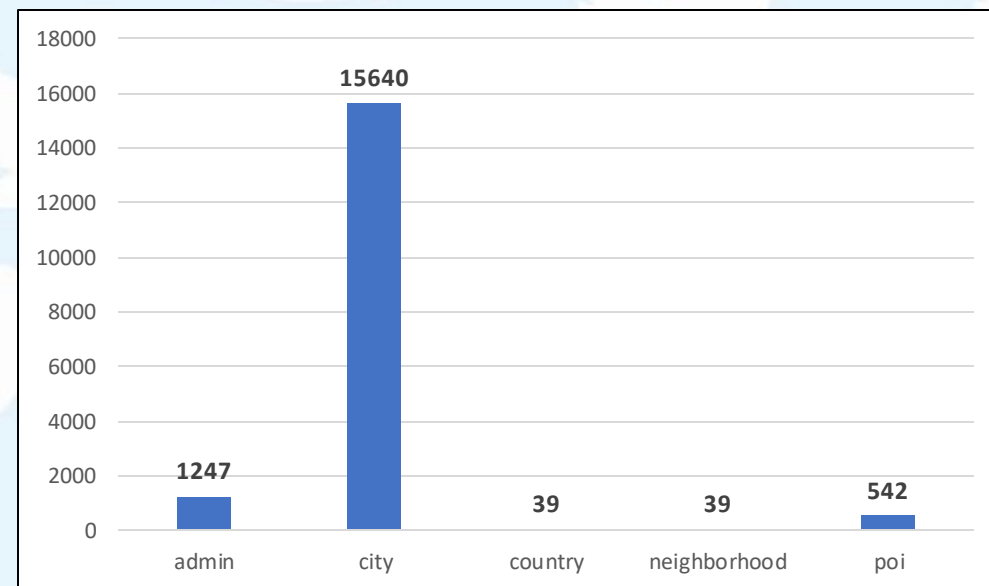
@SouthernRailUK @nationalrailenq @TfL  
@TLRailUK I don't really need further  
details. Commuting through Victoria every  
day I see platform and ticket barrier staff and  
neither have face coverings. You won't get  
commuters to comply if staff don't lead by  
example. <https://t.co/WJhJEzotID>



i dont really need further details commuting  
through victoria every day i see platform and  
ticket barrier staff and neither have face  
coverings you wont get commuters to  
comply if staff dont lead by example

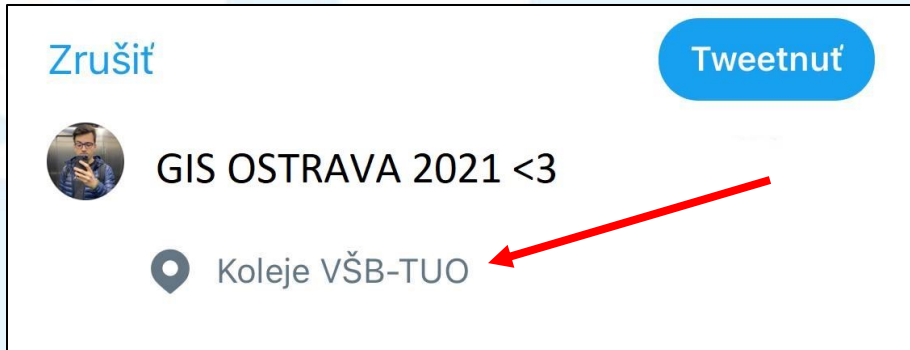
# Metódy lokalizácie

- Testovanie rôznych prístupov lokalizácie z metadát tweetu
- Súradnice tweetu → iba 0,0001% - funkcia vypnutá od 6/2019
- Twitter place → 3% zo vzorky, takmer všetko len na úrovni mesta
- Vypovedá viac o polohe užívateľa ako o polohe sťažnosti, emócie, problému



# Metódy lokalizácie

- Polohu určí manuálne užívateľ – zoznam od forsquare
- Vzniká tu neistota – polohu možno určiť kdekoľvek



place_url	place_name	place_full_name	place_type	country	country_code
https://api.twitter.com/1.1/ge	Koleje VŠB-TUO	Koleje VŠB-TUO	poi	Česká republika	CZ

Odstrániť

Označiť polohu

Hotovo

Prehľadávajte lokality

Koleje VŠB-TUO

Studentská 1770/1 · 20 m



Hotel Garni VŠB

Vrtule Club

Studentská 1770/1 · 90 m

Erasmus Student Network VŠB-TUO

Studentská 1770/1 · 60 m

Posilovna

Studentská 1770/1 · 250 m

Pizzerie na kolejiách

Studentská 1770/1 · 200 m

klaud

Studentská 1a · 150 m

Bufet Vrtule (č.9)

Studentská 1770/1 · 100 m

Kolejní tělocvična

Studentská 1770/1 · 60 m

T2 Club & Steak house

Studentská 1770/1 · 70 m

Ostrava, Česká republika

# Geokódovanie

- Text tweetu
- V texte častá zmienka názvu stanice, linky, spoja
- Problém zo skratkami, slangom, skomoleninami
- V Londýne:
  - 7042 – na úrovni staníc (cca 1%) – len 10 najrušnejších
  - 10493 – na úrovni linky metra (cca 2%)
- V Prahe:
  - 62 – na úrovni staníc (cca 1%)
- V Ostrave:
  - 12 – na úrovni staníc (cca 4%) – manuálne z 257 tweetov

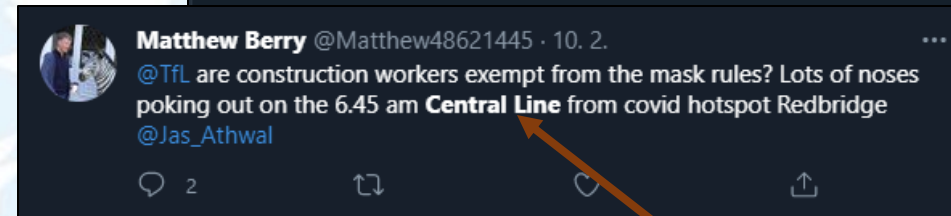


Odpoveď užívateľi @Miss\_Sis\_

Hi there, sorry for the disruption to your journey. Currently, there's no service Kings Cross to Acton Town while emergency services deal with a casualty on the track. Severe delays on the rest of the line. For live updates please see here [tfl.gov.uk/tube-dlr-overg...](https://tfl.gov.uk/tube-dlr-overg...) Thanks, Morpheus

Přeložit Tweet

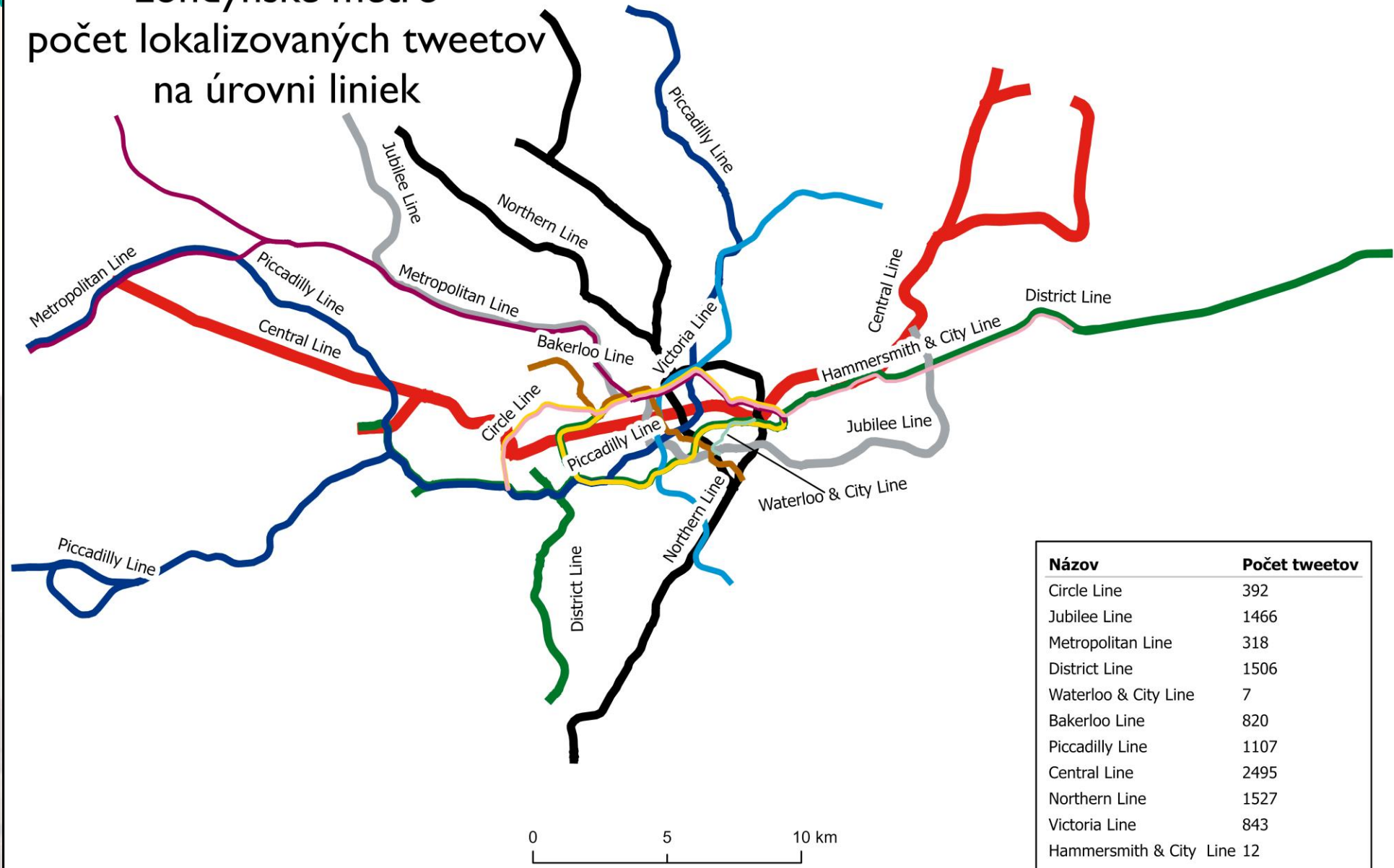
2:28 odp. · 25. 9. 2020 · Clarabridge Engage



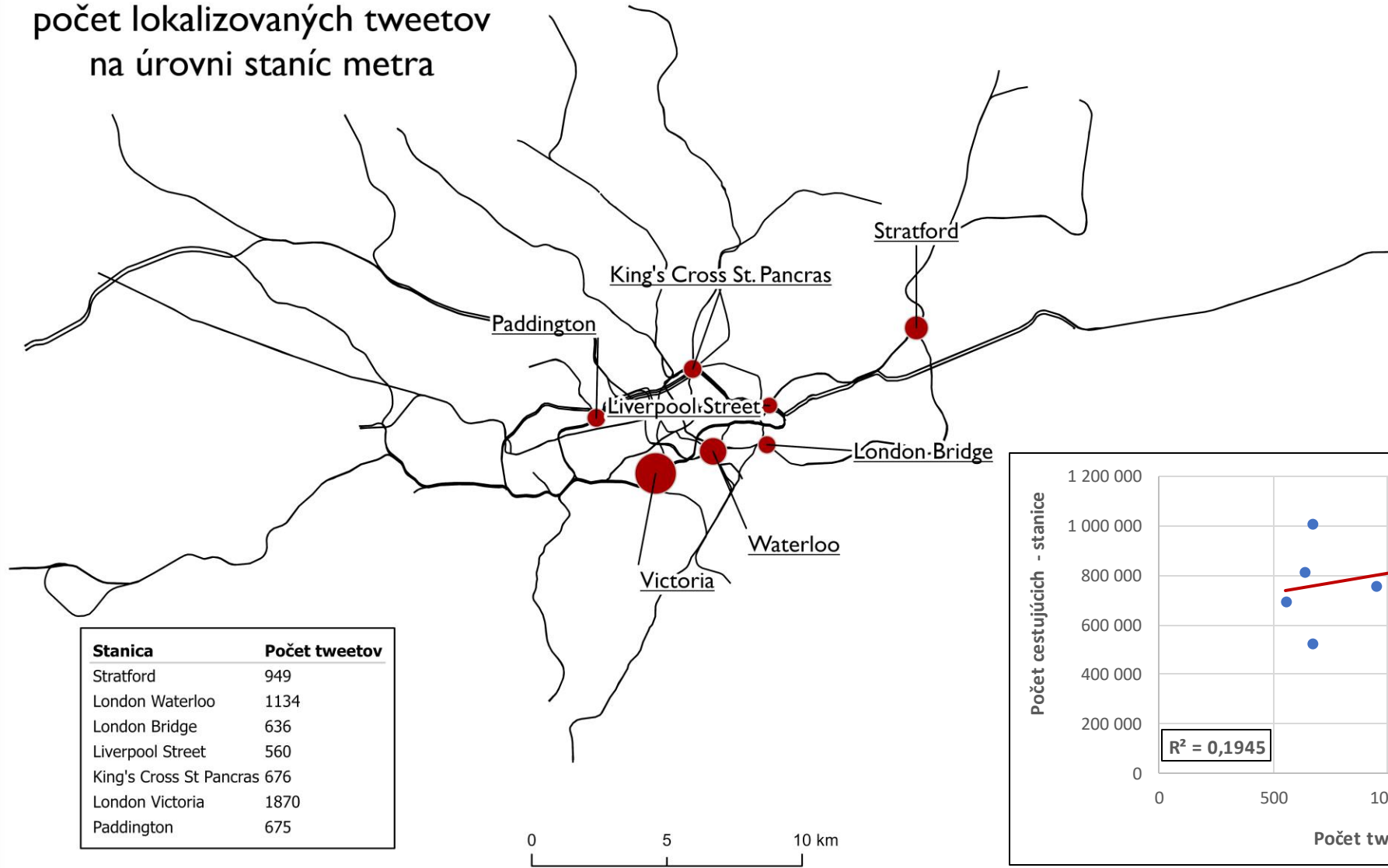


# Londýnske metro

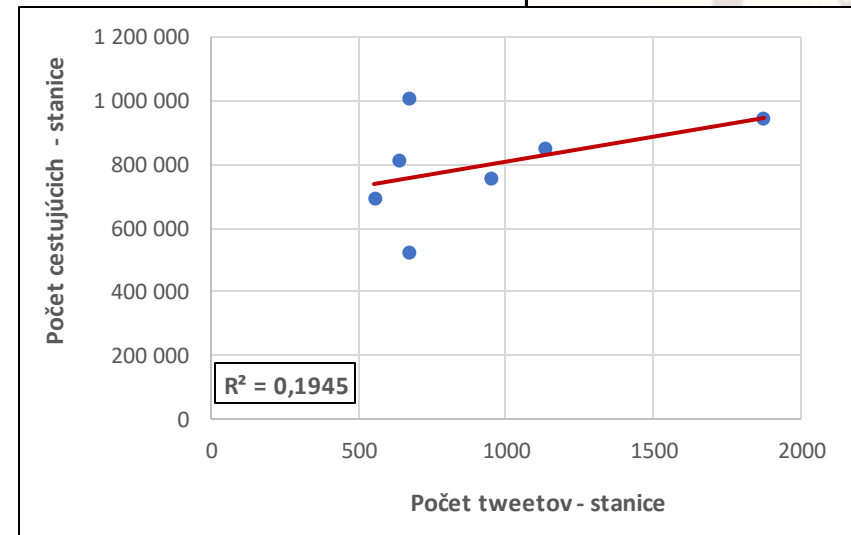
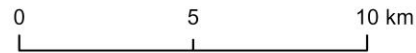
## počet lokalizovaných tweetov na úrovni liniek



## Londýnske metro počet lokalizovaných tweetov na úrovni staníc metra

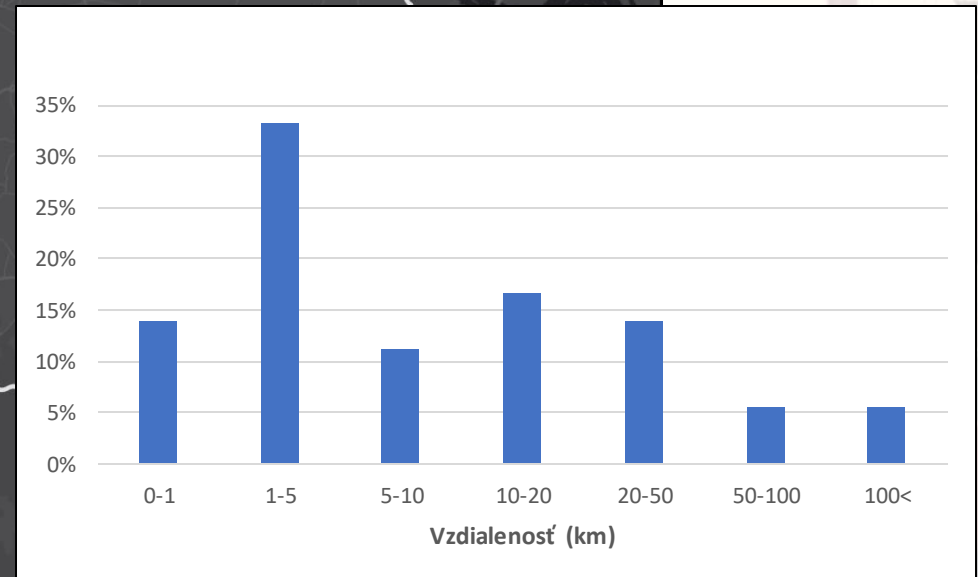
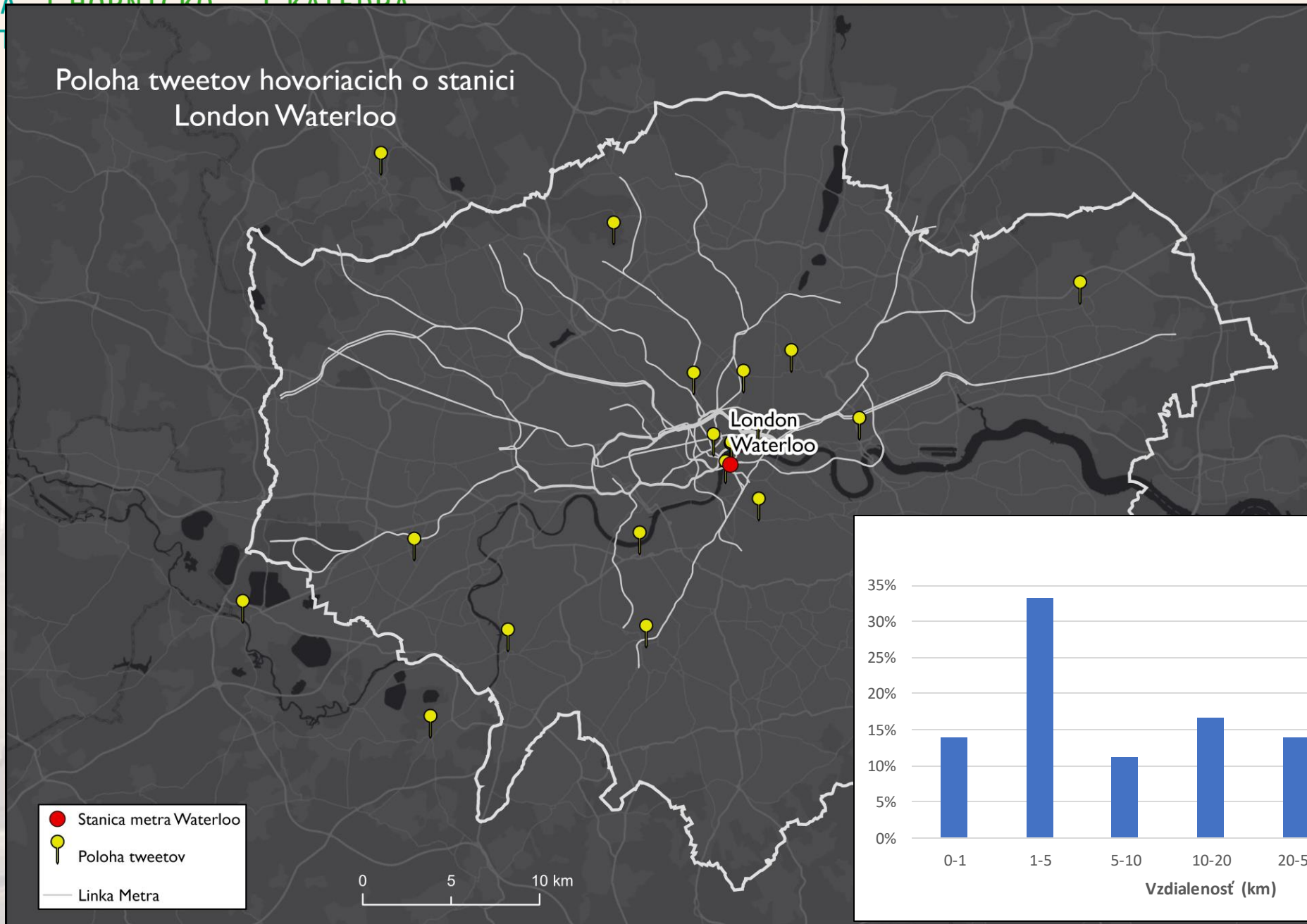


Stanica	Počet tweetov
Stratford	949
London Waterloo	1134
London Bridge	636
Liverpool Street	560
King's Cross St Pancras	676
London Victoria	1870
Paddington	675



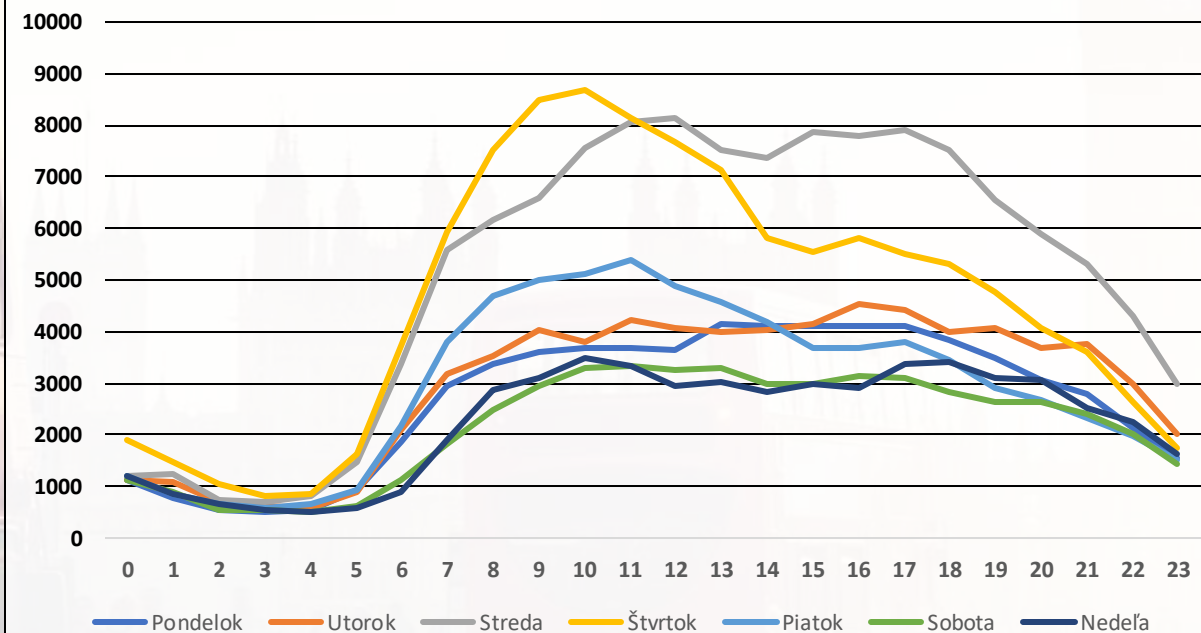


### Poloha tweetov hovoriacich o stanici London Waterloo

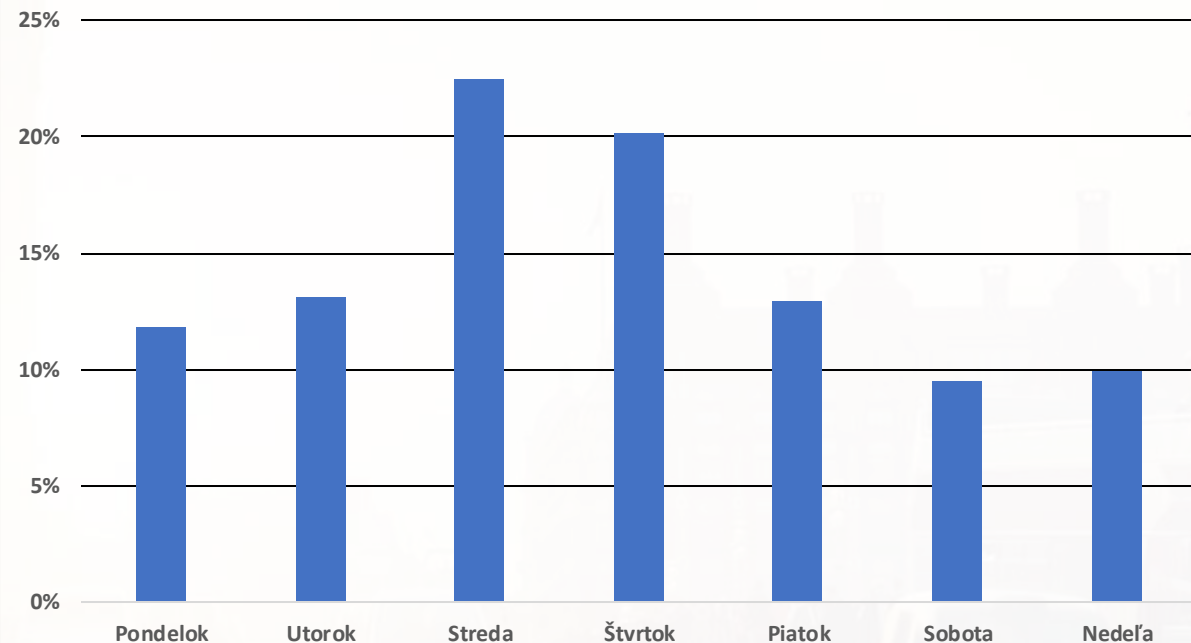


# Aktivita dopravného tweetovania

Dopravné tweetovanie v Londýne – v čase



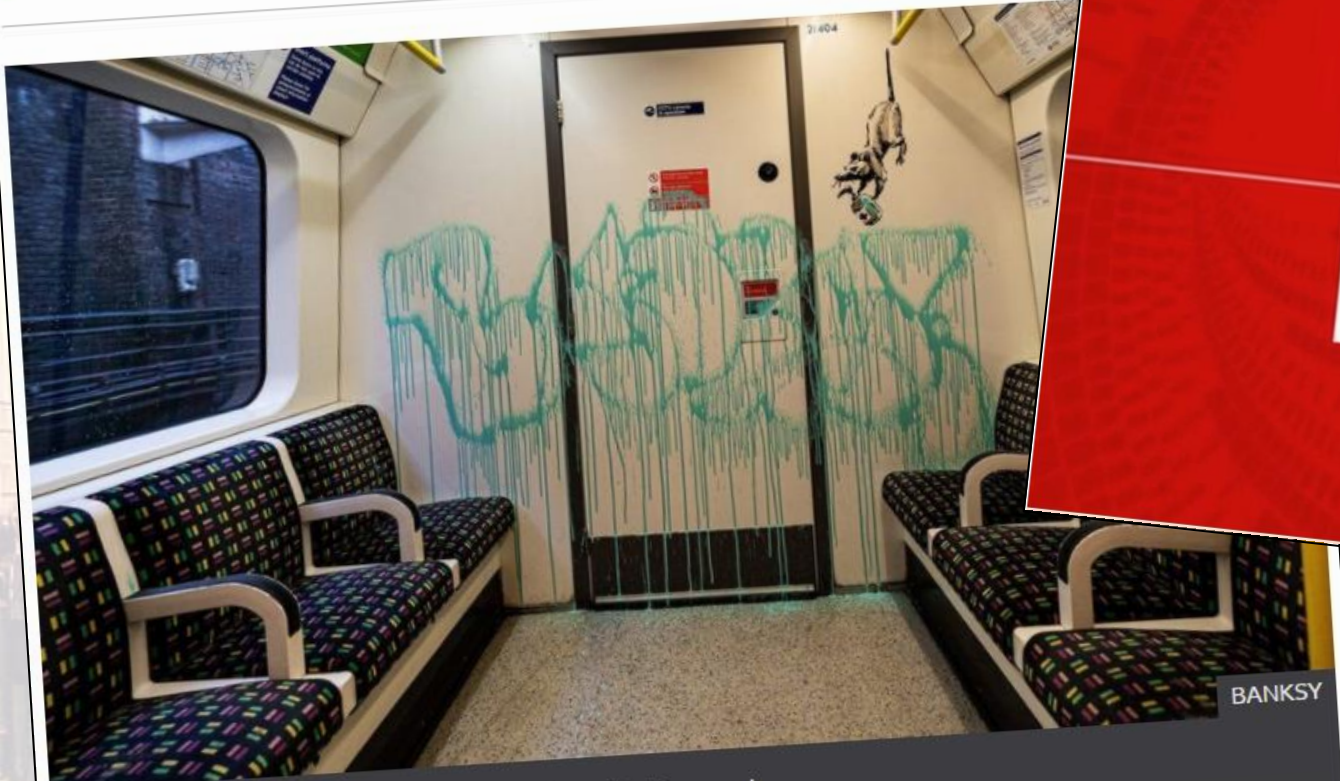
Dopravné tweetovanie v Londýne – v týždni





# Banksy Tube graffiti: Cleaners 'unaware it was by artist'

15 July 2020



Banksy spray painted his tag in the colours of a medical face mask

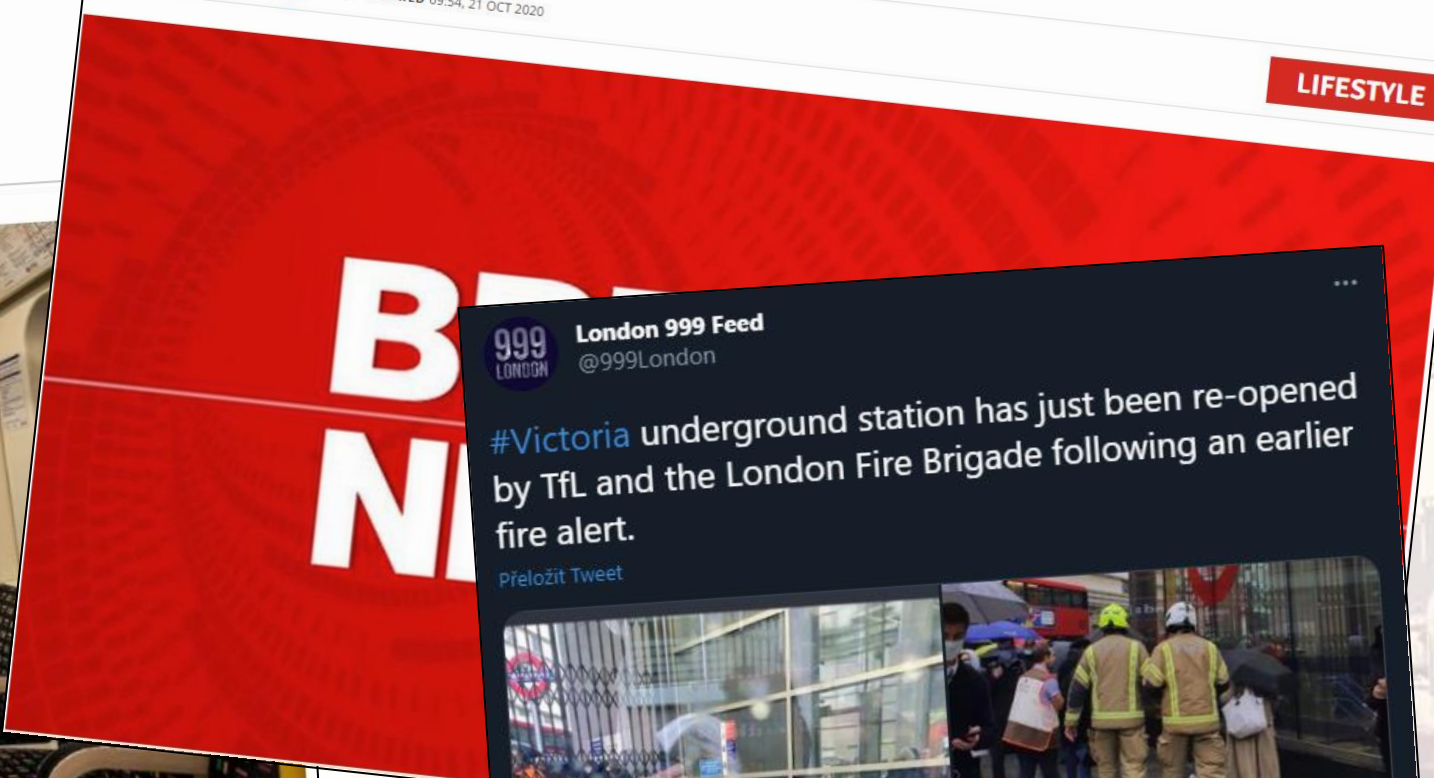
# Victoria station evacuated: Fire alert closed busy London Underground station

One witness advised people to avoid the London Underground



By [Ian Molyneux](#)  
UPDATED 09:54, 21 OCT 2020

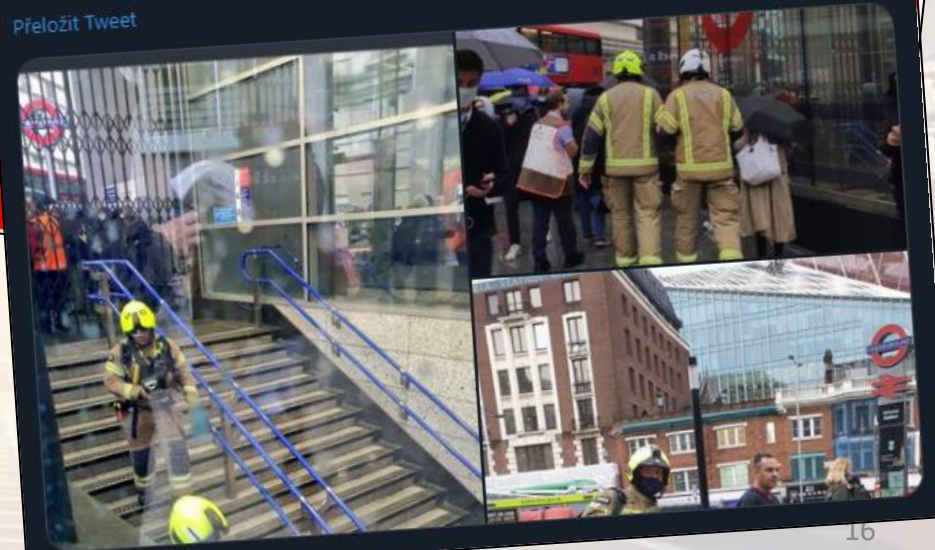
LIFESTYLE



999 LONDON London 999 Feed @999London

#Victoria underground station has just been re-opened by TfL and the London Fire Brigade following an earlier fire alert.

Přeložit Tweet

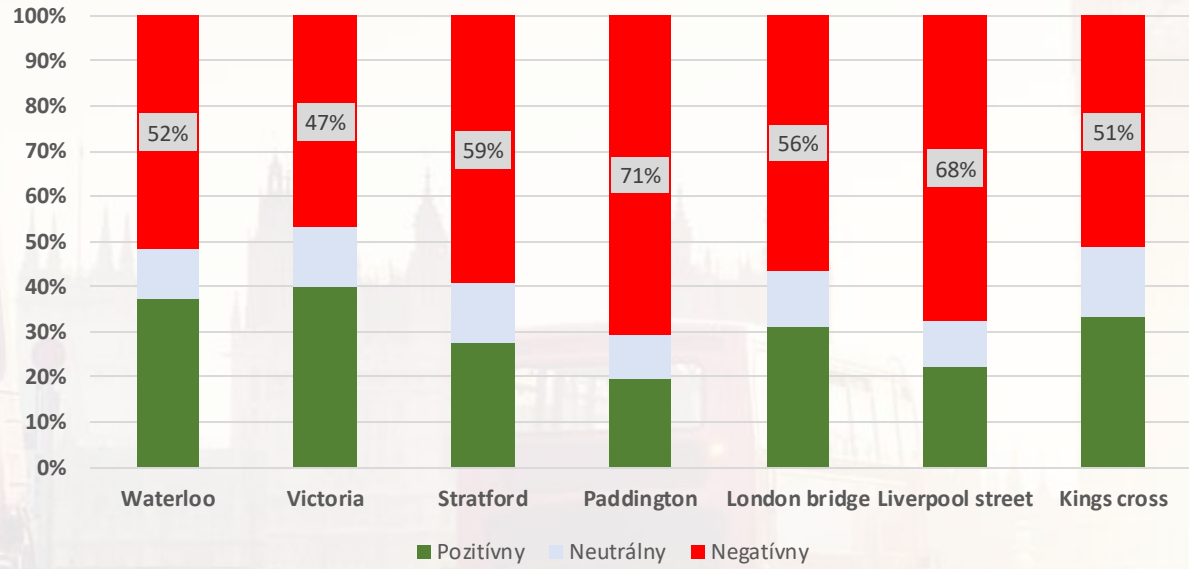


# Analýza pocitu

- Riešenie postavené na R
- Lexikón *BING* – slová kategorizované na pozitívne / negatívne
- Prevedené na hodnoty -1 / 1 → sčítanie sentimentu za celý tweet



Pocity na stanicích



Pocity v týždni



# Problémy na stanicích

- Len negatívne tweety
- Identifikované problémy v doprave → Porucha ; Pohodlie ; Covid-19 ; Preplnenosť ; Dochvilnosť
- Identifikované výrazy vo vzťahu k problémom

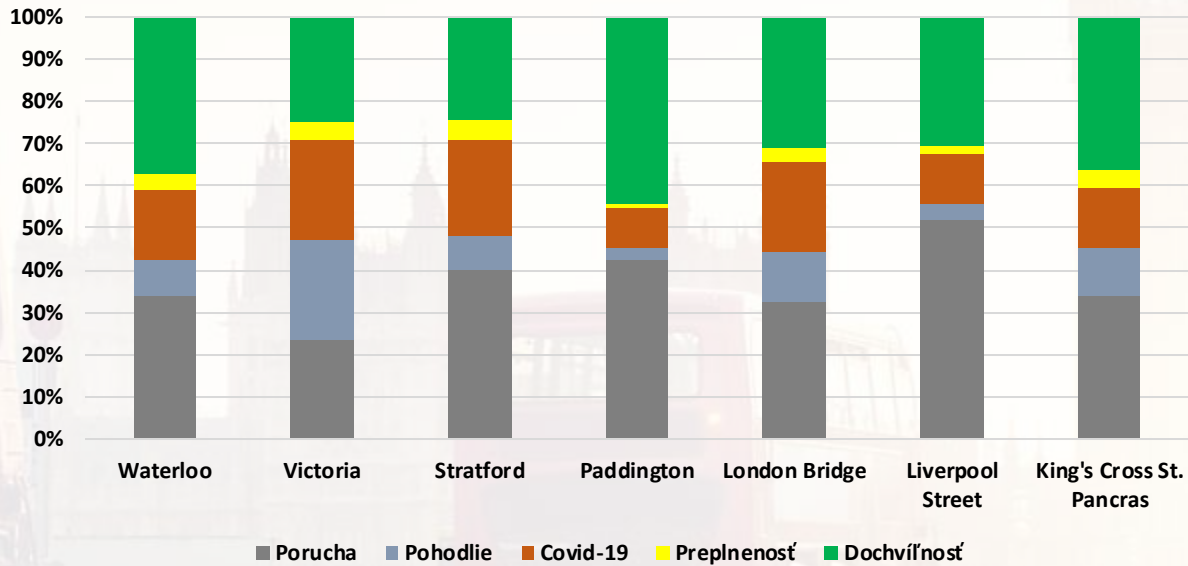
Dochvilnosť
rush
stuck
waiting
delayed
delays
slow
frequency
waiting
wait
time
late
moment
ontime
onschedule

Porucha
faulty
fault
paralyzed
collapse
collapsing
failure
failures
closed
evacuate
blocked
service
interrupted
interruption
suspended

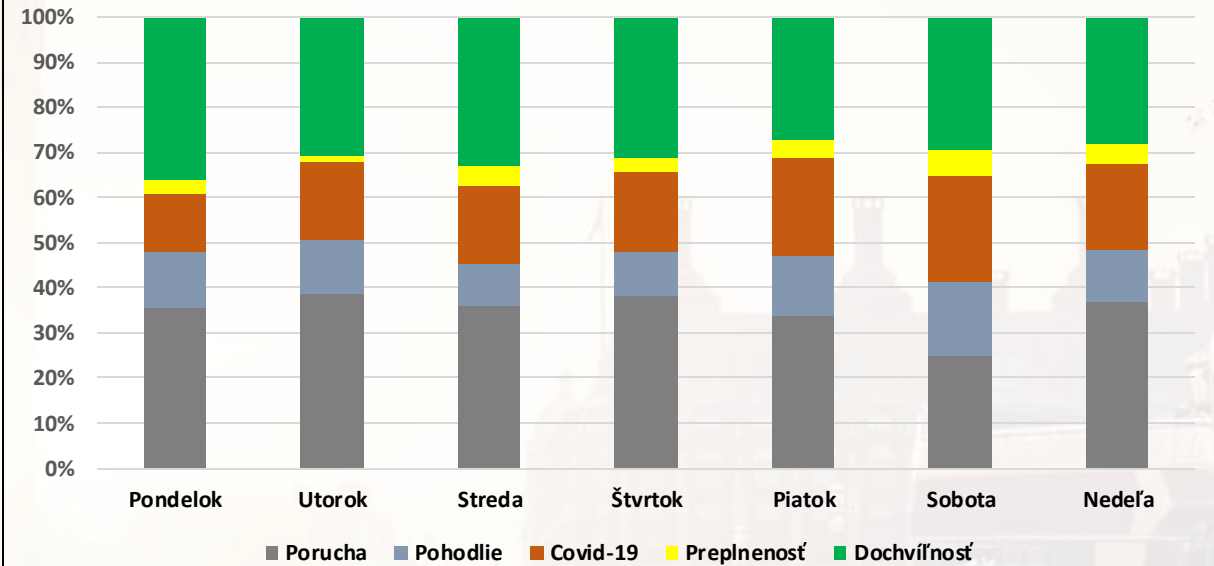
Komfort
ventilation
hot
dirty
smell
homeless
beggars
safety
disgusting
heat
cold
sauna
sweat
sweating
ventilated



Problémy na stanicích



Problémy v týždni



# Modelovanie tém - LDA

- Neriadený algoritmus
- Vstúpe → tweety ; výstup → sú množiny slov reprezentujúce témy v tweetoch
- Iný prístup k hľadaniu problémov v doprave

# Modelovanie tém – LDA

Téma	Výrazy	Problém
1	travel, card, oyster, updates, information, website, journey	Oyster card – platobná karta pre MHD
2	service, platforms, closed, northern, closed, entrance, falut, alert	Porucha
3	delay, minutes, time, running, late	Dochvílnosť
4	wearing, masks, face, wear, distancing, covering, covid19	Covid-19

	Výrazy	Problém
Victoria	belly, mujinga, spat, worker, face, incident, justice	Incident s Belly Mujinga
King's Cross	harry, platform, potter, hogwarts, platform, 3/4	Harry Potter 😊
Bakerloo line	bakerloo, extension, project, mayor, lewisham,	Predĺženie Bakerloo linky

# Záver

- Dáta zo sociálnych sietí ako zdroj dát – veľa nástrah a limitácií
- Interpretovateľné informácie len pre veľké mestá - dostatočná aktivita na twitteri
- Potreba premyslieť prístup k lokalizácii - príspevok / text
- Dôležité - identifikovanie sentimentu a tém v dopravných tweetoch
- Následné cielené vylepšovanie situácie
- Výzvy do budúcnosti:
  - Využitie twitter API v.2 – nové funkcie
  - Rozlíšenie jednotlivých spôsobov hromadnej dopravy (Metro, DLR, Bus, Lod', Vlak, Električka) a vzájomné porovnanie
  - Identifikácia pravidelných „tweetujúcich“ ich kategorizácia



# Ďakujem za pozornosť

*„Nie som závislý na sociálnych sieťach, ja s nimi len nehodlám prestať.“*

**Martin Zajac**

[martin.zajac.st@vsb.cz](mailto:martin.zajac.st@vsb.cz)

[gis.vsb.cz](http://gis.vsb.cz)

[@matozayo](#) | [@katedra\\_geoinformatiky](#)