

Akademia Górniczo-Hutnicza

Eksploracja danych

Analiza fake news na podstawie
tweetów o wyborach
parlamentarnych w Polsce

grupa środa 12:50

inż. Piotr Rząsa, inż. Michał Kawałek

Spis treści

- 1.Cel projektu
- 2.Wybór technologii
- 3.Dane
- 4.Przebieg projektu
- 5.Podsumowanie, wnioski
- 6.Literatura

1. Cel projektu

Celem projektu była analiza tzw. 'fake news' w kontekście wyborów parlamentarnych w Polsce w roku 2018. Miała ona zostać przeprowadzona przy użyciu istniejących już rozwiązań do przetwarzania tego typu danych lub własnego stworzonego przez studentów mechanizmu w oparciu o artykuły naukowe. Dane miały pochodzić z serwisu społecznościowego Twitter i zostać przez studentów samodzielnie zgromadzone w trakcie trwania projektu.

2. Wybór technologii

Z racji popularności w tematyce 'data analysis' jako główne narzędzie do wykonania projektu w kontekście przetwarzania danych został wybrany język programowania Python.

Dodatkowo użyte biblioteki:

- psycopg2 do połączenia się z bazą postgreSQL
<http://initd.org/psycopg/>
- matplotlib do tworzenia wykresów
<https://matplotlib.org>
- tweepy do pobierania danych z twittera
<http://www.tweepy.org/>

W kontekście analizy danych wykorzystaliśmy również program Gephi do wizualizacji grafów i zależności.

<https://gephi.org/>

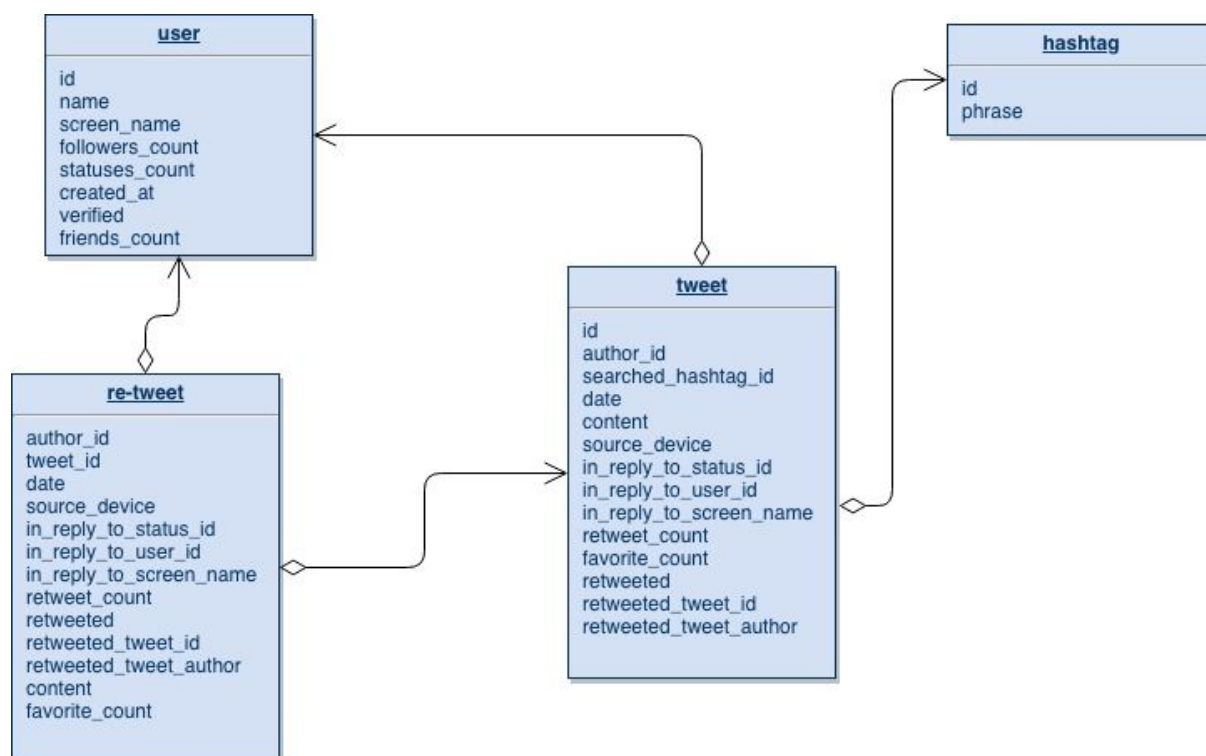
3. Dane

Używając API Twittera do pobierania interesujących nas danych wykonaliśmy następujący preprocessing:

- ustawienie odpowiedniego kodowania - twitter zwraca dane w kodowaniu, które nie wspiera polskich znaków dlatego aby obsłużyć nasz przypadek (polskie wybory) musieliśmy zmienić kodowanie na bardziej dokładne (utf-8).
- w trakcie trwania pracy nie mieliśmy wszystkich niezbędnych parametrów, np. kont czy też samych tweetów. Dlatego często musieliśmy często dociągać potrzebne nam atrybuty.
- tweeter zwracał daty w formacie, który nie był dla nas czytelny więc aby ułatwić sobie analizę skonwertowaliśmy daty do prostszego formatu.
- twitter w dziwny sposób zwraca retweety:
 - raz jest to prefix RT @ przed samym tweetem
 - raz jest to flaga 'retweeted' w JSON'ie zwracanym przez API
 - raz jest to obiekt 'retweet_status' w JSON'ie zwracanym przez API

Dlatego na każdą reprezentację retweetu reagowaliśmy i wstawialiśmy o tym informację do bazy danych

Dane zebrane z portalu Twitter były zapisywane do lokalnie postawionej bazy PostgreSQL przedstawionej na Rysunku 1.



Rysunek 1. - Schemat bazy danych

Baza danych zawiera 4 tabele.

W tabeli 'user' znalazły się informacje na temat użytkowników serwisu, których tweety lub retweety pobraliśmy. Pola tej tabeli nazwami w większości odpowiadają realnym atrybutom pobieranym z Twittera:

- id (bigint) - id twitterowe
- name (varchar) - login użytkownika
- screen_name (varchar) - wyświetlana nazwa użytkownika
- followers_count (int) - liczba śledzących
- statuses_count - liczba wszystkich statusów dodanych przez tego użytkownika
- created_at (date) - data stworzenia konta
- verified (bool) - czy konto jest zweryfikowane przez administrację Twittera (dot. kont znanych organizacji, firm i osób publicznych)
- friends_count (int) - liczba osób obserwowanych przez użytkownika

Tabela 'hashtag' zawiera jedynie hashtag w polu phrase (varchar) i jego id (bigint).

Ostatecznie ta tabela nie była niezbędna i wystarczyłoby w tabeli 'tweet' umieścić treść hashtaga w jednym z pól.

Tabela 'tweet' i 're tweet' są niemal bliźniacze i zawierają odpowiednio tweety, pobierane po hashtagach oraz retweety pobierane również po hashtagach (w treści był cały tweet, który się podawało dalej, więc z tego powodu wyszukiwarka również brała je pod uwagę). Pola w tych tabelach były następujące:

- id (bigint) - id obiektu
- author_id (bigint) - id użytkownika
- date (timestamp) - czas publikacji statusu
- content (varchar) - treść
- source_device - skąd został opublikowany (iPhone, Android, Web, itp)
- in_reply_to_status_id (bigint) - jeśli tweet jest odpowiedzią na jakiś inny tweet to tu jest jego id
- in_reply_to_user_id (bigint) - j.w. tylko id autora tamtego tweeta
- in_reply_to_screen_name (varchar) - j.w. tyle, że login autora
- retweet_count (int) - liczba retweetów
- retweeted (bool) - czy ten tweet był podany dalej
- retweeted_tweet_id (bigint) - jeśli to retweet to id podanego dalej tweeta
- retweeted_tweet_author (bigint) - jeśli to retweet to id autora podanego dalej tweeta
- favourite_count (int) - liczba polubień tweeta

Dane pochodzą z przedziału czasowego 21.10.2018r. - 06.11.2018r.

Lista przykładowych hashtagów:

- #wybory
- #Wybory2018
- #NowyPrezydentWiększeMożliwości
- #WspólnieTworzymyPrzyszłośćKrakowa
- #wyborysamorzadowe2018
- #WyborySamorzadowe

Niestety nie wszystkie partie oraz nie wszystkie hashtagi mogliśmy wykorzystać, niektóre były zbyt ogólne i nie dawały nam danych jakich chcieliśmy.

Ilość pobranych danych:

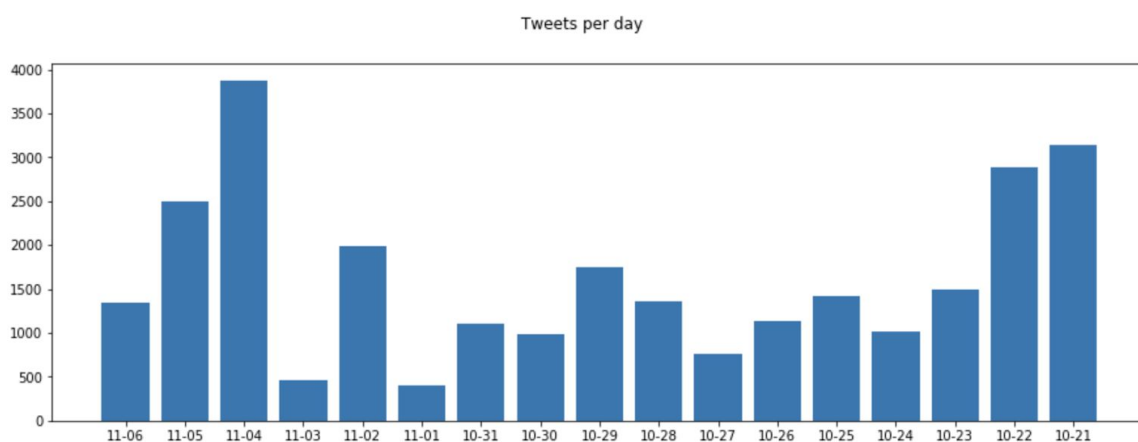
- 8882 autorów/użytkowników
- 31943 retweetów
- 4402 tweetów

4. Przebieg projektu

Etap 1 - pobranie danych

Pierwszym krokiem było pobieranie tweetów, dociągnięcie innych potrzebnych informacji z Twittera o autorach i odpowiedziach do tweetów. Następnie wszystkie te informacje umieszczane były w bazie postgreSQL przy pomocy odpowiednich skryptów.

Ponieważ Twitter pozwala na pobieranie danych po słowie kluczowym tylko do 10 dni wstecz, dane pobieraliśmy na raty. Nie wszystkie potrzebne dane zostały pobrane od razu. Część danych została dociągnięta po pewnym czasie, gdy okazywały się one potrzebne.



Rysunek 2. - Wykres liczby tweetów i retweetów w zależności od dnia

Etap 2 - analiza danych pod kątem 'fake news' i 'fake accounts'

Wątek 1 - Analiza 'fake accounts'

Przeszukując internet w poszukiwaniu metod do analizy fake news udało się znaleźć wiele informacji o tym, że fake news mogą być tworzone przez, np. różnego rodzaju boty lub konta założone specjalnie na potrzeby rozprzestrzeniania fake news.

Użyliśmy metryki, która na podstawie danych o koncie z Twittera miała powiedzieć czy dane konto jest 'fake account'. Sprawdzaliśmy czy konto spełnia pewne określone kryteria - jeżeli tak to mnożyliśmy metrykę przez liczbę podana przy tym kryterium. Lista tych kryteriów prezentuje się następująco:

Attributes	Weight
the account has at least 30 followers	0.53
the account has been geo-localized	0.85
it has been included in another user's favorites	0.85
it has used a hashtag in at least one tweet	0.96
it has logged into Twitter using an iPhone	0.917
a mention by twitter user	1
it has written at least 50 tweets	0.01
it has been included in another user's list	0.45
(2*number followers) _ (number of friends)	0.5
User have at least one Favorite list	0.17

Rysunek 3. - Wagi poszczególnych atrybutów na podstawie art. naukowego ^[5]

Na przykład konto z 40 followersami, które użyło co najmniej jednego hashtaga w co najmniej jednym tweecie będzie miało metrykę $1(\text{podstawa}) * 0.53 * 0.85 = 0,4505$. Konta, które mają metrykę mniejszą niż 0.5 uznawane są za 'fake account'.

Ta metryka pozwoliła wykryć nam kilka podejrzanych kont lecz jak wiemy - stuprocentowej pewności nigdy nie będziemy mieć.

Przykład: https://twitter.com/St_Janecki

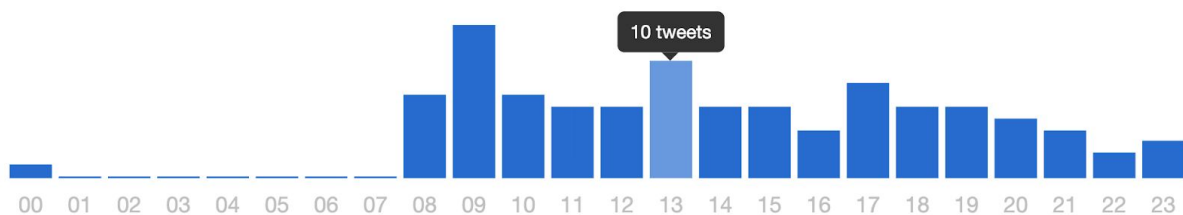
Replies	13	/ 100
Tweets with @mentions	13	/ 100
Tweets with #hashtags	0	/ 100
Retweets	0	/ 100 were retweets by @St_Janecki
Tweets with links	57	/ 100
Tweets with media	8	/ 100
Most linked domain	twitter.com , wpolityce.pl , www.youtube.com , t.co , warszawa.wyborcza.pl	
Followers ratio	281.90 followers per following	

Tabela 1. - Statystyki dla znalezionej fake konta

HAPPY TWEETS VS. SAD TWEETS



Rysunek 4. - Analiza sentymentu tweetów konta ST_Janecki



Wykres 1. - Rozkład tweetowania w ciągu dnia;
źródło powyższych analiz : https://foller.me/st_janecki

Spróbowaliśmy również znaleźć konta, które dodały post zaraz po założeniu tego konta - co jasno wskazałoby, że konto zostało stworzone tylko po to aby dodać dany post. Wyszukaliśmy wszystkie konta, które swój pierwszy post opublikowały nie dłużej niż dzień po rejestracji. Niestety łącznie takich kont było 0 co nie dało nam możliwości dalszej analizy w tym kierunku.

Wątek 2 - Analiza 'fake news'

Najpierw scharakteryzowaliśmy fake news:

- są najczęściej tworzone aby na czymś zyskać (np. finansowo)
- mają charakter opiniotwórczy czy też podburzający (np. aby czytelnicy przeczytali podlinkowany artykuł)
- są częściej re-tweet'owane niż prawdziwe tweety
- często rozprzestrzeniane przez fake konta
- często można znaleźć jakieś wskazówki w komentarzach do tweeta, jeżeli występują tam słowa takie jak np. 'nieprawda'
- rozwiązaniem może być przekazywanie tweetów do ekspertów, którzy stwierdzą czy jest on prawdą czy fałszem - jest to jednak rozwiązanie, które się nie skaluje
- możliwość użycia DBPedia do zweryfikowania prawdy - w przypadku wyborów niestety nie ma możliwości z tego skorzystać ale warto o tym wspomnieć

Podejście 1 do analizy 'fake news'

Następnie mając narzędzie do analizy fake accounts postanowiliśmy je wykorzystać do analizy fake news. Przeszukaliśmy konta, które znaleźliśmy używając powyższej metryki w celu:

- znalezienie fake accounts, które same tweetują fake news
- znalezienie fake accounts, które retweetują inne tweety, które są fake news'ami.

Ilość fake accounts jakie udało nam się znaleźć była na tyle mała, że mogliśmy to sprawdzić ręcznie.

- Krystia80467130
- JakubBroch
- PoczytalniaPL
- 36_osiedli
- adambryndza
- MilewskMateusz

Niestety nie udało nam się znaleźć żadnych takich zależności, możliwe, że z powodu zbyt małej ilości danych.

Podejście 2 do analizy 'fake news'

Drugim podejściem była analiza semantyczna każdego tweet'a i porównanie każdy z każdym w celu znalezienia podobnych ale np. sprzecznych tweetów.

Niestety algorytm^[10] nie sprawdzał się - każdy sposób analizy semantycznej tweetów kończył się niepowodzeniem i znalezione pary w ogóle nie były do siebie podobnie semantycznie. Dlatego postanowiliśmy odpuścić sobie tą metodą i pójść dalej.

Wykorzystane metody:

- bag of words
- obliczanie odległości między słowami używając bag of words zgodnie z metryką z artykułu naukowego^[10]

Przykłady zdań, które algorytm uznał za podobne semantycznie:

- *W trzech obwodowych komisjach w Warszawie frekwencja wyniosła 100 procent!*
- *W Warszawie nie będzie mostów. W Łodzi będzie komisarz. A w Legionowie? #wybory2018*
- *Warszawa kupi udziały PPL w Modlinie? #wybory2018*
- *"Dziękuję, że wybraliście przyjaciół". #Owsiak komentuje wyniki #wybory2018 w #Warszawa*

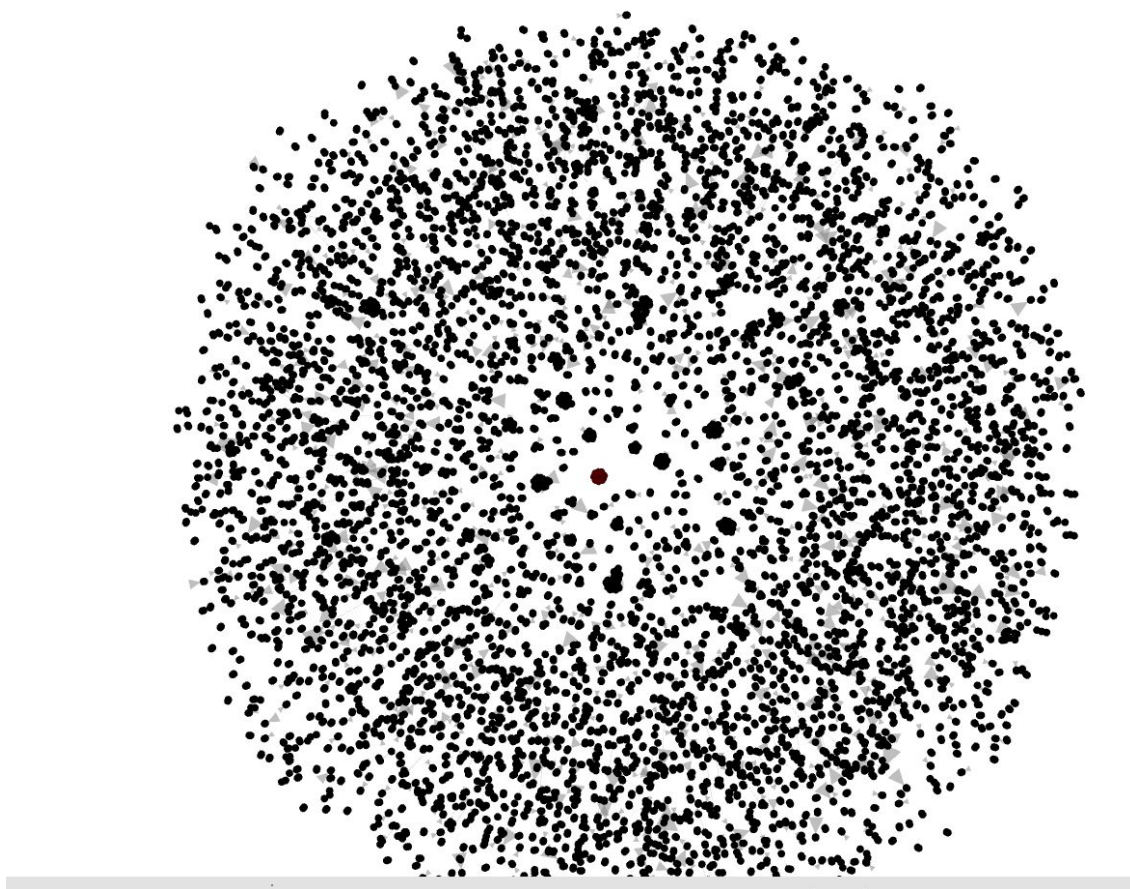
Podejście 3 do analizy 'fake news'

Podejście trzecie polegało na zwizualizowaniu wszystkich tweetów pod kątem poniższych kryteriów i spróbowaniu znalezienia pewnych odchyleń od reszty użytkowników oraz zwizualizowanie tego w programie gephi co pomogłoby nam coś zauważyć.

Kryteria:

- kto jest autorem
- z jakiego urządzenia zostały wrzucone tweety
- czy konto jest zweryfikowane (okazało się, że tylko 26 z 8882 kont jest zweryfikowane)
- liczba tweetów
- czy retweetują jakieś fake konta
- wiek
- data rejestracji

Niestety wizualizacja nie pomogła wykryć żadnych grup, czego przyczyną była zbyt mała ilość danych. Na poniższym wykresie widzimy tweety oraz ich retweety (na szaro strzałki). Nie zauważyliśmy żadnych większych zbiorów czy też wyodrębnionych grup.



Rysunek 5. - Graf wygenerowany w programie Gephi

Podejście 4 do analizy 'fake news'

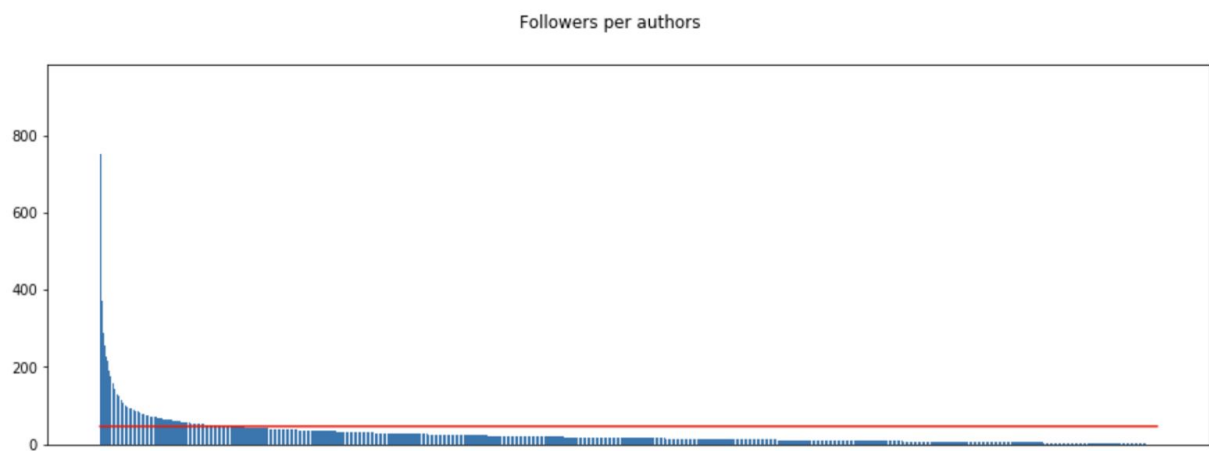
Po uświadomieniu sobie, że dane które posiadamy nie są wystarczające skupiliśmy się na analizie grup i segmentów w pobranych z Twittera autorach i tweetach.

Sposób wykonania tego był następujący: dla każdego użytkownika obliczono średnie wartości z wszystkich tweetów dla wszystkich wcześniej wymienionych kryteriów i zbudowano z niego wektor. Każdy użytkownik, który ma wartość większą niż wartość średnia dla danego 'atrybutu' należy do danego segmentu, np. użytkownik z liczbą followersów 50 gdzie średnia wynosi 25 znajduje się w segmencie "użytkownicy z dużą ilością followersów".

To podejście pozwoliło nam wyodrębnić pewne grupy, np. użytkownicy, którzy często tweetują, często re-tweetują ale rzadko komentują.

Liczba śledzących na autora.

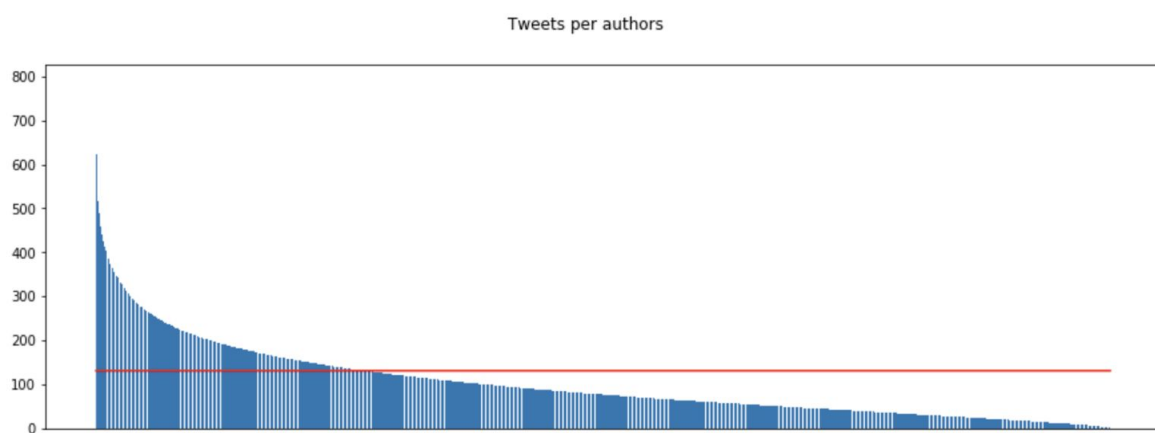
- avg 2184.17
- min 0
- max 875777
- stand dev. 19551.87



Rysunek 6. - Liczba śledzących na autora (dla skali liczba jest spierwiatkowana)

Liczba tweetów na autora:

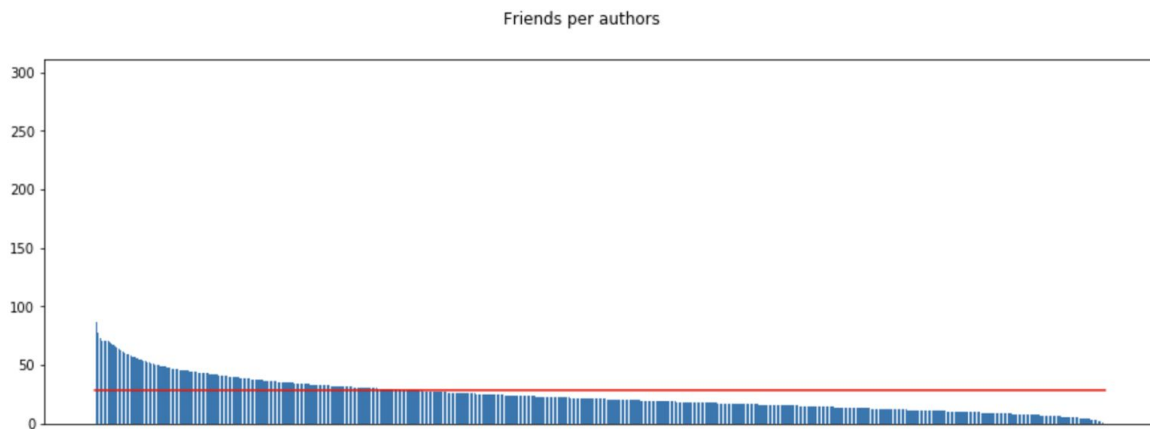
- avg 17173.35
- min 0
- max 620478
- std dev. 32765.08



Rysunek 7. - Liczba tweetów na autora (dla skali liczba jest spierwiatkowana)

Liczba obserwowanych na autora:

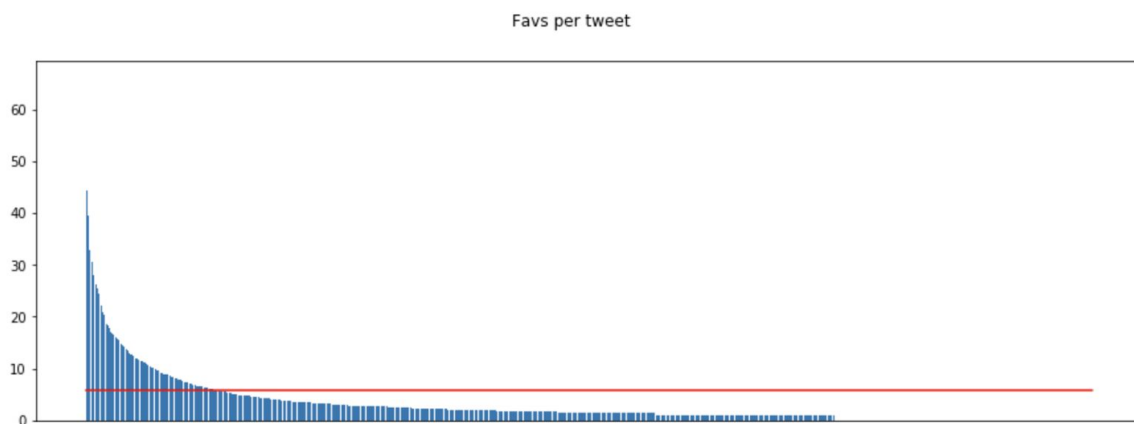
- avg 814.39
- min 0
- max 87814
- std dev. 1418.45



Rysunek 8. - Liczba obserwowanych na autora (dla skali liczba jest spierwiatkowana)

Liczba polubień na tweeta:

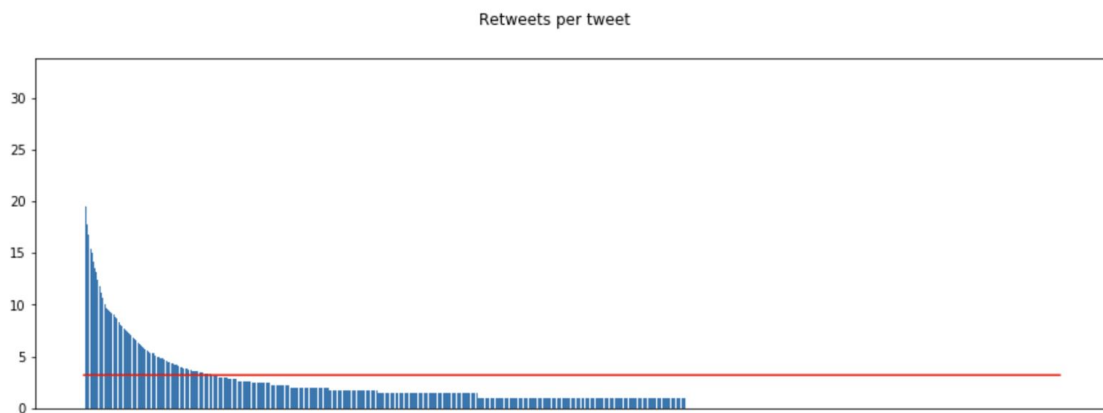
- avg 10.20
- min 0
- max 1038
- std dev. 44.19



Rysunek 9. - Liczba polubień na tweeta (dla skali liczba jest spierwiatkowana)

Liczba retweetów na tweeta:

- avg 33.62
- min 0
- max 4371
- std dev. 155.01



Rysunek 10. - Liczba retweetów na tweeta (dla skali liczba jest spierwiatkowana)

Niestety również i to podejście nie pozwoliło na wykrycie 'fake news' a bardziej na wykryciu pewnych grup użytkowników na portalu Twitter. Możliwe, że gdybyśmy wcześniej przeprowadzili taką analizę - mogłoby to nas doprowadzić do fake newsów.

5. Podsumowanie

1. Zbyt mało danych w słabej jakości.

Niestety nie mieliśmy doświadczenia w pracy z danymi z Twittera. Postanowiliśmy zaciągać dane bazując na hashtagach, a przez fakt, iż nie wszyscy używają hashtagów udało nam się pobrać zaledwie odsetek tweetów na tematy, które nas interesowały. Również nie od razu pobraliśmy wszystko co było nam potrzebne i część danych nie była już później dostępna. Dodatkowo wszyscy użytkownicy, których dane udało nam się pobrać z Twittera nie miało zweryfikowanego konta co według licznej literatury ma wpływ na postrzeganie takich kont jako źródła fake news.

2. Polski Twitter nie jest łatwy do analizy

Na polskim Twitterze w przeciwieństwie do innych państw nie odbywa się aż tak dużo aktywności. Świetnym tego potwierdzeniem może być fakt że nawet profile największych partii politycznych nie są oznaczone jako "zweryfikowane". Dodatkowo język polski nie jest łatwy do analizy tekstu i nie było to możliwe do zastosowania na większą skalę w trakcie trwania projektu.

3. Brak prawdziwych fake newsów ?

Przeprowadzone przez nas analizy nie wskazały jednoznacznie żadnych fake newsów ani fake kont. Znaleźliśmy co prawda konta "podejrzane" ale były to po prostu tzw. "trolle internetowe" lub osoby fanatyczne względem którejś opcji politycznej. Nie publikowały one jednak fake newsów ale stronnicze, politycznie zaangażowane wpisy.

Wątpliwości budzi też, czy wybory samorządowe były na tyle interesujące i ważne by ktoś miał interes w tym by takie fake news produkować.

Źródła:

Repozytorium z kodem projektu na portalu Github:

https://github.com/piter77/data_exploration

Link do strony projektu na Trac'u:

<https://trac.iisg.agh.edu.pl/ed/wiki/projects/2018/TwitterFakeNews>

6. Literatura

1. <https://www.sciencemag.org/news/2018/03/fake-news-spreads-faster-true-news-twitter-thanks-people-not-bots>
2. <https://www.prattlibrary.org/research/tools/index.aspx?cat=90&id=4735>
3. <https://www.prospectmagazine.co.uk/science-and-technology/the-science-of-fake-news>
4. <https://www.media.mit.edu/projects/the-spread-of-false-and-true-info-online/frequently-asked-questions/#faq-whats-the-background-on-this-research>
5. https://pdfs.semanticscholar.org/aa38/1b7151c57d0de9318b0f6b6a76cac50c5e24.pdf?_ga=2.22446203.1722511163.1541529970-1343671376.1541529970
6. <https://www.sciencenews.org/article/can-computer-programs-flag-fake-news>
7. <https://oko.press/sfalszowano-wybory-krzyczaly-w-sieci-podejrzane-konta-tak-dzialaja-rosyjskie-trolle/>
8. <https://klubjagiellonski.pl/2018/11/06/demokracja-w-dobie-fake-newsow-jak-rosyjskie-farmy-trolli-mieszaly-w-amerykanskich-wyborach/>
9. https://www.researchgate.net/publication/280782550_Fake_Twitter_accounts_Profile_characteristics_obtained_using_an_activity-based_pattern_detection_approach
10. https://www.researchgate.net/publication/263928472_Calculating_statistical_similarity_between_sentences