AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

WYDZIAŁ INFORMATYKI, ELEKTRONIKI I TELEKOMUNIKACJI
KATEDRA INFORMATYKI

PRACOWNIA PROBLEMOWA

Wpływowość użytkowników polskiego Twittera publikujących w temacie koronawirusa

Influence of polish Twitter users publishing on the subject of coronavirus

Autorzy: Wojciech Gruszka, Antoni Pięta
Kierunek studiów: Informatyka
Typ studiów: Stacjonarne
Opiekun: dr inż. Jarosław Koźlak

Kraków, 2020

Spis obrazków

1. Wprowadzenie

1.1. Motywacja pracy

Z każdym rokiem rośnie popularność mediów społecznościowych. Powiększające się grono odbiorców zachęca coraz większą liczbę użytkowników do podejmowania prób zdobywania popularności i wpływowości poprzez publikowanie treści. Taką wpływowość można wykorzystywać w wielu celach. Politycy mogą próbować przekuć ją w zwiększenie poparcia w społeczeństwie, przekładającego się na szansę zwycięstwa w wyborach. Influencerzy z kolei wykorzystują swoją popularność w celach zarobkowych. Mogą wiązać się z firmami w celu przeprowadzania kampanii reklamowych.

Co więcej, granica pomiędzy byciem odbiorcą a nadawcą jest bardzo wąska, ponieważ w mediach społecznościowych przeważnie każde konto ma możliwość publikacji treści. Odróżnia to media społecznościowe od tradycyjnych, gdzie bardzo jednoznacznie można rozróżnić osoby publikujące - dziennikarzy, redaktorów, polityków - od odbierających - widzów, czytelników.

Zarówno te aspekty, jak i wiele innych, powodują, że tematyka jest niezwykle ciekawa i w naszej opinii zasługuje na szersze przebadanie, czego usiłujemy dokonać w tej pracy.

1.2 Cele pracy

Za główny cel pracy postawiliśmy przegląd społeczności Twittera przeprowadzony z użyciem różnych technik. Chcemy zdefiniować różne miary popularności i wpływowości, aby móc przeanalizować wpływowość technikami statystycznymi. Chcemy sprawdzić jak rozkładają się wyniki osiągane przez użytkowników, a także skategoryzować ich zależnie od charakteru uzyskanych rezultatów. Kolejną możliwością jest sprawdzenie jak wyniki użytkowników zmieniały się w czasie.

Na podstawie takiej analizy chcemy próbować wyciągnąć wnioski na temat sposobu osiągania wpływowości. Można też wywnioskować jak bardzo wpływowość konkretnych użytkowników została osiągnięta za pomocą Twittera i innych mediów społecznościowych, a w jakiej części została zbudowana w inny sposób.

2. Przegląd dziedziny

2.1 Podobne prace

Koronawirus jest nowym tematem, więc ciężko odnaleźć prace na temat analizy wpływowości użytkowników w mediach społecznościowych na temat koronawirusa.

Jedna ze znalezionych przez nas praco opowiada o negatywnych skutkach używania mediów w czasie kryzysu wirusowego [**coronavirusImpact**]. Zwraca uwagę, że duża część użytkowników zajmuje się nieistotnymi tematami, często powielając informację, które są zarówno nieprawdziwe, jak i szkodliwe. Podkreślono również, że bardzo ważne jest by opierać się na wielu użytkownikach, którzy specjalizują się w różnych dziedzinach.

Jednak, ogólniejszy temat analizy wpływowości użytkowników w mediach społecznościowych jest dużo lepiej opisany. Zapoznaliśmy się z kilkoma pracami na ten temat.

Jedną ze znalezionych przez nas prób analizy w tym temacie na Twitterze jest ta przeprowadzona przy pomocy dekompozycji K-shell [**kshellPaper**]. Wejściem do algorytmu jest graf skierowany użytkowników, w którym krawędzie odpowiadają obserwacji między użytkownikami. Do algorytmu wprowadzono modyfikacje m.in. usunięto rówieśników, czyli krawędzie, które miały obustronne relacje. Jedną z użytych grafowych metryk był PageRank. W wyniku działania algorytmu udało się skutecznie zidentyfikować małe grupy influencerów.

W kolejnej analizowanej przez nas pracy, autorzy skupili się na obserwowaniu dynamiki zdobywania wpływowości **[timeInfluencePaper].** Najważniejszymi metrykami, które zostały użyte do analizy każdego użytkownika były: liczba obserwujących, retweetów oraz wspomnień przez innych użytkowników. Udało się zaobserwować, że liczba obserwujących ma niewiele wspólnego ze zdobytą wpływowością. Co więcej, wpływowość mogła być zdobyta przez użytkownika w kilku tematach. Zdobywanie wpływowości nie było nagłe, lecz wymagało dłuższej koncentracji na jednym temacie.

2.2 Algorytmy klastrowania

Klastrowanie to rozdzielenie danych na grupy, które zawierają dane podobne do siebie. Jest to rodzaj uczenia nienadzorowanego, ponieważ próbujemy uzyskać informację bezpośrednio z danych. Algorytmy klastrowanie dzielimy ze względu na metodą, którą

próbujemy uzyskać wiedzę na: hierarchiczne, oparte na centoridach oraz oparte na rozkładzie.

Najpopularniejszym algorytmem klastrowania jest *k-means*, który polega na iteracyjnym poprawianiu środka grupy i na jego podstawie wyłapywaniu grup, które leżą najbliżej centrum. Jest to algorytm na podstawie centroidów.

Kolejnym algorytmem jest *Gaussian mixture*, który polega na stworzeniu określonej liczby losowo zainicjalizowanych rozkładów Gaussa, a następnie poprawianie ich iteracyjnie, tak by lepiej pasowały do danych wejściowych. Jest to oczywiście algorytm, który jest oparty na rozkładzie.**[clusterAlgorithms]**

3. Koncepcja

3.1 Pobieranie danych

Zbiór danych stworzyliśmy poprzez odpytywanie API Twittera według słów kluczowych. Słowa te wybraliśmy w taki sposób, aby objęły jak najszerszy zbiór tweetów związanych z wirusem, ale jednocześnie unikając innych tematów. Ich lista to:

- koronawirus
- wirus
- COVID19
- COVID
- COVID-19
- Sars-Cov-2
- epidemia
- pandemia
- GIS
- kwarantanna
- współistniejące
- zakażenie
- zakaźny

a także ich odmianę przez przypadki.

Zbiór składa się z dwóch podzbiorów. Okres od 25.03.2020 do 3.06.2020 jest kompletnym zbiorem polskich tweetów zawierające dowolne z podanych słów kluczowych, które zostały opublikowane w tym okresie.
Z kolei z dni od 8.03.2020 do 25.03.2020 mogliśmy ze względu na ograniczenia API pobrać tylko część tweetów. W związku z tym stworzyliśmy listy 100 użytkowników z największą liczba obserwujących oraz 100 użytkowników z największą średnią liczbą polubień w okresie od 25.03.2020 do 2.04.2020 z dolnym limitem 5 postów opublikowanych w tym czasie. Następnie pobraliśmy tweety tych użytkowników z podanego okresu.

Zebraliśmy w ten sposób ponad 1,6 mln tweetów od 18 tys. unikalnych użytkowników.

3.2 Metryki

W celu analizy zbioru wyliczyliśmy zbiór metryk, które miały za zadanie reprezentować różne aspekty wpływowości.

Odrzuciliśmy użytkowników, którzy podczas okresu 23.03-21.04 opublikowali mniej niż 6 postów o tej tematyce (a więc publikujących raz na dwa dni lub rzadziej). Uznaliśmy bowiem, że takich użytkowników nie można uznać za regularnie wypowiadających się w tym temacie. W przeciwnym wypadku mogliby oni uzyskać dobre wyniki metryk przez generalną popularność swojego profilu i znaleźć się na naszej liście, nie będąc istotnymi w kontekście tematu koronawirusa.

Przyjęliśmy i skategoryzowaliśmy następujące metryki:

Proste wartości:

- całkowita **liczba followersów** (followers_count) - pozwala przejrzeć najbardziej popularne profile polskiego Twittera, które od długiego czasu budowały swoją pozycję i są uznawane za autorytety. W tej grupie jest dużo profili mediów - telewizji, gazet, portali informacyjnych, a także renomowanych dziennikarzy i najpopularniejszych polityków.
- całkowita **liczba tweetów** opublikowanych w tym temacie (tweet_count) - pozwoli sprawdzić czy wysoka aktywność wpływa na popularność - najbardziej aktywne profile publikują do 100 tweetów dziennie. Tu także jest dużo mediów, ale pojawiają się też niezależni influencerzy
- **całkowita liczba polubień** (sum_favorite) - metryka podobna do powyższych, ale sumująca wszystkie zebrane polubienia. Pozwoli na odnalezienie tych użytkowników, którzy pomimo dużej liczby wrzucanych materiałów potrafią utrzymać wysoką aktywność odbiorców. W tej metryce w przeciwieństwie do średniej może być trudniej osiągnąć dobry rezultat, ponieważ należy jednocześnie dbać o częste publikowanie materiałów, jak i wysoką jakość.
- **całkowita liczba retweetów** (sum_retweet) - metryka analogiczna do powyższej.
- **liczba unikalnych retweetowiczów** (unique_retweet_count) - liczba użytkowników, którzy co najmniej raz retweetowali post użytkownika. Pozwala na określenie wielkości grona aktywnych odbiorców danego użytkownika. Pomoże odróżnić użytkowników o wąskim, ale aktywnym gronie od użytkowników docierających do wielu mało zaangażowanych osób.

Średnie i mediany:

- **średnia liczba polubień** na tweet (avg_favorite) - pozwala śledzić które profile generują obecnie największą aktywność odbiorców. Ta metryka powinna w większym stopniu zależeć od atrakcyjności i stopnia angażowania obecnie wrzucanych treści, a w mniejszym od pozycji zbudowanej w społeczności.
- **średnia liczba retweetów** na tweet (avg_retweet) - jak wyżej. Ta metryka w dużej mierze pokrywa się z poprzednią, jednak w niektórych sytuacjach niekoniecznie musi iść z nią w parze. Przykładowo dla ważnych treści informacyjnych wzbudzających negatywne emocje polubień powinno być mało, a retweetów dużo.
- **mediana polubień** (med_favorite) - w zestawieniu ze średnią polubień pomaga sprawdzić skośność rozkładu polubień, a w konsekwencji ocenić jak równomiernie były zdobywane polubienia.
- **mediana retweetów** (med_retweet) - w zestawieniu ze średnią retweetów pomaga sprawdzić skośność rozkładu retweetów, a w konsekwencji ocenić jak równomiernie były zdobywane retweetów.

Średnie ważone i średnie stosuniki metryk:

- **ważona suma polubień i retweetów** (weighed_sum) - metryka będąca kombinacją sum_favorite i sum_retweet, a więc oceniająca w formie jednej liczby wyniki aktywności pod postami danego użytkownika.
- **ważona średnia polubień i retweetów** (weighed_sum_with_cost) - metryka analogiczna do powyższej, ale oceniająca średnią aktywność na post danego użytkownika. Jest więc kombinacją metryk avg_favorite i avg_retweet.
- **średnia liczba polubień / liczba followersów** (fav_to_fol_ratio) - kombinacja avg_favorite i followers_count, pozwala ocenić stopień zaangażowania odbiorców bez względu na ich liczbę. Metryka ta nie daje zadowalających rezultatów, ponieważ łatwo jest osiągnąć wysoki wynik z powodu bardzo małej liczby followersów.
- **Średnia liczba retweetów / liczba followersów** (ret_to_fol_ratio) - analogicznie do powyższej, ale biorąc pod uwagę średnią liczbę retweetów.

Na podstawie grafu skierowanego retweetów (prowadzi od osoby retweetującej do retweetowanej), gdzie wagą jest liczba retweetów:

- **Closeness centrality**
- **Betweenness centrality**
- **PageRank**
- **Liczba wejściowych krawędzie w grafie**
- **Liczba wejściowych krawędzie w grafie z wagą większą lub równą 5**
- **Liczba wejściowych krawędzie w grafie z wagą większą lub równą 10**

Przyjmując następujące oznaczenia:

FOL_u – zbiór użytkowników followujących użytkownika u

RET_u – zbiór użytkowników, którzy zretweetowali co najmniej jeden tweet użytkownika u

T_u – zbiór tweetów opublikowanych przez użytkownika u

$f_t(r_t)$ – liczba polubień (retweetów) uzyskanych przez tweet t

$w_f(w_r)$ – przyjęta waga jednego polubienia (retweeta)

$$followers\_count = |FOL_u|$$

$$tweet\_count = |T_u|$$

$$sum\_favorite = \sum_{t \in T_u} f_t$$

$$sum\_retweet = \sum_{t \in T_u} r_t$$

$$avg\_favorite = \frac{sum\_favorite}{tweet\_count} = \frac{\sum_{t \in T_u} f_t}{|T_u|}$$

$$avg\_retweet = \frac{sum\_retweet}{tweet\_count} = \frac{\sum_{t \in T_u} r_t}{|T_u|}$$

$$weighed\_sum = w_f * sum\_favorite + w_r * sum\_retweet = w_f * \sum_{t \in T_u} f_t + w_r * \sum_{t \in T_u} r_t$$

$$weighed\_avg = \frac{weighed\_sum}{tweet\_count} = \frac{w_f * \sum_{t \in T_u} f_t + w_r * \sum_{t \in T_u} r_t}{|T_u|}$$

$$retweeters\_count = |RET_u|$$

$$med\_favorite = median\left(\bigcup_{t \in T_u} f_t\right)$$

$$med\_retweet = median\left(\bigcup_{t \in T_u} r_t\right)$$

$$fav\_to\_fol\_ratio = \frac{avg\_favorite}{followers\_count} = \frac{\sum_{t \in T_u} f_t}{|T_u| * |FOL_u|}$$

$$ret\_to\_fol\_ratio = \frac{avg\_retweet}{followers\_count} = \frac{\sum_{t \in T_u} r_t}{|T_u| * |FOL_u|}$$

Rys. 1 - Wzory metryk

Wzory metryk możemy obserwować na rysunku 1. Do wyliczania podstawowych metryk nie są brane retweety użytkownika.

Listy 50 najlepszych użytkowników z każdej metryki znajduje się w linku.
[Pracownia problemowa 2020 Twitter - ID użytkowników]

Część użytkowników znajduje się w więcej niż jednym rankingu, więc ostatecznie mamy na ten moment 168 użytkowników z wszystkich kategorii

3.3 Taksonomia

Użytkowników, którzy znajdowali się w dowolnym z kilku pierwszych rankingów podzieliliśmy ręcznie na kategorie w zależności od roli pełnionej w społeczeństwie. Wyróżniliśmy 4 duże grupy dzielące się na podgrupy, a także dwie grupy dodatkowe, które służą do odfiltrowania botów i profili błędnie uwzględnionych w zbiorze, najczęściej przez błędne uznanie profilu zagranicznego jako polski.

Wyodrębnione grupy to:

- Media – gazety, telewizje, portale informacyjne
    - związane z rządem - mr
    - związane z opozycją - ma
    - newsowe - mn
    - muzyczne - mm
    - sportowe - ms
    - technologiczne - mt
- Politycy postujący nie z ramienia funkcji rządzącej
    - aktywni:
        - związani z rządem - pr
        - związani z opozycją - po
    - byli:
        - obecnie związani z rządem - bpr
        - obecnie związani z opozycją - bpo
- Dziennikarze – profile osób fizycznych informujących i komentujących wydarzenia
    - związani z rządem - dr
    - związani z opozycją - do
    - newsowi - dn
    - sportowi - ds
- Influencerzy - kategoria bez ścisłej definicji.
    - ze świata sportu - is
    - o profilu humorystycznym - ih
    - związani z Kościołem - ik
    - prowadzący profil jak na Facebooku - if
    - związani z technologią - it
    - związani z rządem - ir
    - związani z opozycją - io
    - związanie ze światem medycznym - im
- Boty - b
- Błędne profile - najczęściej nie polskie - x

3.4 Klastrowanie

Istotną część analizy społeczności stanowi klastrowanie. Metody tego typu zostały opisane w podrozdziale 2.2. Wykorzystaliśmy je w celu wyodrębnienia grup użytkowników o podobnej charakterystyce. Pożądane było zarówno wyodrębnienie grup uzyskujących wyjątkowo wysokich, jak i wyjątkowo niskich wartości metryk świadczących o wpływowości.

Przeprowadziliśmy kilka klastrowań w celu przebadania społeczności pod różnymi kątami. Możemy wyróżnić następujące sposoby dywersyfikowania metod klastrowania:

- ze względu na czas:
    - statyczne
    - dynamiczne
- ze względu na metodę klastrowania:
    - KMeans
    - EM Clustering
- ze względu na liczbę klastrów
- ze względu na podzbiór uwzględnianych metryk.

Klastrowanie statyczne nie uwzględnia zmienności społeczności w czasie i traktuje tweety z całego okresu badań jako jednolity zbiór danych. Z kolei dynamiczne ma za zadanie zbadać nie tylko grupy występujące w zbiorze, ale także ich zmienność. Z kolei manipulowanie metrykami branymi pod uwagę przy klastrowaniu pozwala na wyodrębnienie grup mających określone przez nie cechy.

Scenariusze klastrowania utworzyliśmy poprzez odpowiednie dobieranie powyższych parametrów tak, aby przebadać zbiór pod różnymi kątami. Zależało nam między innymi na zbadaniu grup o mniejszym i większym poziomie szczegółowości, na analizie pod kątem cech, które potencjalnie mogą charakteryzować poszczególne kategorie użytkowników, a także badać dynamikę klastrów. Dokładne scenariusze opisane są w podrozdziale 5.5.1.

4. Realizacja

4.1 Pobieranie tweetów

Tweety pobieramy z udostępnionego API do wyszukiwania tweetów - Twitter Search API [**TwitterSearchAPI**]. Każda frazę pobieramy indywidualnie, a aby pobierać dane z odpowiedniego okresu, regulujemy parametry *until* i *since_id.*

4.2 Zapisywanie danych

Dane zapisujemy do bazy NoSQL MongoDB. Adres do bazy podajemy za pomocą hosta, więc baza może być zarówno lokalna, jak i znajdować się w zewnętrznym hostingu. W bazie przechowujemy trzy rodzaje rekordów: tweety, użytkowników oraz dane, po których wyszukujemy. Przed dodaniem tweeta do bazy, sprawdzamy czy dany obiekt nie znajduje się już bazie.

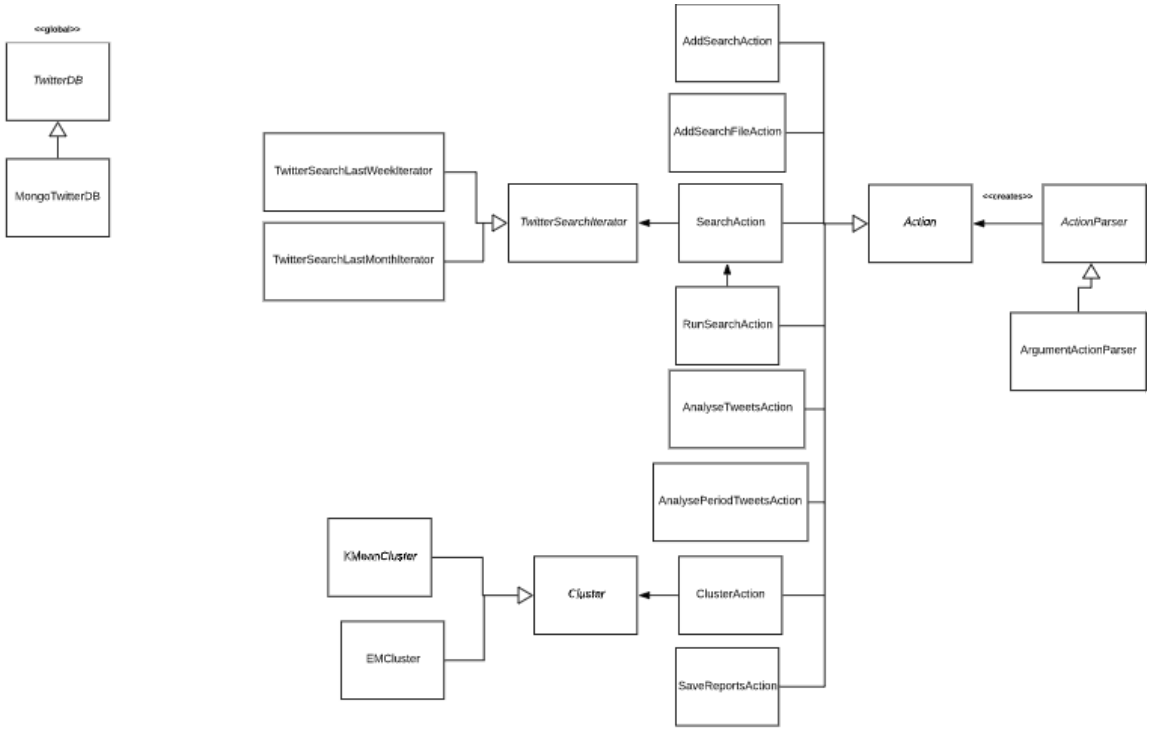
4.3 Wyliczenia statystyk i metryk

Do wyliczenia statystyk i metryk stosujemy wbudowane narzędzia do pythona, biblioteki Numpy [**numpy**] czy statistics [**statistics**] oraz agregację MongoDB. Wbudowane narzędzia pythona oraz biblioteka statistics służą do wyliczenia niezbyt skomplikowanych statystyk jak np. mediana czy odchylenie standardowe. Bibliotek numpy pozwala na bardziej skomlikowane obliczenia, wymagające operowania na złożonych tablicach. Agregacja MongoDB pozwala w zoptymalizowany sposób wyliczać metryki poprzez łączenia między różnymi tabelami w bazie danych.

4.4 Klastrowanie

Do klastrowania użyliśmy biblioteki *sklearn.cluster* [**sklearn**], z których używamy k-średnich oraz algorytm EM. Klastrujemy na podstawie metryk, które są zapisane są w bazie danych.
Przypisane *labele* zapisujemy w bazie, a wygenerowane wykresy zapisujemy na dysku.

4.5 Diagram klas



Rys. 2 - Diagram klas

Na rysunku nr. 2 możemy zobaczyć, jak wygląda diagram klas naszego projektu. Wyróżniamy bazową klasę *Action,* która odpowiada różnego rodzaju akcją, które wykonuje nasz program. Klasy odpowiadające konkretnych zachowaniom dziedziczą po bazowej klasie. Wyróżniamy następujące klasy dziedziczące po *Action:*

- *AddSearchAction* - Dodanie pojedynczej kwerendy wyszukiwania.
- *AddSearchFileAction* - Dodanie kwerend wyszukiwania, które są zapisane w pliku.
- *SearchAction* - Wyszukanie tweetów na podstawie podanej kwerendy, która już jest zapisana w bazie danych.
- *RunSearchAction* - Domyślna akcja projektu. Zajmuje się ciągłym pobieraniem tweetów na podstawie wszystkich zapisanych kwerend.
- *AnalyseTweetsAction* - Wylicza metryki oraz wykresy dla całego przedziału czasu dla tweetów w bazie, następnie zapisuje je w bazie.
- *AnalysePeriodTweetsAction* - Wylicza metryki oraz wykresy dla określonych przedziałów czasu dla tweetów w bazie, następnie zapisuje je w bazie.
- *ClusterAction* - Przeprowadza klastrowanie, wyniki zapisuje w bazie.
- *SabeReportsAction* - Na podstawie wyliczonych metryk oraz danych klastrowania, generuje raport.

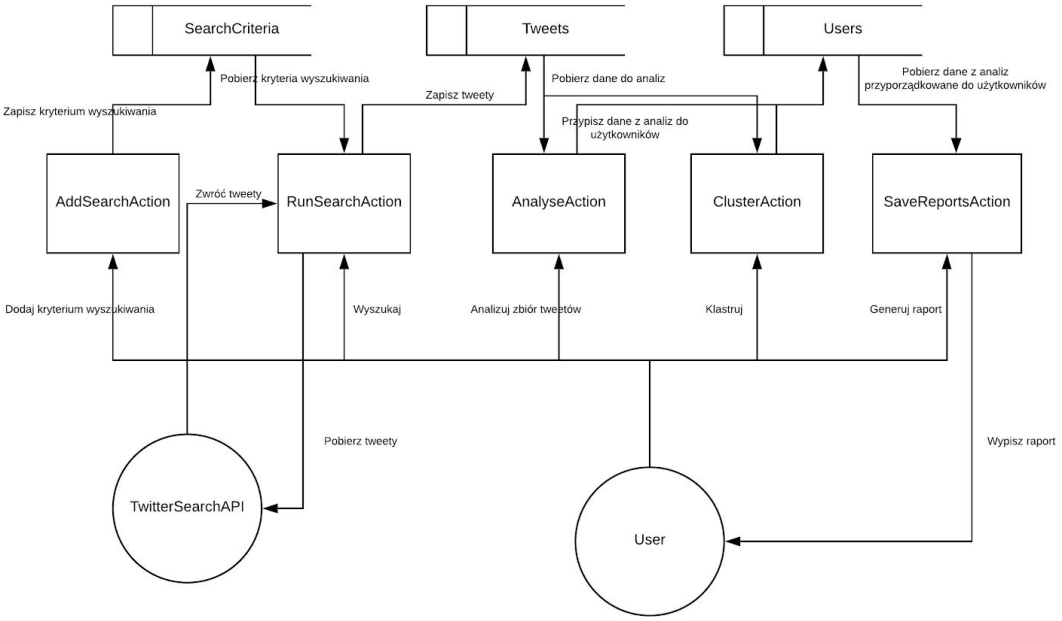
Tworzeniu akcji opowiada interface *ActionParser.* Jedyną jego implementacją jest Argument*ActionParser,* który parsuje dane z konsoli i na ich podstawie tworzy akcje.

Kolejną bazowa klasą jest *Cluster*, której implementację zajmują się klastrowanie różnymi metodami. Wyróżniamy *KMeansCluster*, która klastruje metodą k-średnich oraz *EMCluster*, która klastruje metodą *Gaussian mixture.*

Pomocniczą klasami są bazowa *TwitterSearchIterator* oraz jej implementacje *TwitterLastMonthSearchIterator i TwitterLastWeekSearchIterator.* Zajmują się one pobieraniem danych z Twittera. Pierwsza służy pobieraniu danych z ostatniego miesiąca, druga z ostatniego tygodnia. Rozdzielenie tych klas wynika z różnic w polityce ściągania tweetów w tych okresach.

Ostanimi klasmi są bazowa *TwitterDB* i jej implementacja *MongoTwitterDB,* która jest inicjalizowana jako singleton, a przez to dostępna globalnie. Bazowa klasa udostępnia interface do komunikacją z baża danych, a *MongoTwitterDB* implementuje te zachowania w kontekście *MongoDB*.

4.6 Diagram przepływu danych



Rys. 3 - Diagram przepływu danych

W diagramie przepływu danych przedstawionym na rysunku nr. 3 możemy wyróżnić trzy typy obiektów: rekordy, akcje oraz użytkowników. Co więcej, mamy połączenia między nimi, które odpowiadają przepływowi danych między nimi. Akcje zostały opisane w poprzednim rozdziale, więc skupimy się na opisaniu pozostałych obiektów.

W bazie danych wyróżniamy trzy rodzaje obiektów:

- *SearchCriteria* - Pojedyncze kryterium wyszukiwania. Zawiera kryterium, po którym wyszukuje się w API Twittera, date ostatnio pobranego tweeta, używając tego tweeta oraz ostatni pobrany identyfikatora tweeta.
- *Tweets* - Pobrane tweety.
- *Users* - Dane użytkowników twittera wraz z przypisanymi im metrykami.

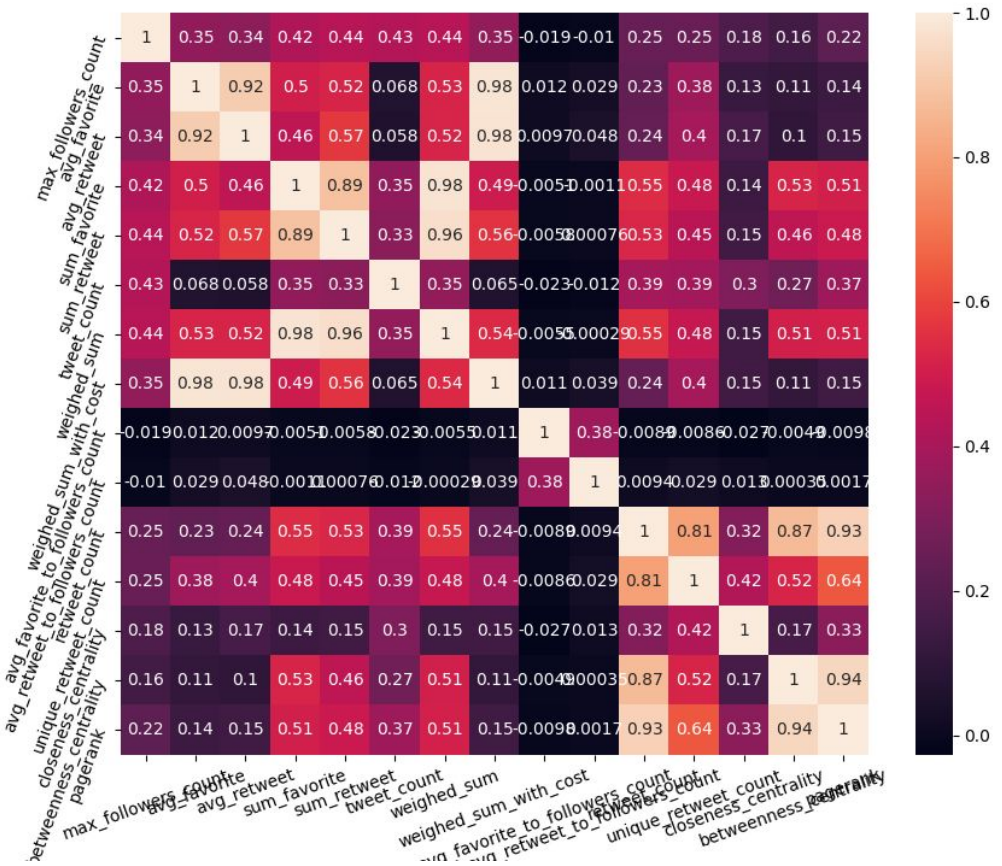
Wyróżniamy dwóch użytkowników w systemie:

- *TwitterSearchAPI* - API Twittera do pobieranie tweetów. System pobiera dane na podstawie konkretnych kryteriów.
- User - Użytkownik, który używając konsoli, wywołuje konkretne akcje.

5. Ewaluacja

5.1 Korelacje

Wyliczyliśmy korelacje między wynikami metryk i przedstawiliśmy je na rysunku 4.



Rys. 4 - Macierz korelacji pomiędzy metrykami użytkowników

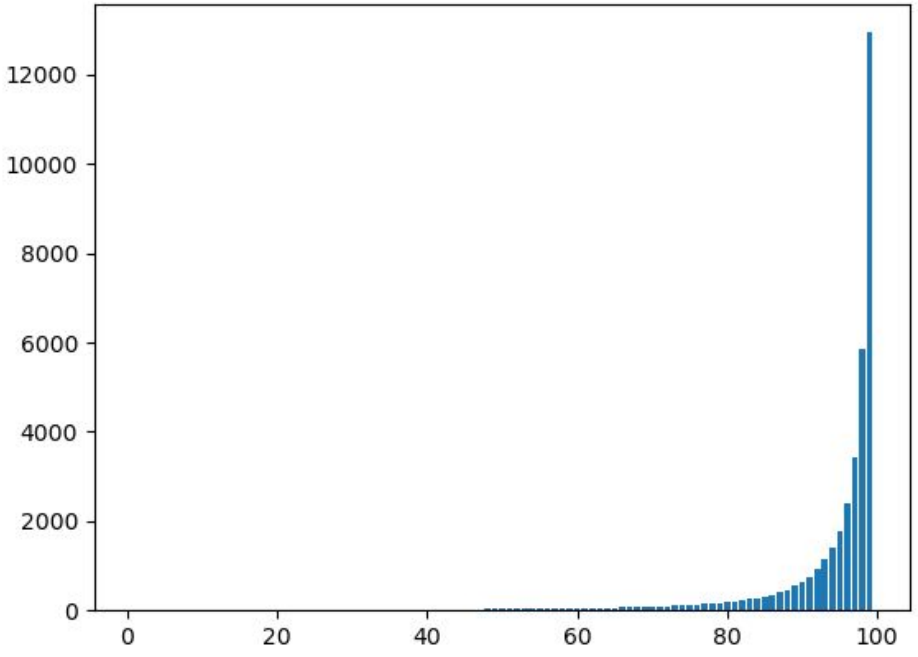
Jak widać wartości polubień i retweetów bardzo silnie korelują ze sobą, a także oczywiście z metrykami będącymi kombinacją tych wartości. Liczba followersów wykazuje średnią korelację z każdą miarą. Zależności pomiędzy sumą a średnią danego parametru także jest średnia.

Wartości metryk stosunku średniej liczby polubień do liczby followersów (avg_favorite_to_follower) oraz stosunku średniej liczby retweetów do liczby followersów (avg_retweet_to_follower) nie korelują z żadną metryką, poza wzajemną korelacją. Po dokładniejszej analizie okazało się, że metryka ta nie daje zadowalających rezultatów, ponieważ użytkownikom niepopularnym bardzo łatwo uzyskać w niej wysokie wyniki.

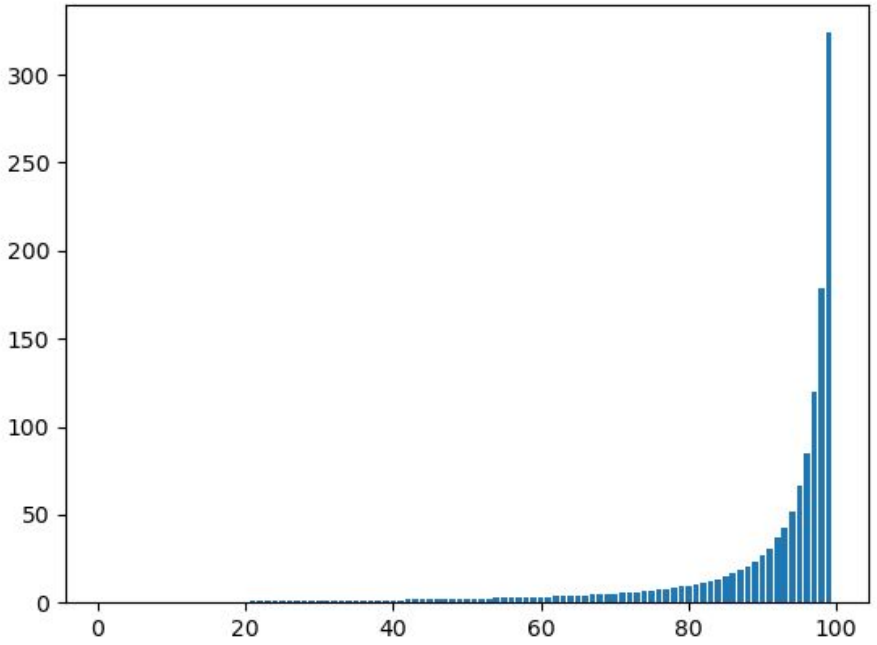
5.2 Rozkład wyników w populacji

Aby zobaczyć jak rozkładają się wartości zmiennych narysowaliśmy na podstawie wyników metryk dla całej populacji wykresy słupkowe o dokładności 1/100. Wszystkie w dużym stopniu spełniają power law, gdzie skrajnie mała część populacji osiąga skrajnie duże wyniki. Poniżej przedstawiamy szczegóły wyników dla poszczególnych metryk.

Analogiczne metryki polubień i retweetów okazują się nie tylko korelować ze sobą, ale też mieć podobną charakterystykę. Różnica występuje tylko pomiędzy metrykami sumarycznymi, a średnimi. Zamieszczamy więc na rysunkach 5 i 6 porównanie tych dwóch grup na przykładzie ważonej sumy polubień i retweetów oraz tej sumy podzielonej przez koszt publikowania. Dla sumy średnia równa 1050 wypada w okolicach 93 centyla. Dla sumy podzielonej przez liczbę tweetów średnia równa 19 wypada w okolicach 87 centyla.

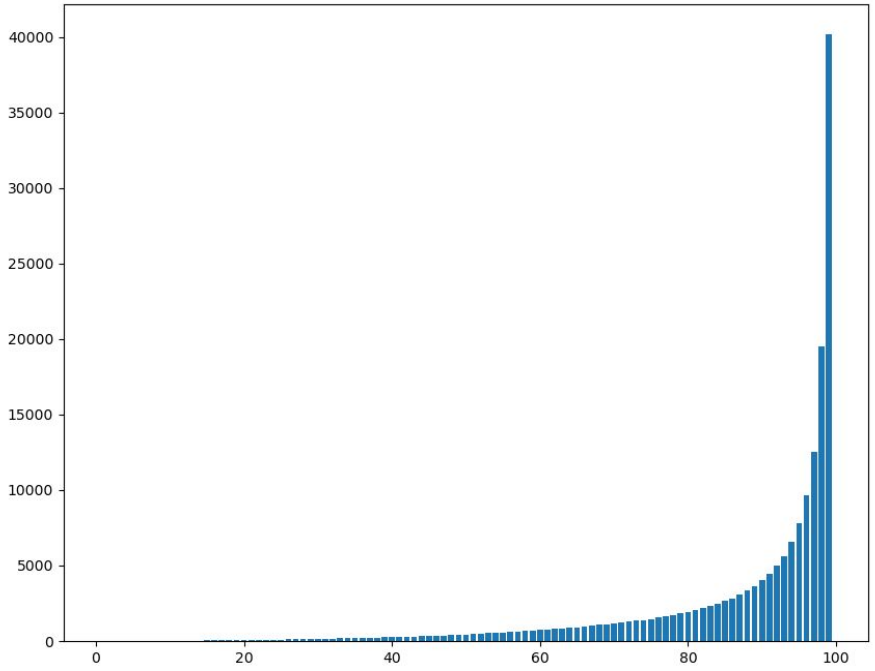


Rys. 5 - Rozkład metryki sumy ważonej polubień i retweetów



Rys. 6 - Rozkład metryki sumy ważonej polubień i retweetów podzielonej przez liczbę tweetów

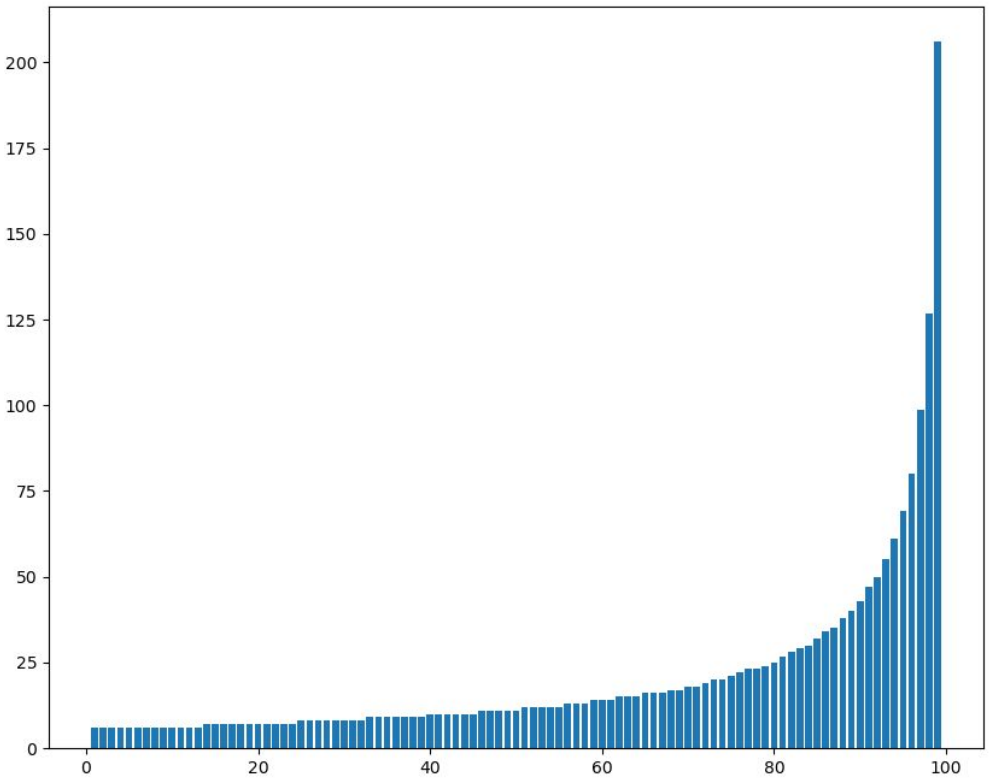
Jak widać podzielenie metryki sumy ważonej przez koszt publikacji tweetów powoduje że wykres jest nieco mniej zakrzywiony. Oznacza to prawdopodobnie, że użytkownicy osiągający najlepsze wyniki dużo publikują, a więc ich wynik spada po uwzględnieniu pracy włożonej w publikacje.



Rys. 7 - Rozkład metryki liczby followersów

Wykres liczby followersów przedstawiony na rysunku 7 jest nieco łagodniejszy od poprzednich metryk. Może to oznaczać, że łatwiej jest zdobyć od kogoś following, niż utrzymać jego uwagę. Średnia równa 3400 wypada w okolicach 89 centyla.

Ostatnim wykresem jest liczba tweetów w tym okresie przedstawiona na rysunku 8.



Rys. 8 - Rozkład metryki liczby opublikowanych tweetów

Mimo odrzucenia użytkowników, którzy opublikowali mniej niż 6 postów w tym okresie widać że w dalszym ciągu metryka spełnia power law - zaledwie 20% użytkowników opublikowało więcej niż 25 postów. Średnia równa 24 wypada w okolicach 80 centyla.

5.3 Charakterystyka grafu retweetów

Graf retweetów jest niespójny, więc nie mogliśmy policzyć dla niego promienia czy średnicy.

Aby użytkownik był brany pod uwagę przy budowie grafu, musi mieć co najmniej 5 tweetów w naszej bazie.

Liczba krawędzi: 415274

Liczba wierzchołków: 21763

Największa liczba połączonych wierzchołków: 78

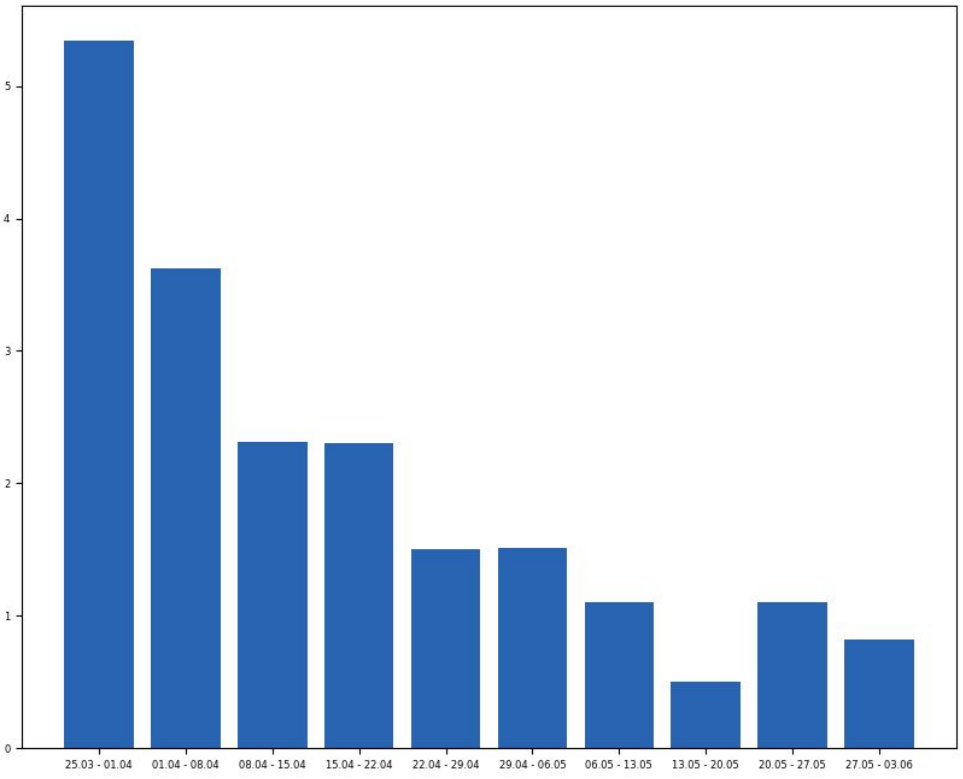
5.4 Analiza średnich tygodniowych metryk

Stworzyliśmy także wykresy, które przedstawiają średnie wartości metryk dla użytkowników w kolejnych tygodniach. Dane na podstawie, których tworzyliśmy wykresy, mogą nie być kompletne, ponieważ kompletne dane posiadamy od 25.03, a o sprawie koronawirusa zaczęto pisać wcześniej.

Celem stworzenia wykresów było dopasowanie zmian, które będą zwizualizowane na wykresie do realnych wydarzeń, związanych ze sprawą koronawirusa. Do rozpoznania, jakie zdarzenia działy się w danym okresie, korzystamy z kalendarium [**kalendarium**].

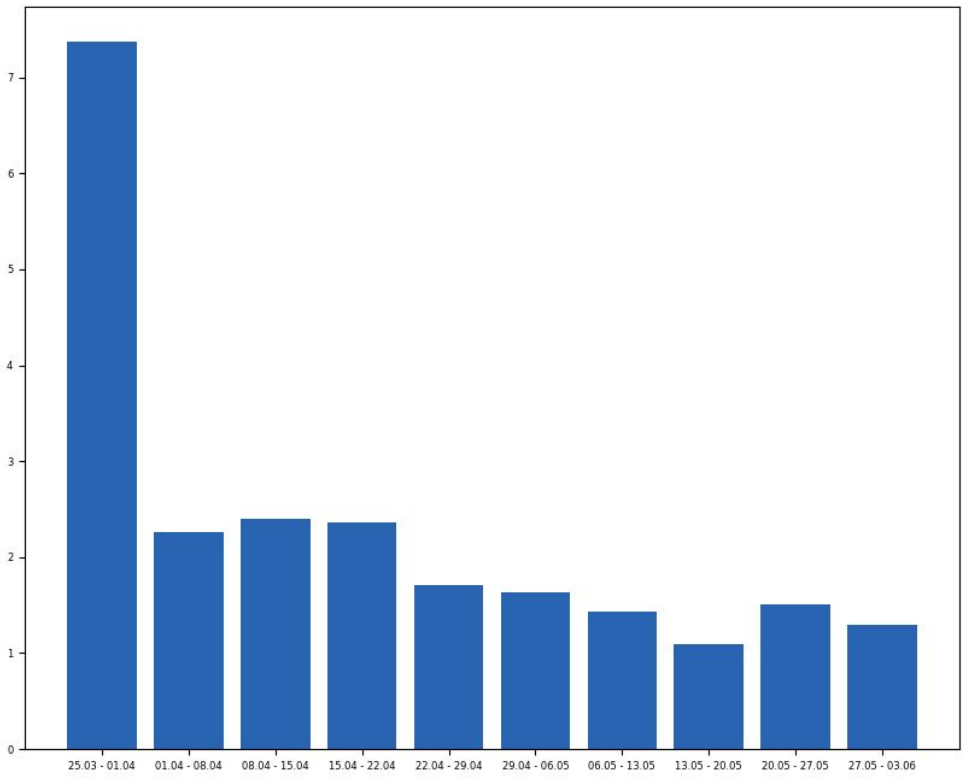
Rozpocznijmy od analizy wykresu nr. 9 - średniej liczby tweetów użytkowników w tygodniowych okresach. Możemy zaobserwować, że największa liczba tweetów w temacie koronawirusa, powstawała na samym początku, następnie malała. W okresie 25.03 - 1.04 osiągała wartość powyżej pięciu, by zmaleć w okresie 13.05 - 20.05 zmaleć do około połowy tweeta. Następnie nieco się odbiła do poziomu około jednego tweeta.

Wysokie wyniki na początku zgadzają się z dużym zainteresowaniem tematem na początku, które zaczęło z czasem naturalnie maleć wraz z przyzwyczajeniem się społeczeństwa do tematu. Zmiana tendencja i ponowny wzrost zainteresowaniem tematem w tygodniu 20.05 - 27.05, w którym rozmrażanie gospodarki wchodziło w kluczową fazę, a krzywa zachorowań nie spadała.



Rys. 9 - Średnia tygodniowa liczba tweetów użytkowników w temacie koronawirusa

Obserwując wykres nr. 10, możemy wysnuć podobne wnioski jak z analizy wykresu nr. 7. Jedyną dodatkową obserwacją jest utrzymywanie się podobnej liczby średniej lajków między 1.04 - 22.04, co może sugerować, że po początkowym bardzo wysokim zainteresowaniu, uwaga utrzymywała się na stałym, nieco niższym poziomie.



Rys. 10 - Średnia liczba lajków pod tweetem w temacie koronawirusa w tygodniowym okresie

5.5 Klastrowanie

5.5.1 Cele

Początkowa hipoteza przy klastrowaniu brzmiała następująco:

Wyróżniamy 3 grupy użytkowników:

1) 5-10% użytkowników, wysoki wpływ, dużo retweetów oraz polubień

2) ~20% użytkowników, średni wpływ, często duża liczba tweetów, mniejsze środowiska, mniejsza liczba reakcji

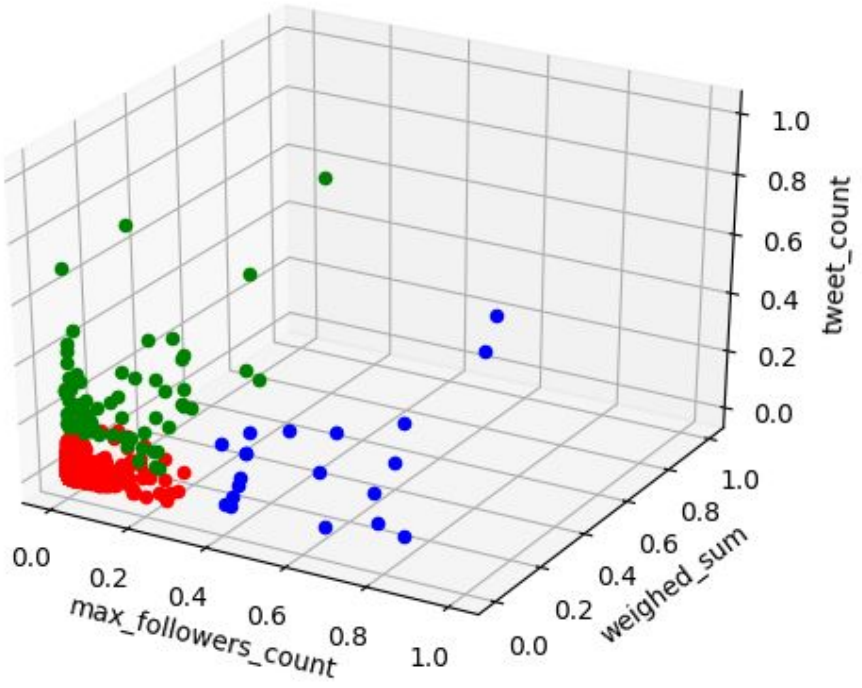
3) ~70% użytkowników, niski wpływ, małe zaangażowanie wokół tweetów takiego użytkownika

Podjęliśmy kilka prób klastrowania ze względu na różne metryki, metody, liczbę klastrów i sposób agregowania użytkowników.

Zaplanowaliśmy następujące scenariusze:

- Klastrowanie wysokopoziomowe - zaprojektowane dla małej liczby klastrów i podstawowych metryk w celu wychwycenia najbardziej istotnych grup, które dalej można dzielić na podgrupy.
- Klastrowanie niskopoziomowe - zaprojektowane dla większej liczby mniejszych klastrów w celu wychwycenia mniejszych podgrup klastrów wyłonionych w klastrowaniu wysokopoziomowym.
- Klastrowanie według metryk charakteryzujących grupy - zaprojektowane w celu utworzenia klastrów, w których znajduje się najwięcej użytkowników z tej samej kategorii. W tym celu zbadaliśmy które metryki najlepiej charakteryzują dane kategorie, a następnie użyliśmy ich przy klastrowaniu.
- Klastrowanie metodą EM - zaprojektowane w celu zbadania alternatywnego podziału zbioru niż poprzez typowe KMeans używane w pozostałych scenariuszach.
- Klastrowanie tygodniowe całościowe - klastrowanie z uwzględnieniem czasu, gdzie tweety użytkownika z każdego tygodnia są traktowane jako osobny punkt danych. Miało na celu scharakteryzowanie wyników możliwych do osiągnięcia w przeciągu krótkiego okresu.
- Klastrowanie tygodniowe dynamiczne - zbiór klastrowań z uwzględnieniem czasu, gdzie dane z każdego tygodnia klastrowane są w identyczny sposób w celu przebadania ich zmienności w czasie.

5.5.2 Klastrowanie wysokopoziomowe



Rys. 11 - Wyniki użytkowników z poszczególnych klastrów w metrykach uwzględnianych w klastrowaniu

W pierwszej kolejności użyliśmy trzech metryk, które wydawały się wnosić dużo niezależnych informacji:

- Liczba followersów - pozwala oddzielić użytkowników popularnych obecnie lub w przeszłości od takich, którzy popularności nigdy nie zdobyli. Wysoka wartość tej metryki może wskazywać na użytkowników o szerokiej grupie docelowej i popularnych poza Twitterem.
- Suma ważona polubień i retweetów - pozwala zidentyfikować użytkowników cieszących się dużą popularnością i aktywnością fanów
- Liczba tweetów - pozwala oddzielić użytkowników, którzy próbują zdobyć wpływowość poprzez wysoką aktywność (być może portale informacyjne, dziennikarze) od użytkowników mniej aktywnych lub stawiających na jakość (np. satyryków)

Do metryk które także rozważaliśmy w pierwszej fazie należały:

- Liczba followowanych - identyfikacja użytkowników, którzy próbują zdobyć followersów poprzez wzajemne followowania

- Betweenness centrality - wyróżnia użytkowników, którzy potencjalnie mogą trafić do największej liczby osób

Na rysunku 11 przedstawiona jest wizualizacja klastrowania użytkowników na 3 klastry używając trzech metryk wyłonionych w poprzednim punkcie.

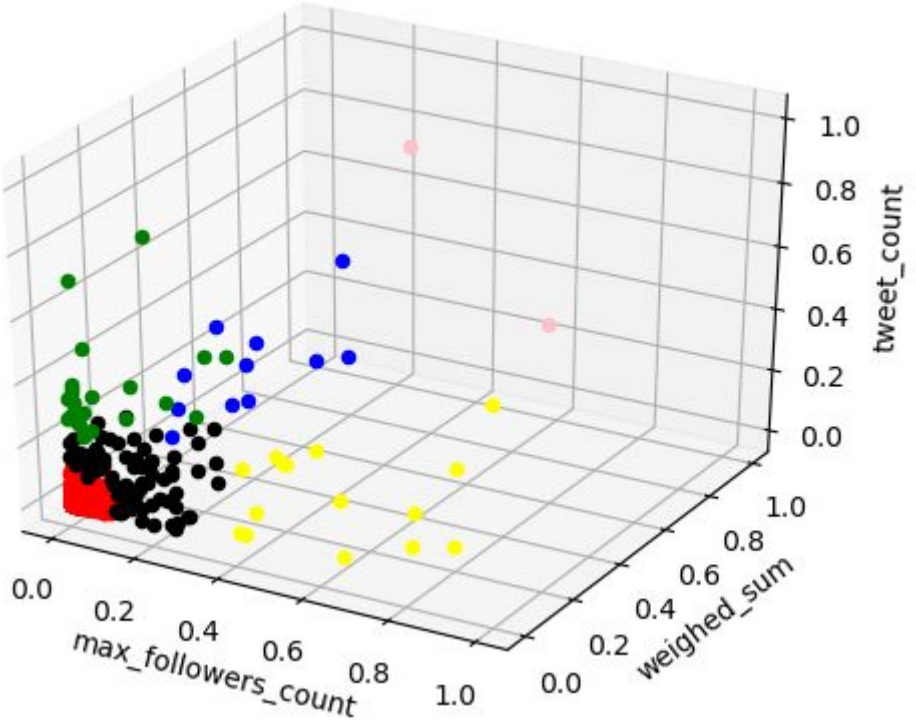
Statystyki:

- Grupa: 0, kolor: czerwony, procent całości : 99.63759263783696%
- Grupa: 1, kolor: niebieski, procent całości: 0.08143985666585227%
- Grupa: 2, kolor: zielony, procent całości: 0.2809675054971903%

Już na tym etapie widać, że prawie wszyscy użytkownicy zostali umieszczeni w jednej grupie, a tylko użytkownicy o wyjątkowo dobrych wynikach stworzyli dwa pozostałe klastry.

5.5.3 Klastrowanie niskopoziomowe

Aby uzyskać większą szczegółowość podziału przeprowadziliśmy klastrowanie jeszcze raz, ale dla 6 klastrów. Wyniki przedstawiono na rysunku 12.



Rys. 12 - Wyniki użytkowników klastrowanych wg. sumy ważonej, liczby followersów i liczby tweetów na 6 grup

color	count	percentage	top_50	top_10	p	m	d	i
red	26540	99,5	1,0	0,2	0,1	0,0	0,0	0,1
blue	11	0,0	100,0	100,0	27,3	18,2	18,2	36,4
green	24	0,1	100,0	33,3	0,0	41,7	8,3	41,7
black	83	0,3	66,3	9,6	15,7	8,4	9,6	20,5
yellow	15	0,1	100,0	53,3	26,7	26,7	33,3	13,3
pink	2	0,0	100,0	100,0	0,0	100,0	0,0	0,0

color	max_followers_count				weighed_sum				tweet_count			
	mean	min	max	std	mean	min	max	std	mean	min	max	std
red	1723	0	139884	6041	463	0	137110	3038	24	6	775	42
blue	108987	16731	251250	93995	807927	481030	1376957	276503	1045	259	1955	493
green	93390	525	429489	128450	59929	2689	333858	85771	2310	1420	5327	881
black	149872	1026	376853	110707	96572	195	417736	110435	580	6	1374	416
yellow	775056	489695	1133132	215967	218608	1596	877208	225955	756	53	1892	525
pink	1057060	695265	1418855	361795	520897	281726	760068	239171	5550	4807	6292	743

Rys. 13 - Statystyki klastrów i średnie wyników uzyskiwanych w poszczególnych metrykach dla każdego klastra

Tabela na rysunku 13 (podzielona dla czytelności na dwie) opisuje zidentyfikowane klastry. Każdy wiersz reprezentuje jeden klaster, a każda kolumna pewną cechę. Kolejne kolumny:

- color - kolor, którym reprezentowani są użytkownicy z danego klastra na wykresie punktowym,
- count - liczba użytkowników należąca do klastra,
- percentage - procent populacji należący do klastra,
- top_50 - procent użytkowników należących do tego klastra, którzy uzyskali w rankingu którejkolwiek ze zdefiniowanych przez nas miar pozycję z przedziału 1-50,
- top_10 - procent użytkowników należących do tego klastra, którzy uzyskali w rankingu którejkolwiek ze zdefiniowanych przez nas miar pozycję z przedziału 1-10,
- p, m, d, i - procent użytkowników należących do tego klastra, który stanowią politycy, media, dziennikarze i influencerzy. Należy pamiętać, że tylko niewielkiej części użytkowników przypisaliśmy kategorię, gdyż był to proces manualny. Proporcje te pozwalają określić, jacy użytkownicy znaleźli się w danym klastrze,
- następnie dla każdej z miar uwzględnionych w klastrowaniu wypisane są cztery kolumny Wartości pozwalają na zbadanie jakie wartości miar charakteryzowały ten klaster. Przykładowo dla miary liczby followersów obliczone są:

- mean - średnia liczba followersów użytkowników należących do danego klastra,
- min, max - minimalna/maksymalna liczba followersów użytkowników należących do danego klastra,
- std - odchylenie standardowe liczby followersów użytkowników należących do danego klastra

5.5.4 Identyfikacja klastrów

Dla ułatwienia dalszej analizy przy kolejnych klastrowaniach wprowadzamy następujące nazwy klastrów (w nawiasie kolor na powyższym wykresie, w kolejnych klastrowaniach kolory nie są zachowane). Zauważamy, że ich charakterystyka pozwala podzielić je na dwie nadgrupy:

1. Klastry o równomiernych wynikach w metrykach:
   - **użytkownicy zwykli** (klaster czerwony) - największy klaster stanowiący zdecydowaną większość grupy, osiągający najniższe wartości w każdej z metryk klastrowania zarówno pod względem średniej, wartości minimalnej i maksymalnej oraz odchylenia standardowego. Już na tym etapie można zinterpretować tę grupę jako ogon społeczności, której rozkład zachowuje się zgodnie z power law
   - **użytkownicy ponadprzeciętni** (klaster czarny) - druga najliczniejsza grupa. Osiągają oni wyniki wyższe niż użytkownicy przeciętni, jednak niemające wybijających się wyników w którejś z miar.
   - **użytkownicy ekstremalni** (klaster różowy) - bardzo mała, w tym przypadku zaledwie dwuelementowa grupa użytkowników, którzy w każdej z miar klastrowania osiągnęli wyniki ekstremalnie wysokie. Średnia wyników w każdej z miar klastrowania jest najwyższa ze wszystkich klastrów.
2. Klastry wyróżniające się w jednej z metryk:
   - **użytkownicy aktywni** (klaster zielony) - użytkownicy, którzy dodali najwięcej tweetów. W statystykach tego klastra widać, że średnia liczba tweetów w badanym okresie wynosiła aż 2310 i była to najwyższa średnia ze wszystkich klastrów poza klastrem ekstremalnym. Co ciekawe, widać wysoki odsetek mediów i influencerów.
   - **użytkownicy popularni** (klaster żółty) - użytkownicy, którzy mają najwięcej followersów, średnio 775 tysięcy - najwięcej ze wszystkich klastrów poza ekstremalnym. Widać bardzo wyraźną granicę pomiędzy tym klastrem a pozostałymi, gdyż użytkownik o najmniejszej liczbie followersów należący do tej grupy ma więcej followersów niż którykolwiek z użytkowników nienależących do tej grupy (nie licząc ekstremalnej).

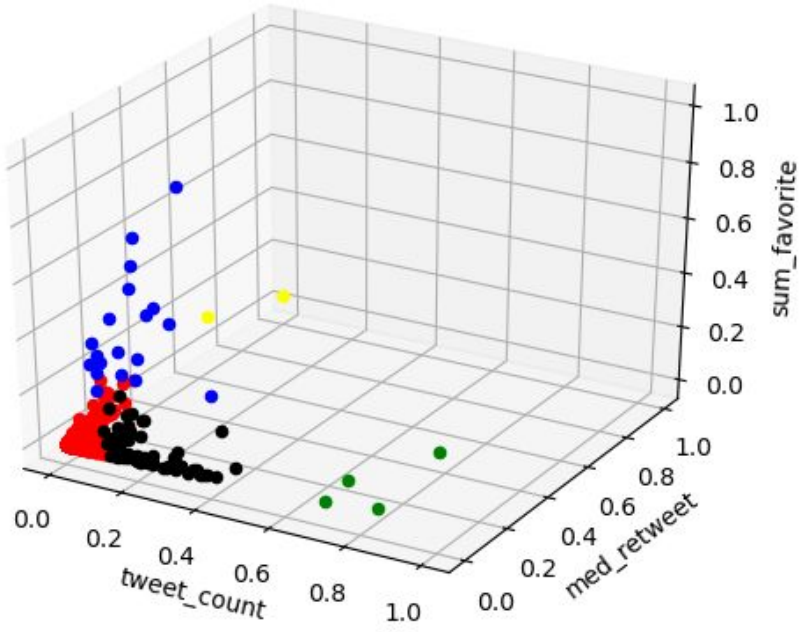
- **użytkownicy wpływowi** (klaster niebieski) - użytkownicy, którzy uzyskali największe wartości sumy ważonej liczby polubień i retweetów. W tym przypadku jednak nie była to jedyna grupa osiągająca wysokie wyniki w tej mierze, gdyż było dużo użytkowników z grupy popularnej, którzy także osiągali wyniki wysokie.

Z pierwszych prób klastrowania wyciągnęliśmy wniosek, że prawidłowo udaje się rozpoznać pewne grupy użytkowników i łatwo można uzasadnić ich odrębność. Szczególnie ważną cechą jest fakt, że grupa aktywna wyodrębniając użytkowników według dużej liczby tweetów jednocześnie wyodrębniła media i influencerów. Odczytaliśmy to jako pierwszy znak że zbliżamy się do wyodrębnienia klastrów według typu użytkowników.

5.5.5 Klastrowanie według metryk charakteryzujących grupy

Aby wzmocnić efekt wyodrębnienia użytkowników z danej kategorii, w kolejnej próbie obliczyliśmy średnie wyniki metryk dla każdej dużej grupy użytkowników (polityków, mediów, dziennikarzy i influencerów), a także użytkowników nieskategoryzowanych. Miało to na celu wyznaczenie tych miar, których wynik dla danej grupy najbardziej się wybijał. Przykładowo grupa polityków zdobywała w metryce “mediana retweetów“ 3 razy więcej niż dziennikarze, 4 razy więcej niż influencerzy i 12 razy więcej niż media. Spodziewaliśmy się więc, że wykorzystanie tych metryk najlepiej się sprawdzi przy klastrowaniu.

Użyliśmy więc metryk: 'tweet_count', 'med_retweet', 'sum_favorite', ‘avg_retweet_to_followers_count', które miały wyodrębnić odpowiednio grupy: mediów, polityków, dziennikarzy i influencerów. Wyniki przedstawione zostały na rysunkach 14 i 15.



Rys. 14 - Wyniki użytkowników klastrowanych na 6 grup wg. miar charakteryzujących kategorie użytkowników

color	count	percentage	top_50	top_10	p	m	d	i
red	26586	99,67	1,11	0,18	0,11	0,01	0,04	0,09
blue	19	0,07	100,00	94,74	21,05	10,53	21,05	47,37
green	4	0,01	100,00	100,00	0,00	75,00	0,00	25,00
black	63	0,24	77,78	14,29	6,35	28,57	9,52	30,16
yellow	2	0,01	100,00	100,00	50,00	0,00	0,00	0,00
pink	1	0,00	100,00	100,00	0,00	0,00	0,00	0,00

color	tweet_count				med_retweet				sum_favorite				g_retweet_to_followers_co			
	mean	min	max	std	mean	min	max	std	mean	min	max	std	mean	min	max	std
red	24,84	6,00	713,00	43,92	0,49	0,00	175,00	4,03	311,89	0,00	204032,00	3387,73	0,00	0,00	0,83	0,01
blue	952,79	69,00	2550,00	612,68	33,89	0,00	236,00	55,22	426155,63	225650,00	982217,00	187637,32	0,00	0,00	0,00	0,00
green	5215,75	4437,00	6292,00	697,18	2,75	0,00	5,00	2,28	113901,50	18851,00	288213,00	105983,97	0,00	0,00	0,00	0,00
black	1376,98	577,00	2957,00	568,75	1,95	0,00	34,00	5,39	42496,19	384,00	213783,00	53052,36	0,00	0,00	0,01	0,00
yellow	32,00	11,00	53,00	21,00	508,50	391,00	626,00	117,50	94783,00	39460,00	150106,00	55323,00	0,01	0,00	0,01	0,01
pink	11,00	11,00	11,00	0,00	0,00	0,00	0,00	0,00	49,00	49,00	49,00	0,00	2,09	2,09	2,09	0,00

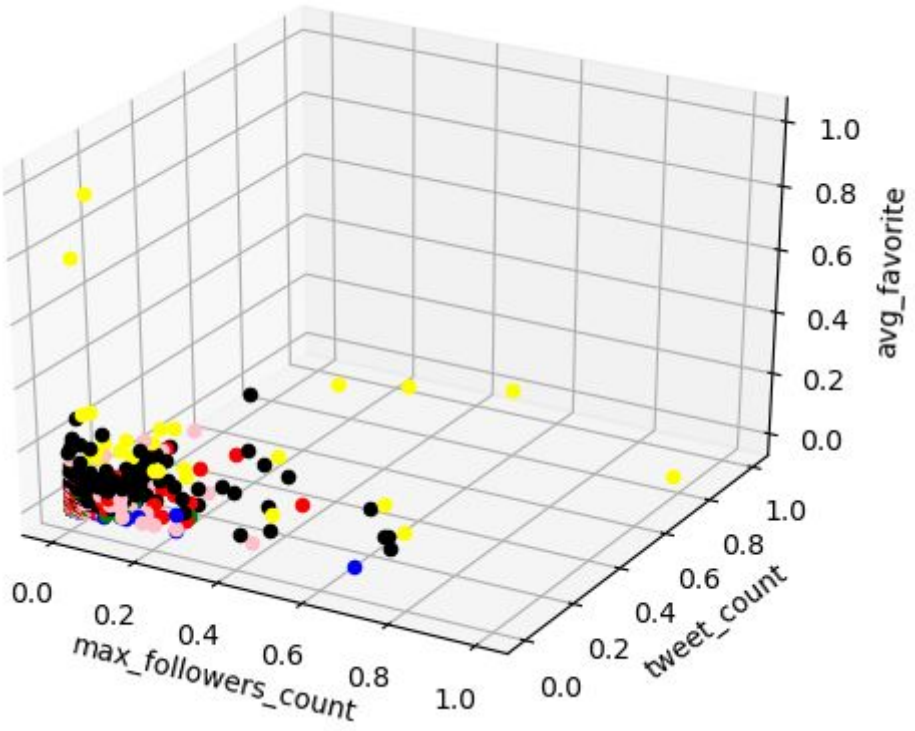
Rys. 15 - Statystyki klastrów i średnie wyników uzyskiwanych w poszczególnych metrykach dla każdego klastra

Wnioski:

- Ponownie pojawiły się klastry użytkowników zwykłych i ponadprzeciętnych (czerwony i czarny)
- Ponownie grupy niebieska, zielona, żółta i różowa maksymalizowały poszczególne metryki, jednak nie udało się osiągnąć bardziej wyrazistej charakterystyki dla żadnego z klastrów. Co więcej często klastry były bardzo małe - 4, 2 i 1 użytkownik.

5.5.6 Klastrowanie po wszystkich metrykach

Eksperymentalnie spróbowaliśmy też klastrować po wszystkich dostępnych metrykach. Wyniki przedstawiono na rysunkach 16 i 17



Rys. 16 - Wyniki użytkowników klastrowanych na 6 grup wg.wszystkich dostępnych metryk

color	count	percentage	top_50	top_10	p	m	d	i
red	4199	15,7	0,9	0,1	0,0	0,0	0,1	0,1
blue	12149	45,5	0,4	0,1	0,0	0,0	0,0	0,0
green	3472	13,0	0,8	0,1	0,0	0,0	0,0	0,1
black	631	2,4	26,5	3,8	4,3	2,5	1,9	4,6
yellow	33	0,1	100,0	87,9	30,3	15,2	12,1	30,3
pink	6191	23,2	0,8	0,2	0,0	0,0	0,0	0,1

Rys. 17 - Statystyki klastrów dla klastrowania wg. wszystkich dostępnych metryk

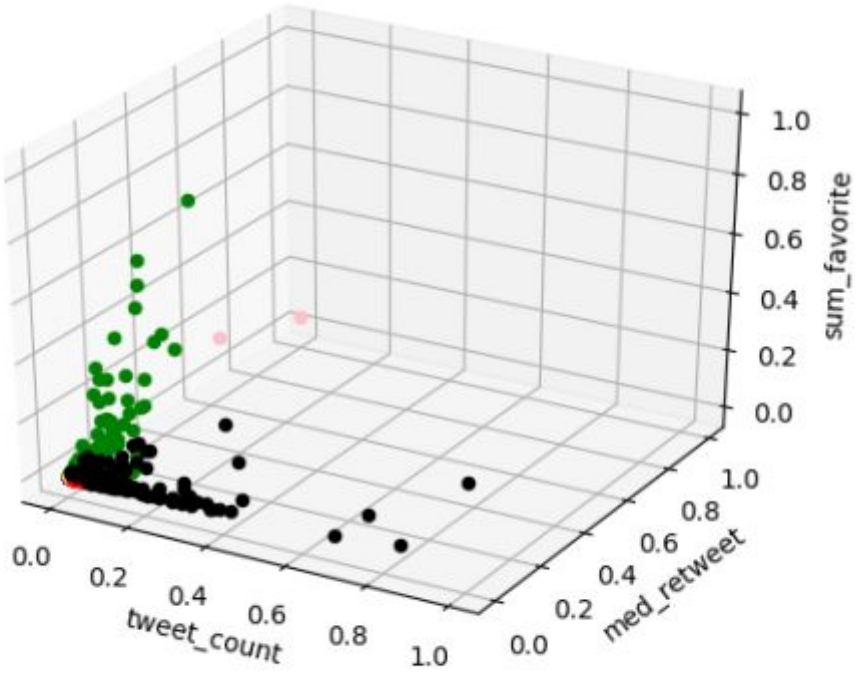
Nie dało to jednak zadowalającego rezultatu. Wcześniej wyodrębniona grupa zwykłych użytkowników została podzielona na kilka i w efekcie teraz tylko grupa żółta i częściowo czarna zawierała użytkowników o dobrych wynikach.

5.5.7 Klastrowanie metodą EM

Postanowiliśmy zastosować inną metodę klastrowania - *Expectation–maximization algorithm*, który jest metodą iteracyjną, mającą na celu znalezienie maksymalnego prawdopodobieństwa oszacowanych parametrów w modelu statystycznym, gdzie model zależy od zmiennych ukrytych[**EM]**.

Do klastrowania zostały użyte następujące metryki: Użyliśmy więc metryk: 'tweet_count', 'med_retweet', 'sum_favorite'.

Wyniki przedstawiono na rysunkach 18 i 19.



Rys. 18 - Wyniki użytkowników klastrowanych metodą EM na 6 grup wg. trzech podstawowych metryk

color	count	percentage	top_50	top_10	p	m	d	i
red	2 562	9.42	001.29	0.12	0.0	0.08	0.0	0.04
blue	23 655	87.02	0.23	0.04	0.0	0.0	0.0	0.01
green	129	0.47	81.39	26.36	24.80	2.33	009.30	16.28
black	221	0.81	37.10	7.69	1.36	9.50	004.07	10.86
yellow	615	002.26	4.39	0.33	0.49	0.0	0.0	0.81
pink	2	0.01	100.0	100.0	50.0	0.0	0.0	0.0

\

color	tweet_count				med_retweet				sum_favorite			
	mean	min	max	std	mean	min	max	std	mean	min	max	std
red	94.71	6.0	339.0	59.20	0.54	0.0	5.0	001.02	493.40	0.0	5798.0	750.79
blue	14.70	6.0	51.0	9.88	0.07	0.0	43866.00	0.30	59.98	0.0	4779.0	202.86
green	256.16	6.0	2103.0	373.78	41.72	0.0	236.0	41.62	93889.55	674.0	1038916.0	170299.38
black	783.65	9.0	6672.0	879.55	0.71	0.0	5.0	001.15	20209.31	5.0	318777.0	43383.97
yellow	25.37	6.0	119.0	21.36	009.07	0.0	41.5	7.48	2148.31	10.0	13455.0	2469.20
pink	34.5	11.0	58.0	23.5	510.25	394.5	626.0	115.75	100656.0	39460.0	161852.0	61196.0

Rys. 19 - Statystyki klastrów dla klastrowania metodą EM na 6 grup wg. trzech podstawowych metryk

Główna różnica między klastorwaniem k-średnich, a klastrowniem EM przy tych samych parametrach jest taka, że w klastrowaniu metodą EM udało się wyodrębnić z

największej grupy, jedną grupę, która zawiera 10% użytkowników. Ta grupa zdobyła nieco lepsze wyniki w metrykach niż największa grupa, a wśród jej użytkowników mało jest skategoryzowanych użytkowników oraz użytkowników, którzy znaleźli się w topie którejkolwiek z metryk. Interesującym wyodrębnieniem z największej grupy jest grupa żółta, która zawiera użytkowników o stosunkowo dużej sumie lajków, a niewyróżniających się innymi metrykami.

Grupa zielona cechuje się tym, że zawiera użytkowników o największej średniej sumy lajków, a grupa czarna - liczbie tweetów oraz medianie retweetów. Grupa różowa jest zbyt mała, by ją rozważać.

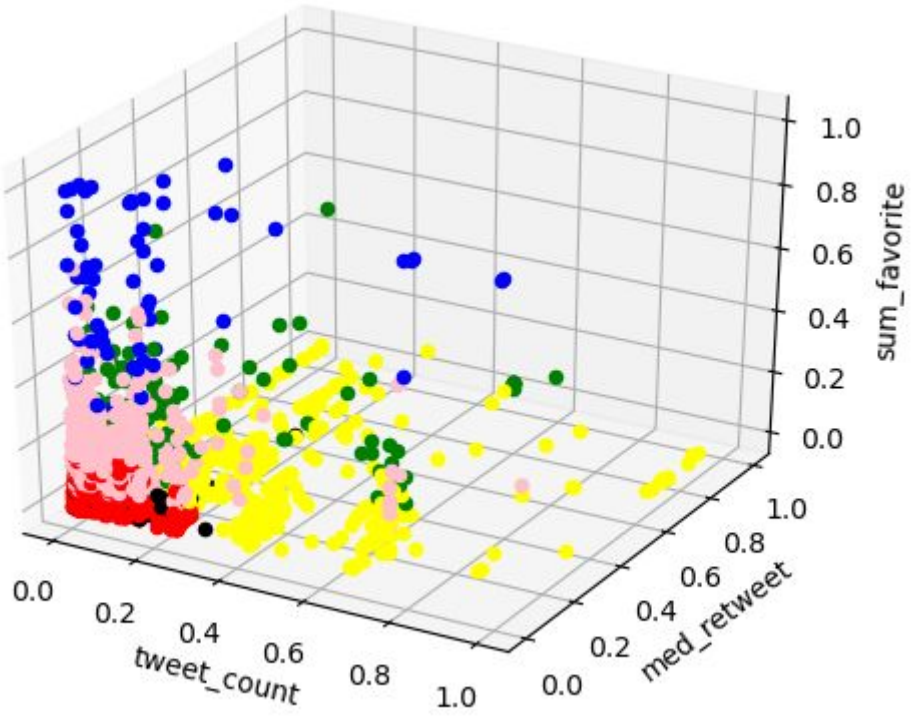
Wyniki klastrowania dla tej metody nie różnią się znacznie od metody k-średnich, więc nie pracowaliśmy dalej na tą metodą

5.5.8 Klastrowanie tygodniowe całościowe

Aby dokładniej zbadać zbiór kolejnym krokiem była próba klastrowania z uwzględnieniem czasu. W pierwszym kroku chcieliśmy przebadać czy wystąpiły wyróżniające się tygodniowe okresy, gdzie dany użytkownik uzyskiwał wyniki o określonej charakterystyce. W tym celu obliczyliśmy wartości metryk dla danego użytkownika dla każdego tygodnia osobno. Taki wynik nazywamy użytkownikotygodniem (wynikiem danego użytkownika w danym tygodniu). Każdy użytkownikotydzień traktowaliśmy jako osobny punkt w klastrowaniu. Klastrowanie przeprowadziliśmy jedno, dla wszystkich użytkownikotygodni zbiorczo. Oznacza to, że nie służy ono do zbadania dynamiki danego użytkownika bądź klastra w czasie, a do sprawdzenia jakie wyniki można było osiągnąć w krótkim terminie.

Użyliśmy ponownie metryk wyznaczonych obliczeniami jako najbardziej charakterystyczne: tweet_count, med_retweet, sum_favorite, avg_retweet_to_followers_count.

Wyniki przedstawiono na rysunkach 20 i 21.



Rys. 20 - Wyniki użytkownikotygodni klastrowanych na 6 grup wg.wszystkich dostępnych metryk

color	count	percentage	top_50	top_10	p	m	d	i
red	50207	95,3	1,7	0,2	0,1	0,2	0,1	0,3
blue	63	0,1	100,0	79,4	57,1	9,5	6,3	22,2
green	112	0,2	99,1	88,4	21,4	17,9	30,4	27,7
black	1450	2,8	13,7	3,4	0,3	1,0	0,0	1,1
yellow	431	0,8	97,2	37,4	14,2	37,6	15,3	26,9
pink	420	0,8	71,9	21,9	24,5	1,0	12,6	17,9

color	tweet_count				med_retweet				sum_favorite				avg_retweet_to_followers_coun			
	mean	min	max	std	mean	min	max	std	mean	min	max	std	mean	min	max	std
red	3410,2	0,0	376853,0	16687,0	9,0	1,0	248,0	14,0	7,2	0,0	834,0	27,2	1,5	0,0	149,4	5,7
blue	230549,6	3738,0	1134498,0	300787,9	21,4	1,0	129,0	21,7	1631,0	550,6	5944,0	1207,6	234,9	37,8	843,0	192,7
green	263982,1	14171,0	1005660,0	311405,7	152,1	21,0	767,0	160,0	510,7	44,7	1512,6	291,6	64,3	12,0	247,7	44,6
black	9856,1	162,0	358334,0	27056,0	15,8	1,0	219,0	21,5	16,5	0,0	447,5	36,6	3,8	0,0	81,4	7,0
yellow	389974,8	221,0	1418855,0	344440,2	175,1	1,0	782,0	146,9	75,8	0,1	653,3	108,3	8,6	0,0	73,1	12,6
pink	110867,4	238,0	1055358,0	175067,2	19,1	1,0	100,0	19,2	462,2	49,3	2565,5	323,7	83,2	6,0	471,0	62,9

Rys. 21 - Statystyki klastrów dla klastrowania według tygodni

- Ponownie zostały wyodrębnione grupy użytkowników zwykłych i ponadprzeciętnych
- Pozostałe klastry zachowały się nieco inaczej niż wcześniej. Zamiast maksymalizować każdą z metryk w jednym klastrze, tym razem użytkownicy uzyskujący dużo polubień zostali podzieleni na dwie grupy (różową o niższych wynikach i niebieską o wyższych). Grupa niebieska ma wysoką zawartość

polityków i cechuje się wysokim wynikem metryki avg_retweet_to_followers_count.
- Odbyło się to kosztem reprezentowania metryk tweet_count i med_retweet - użytkownicy o wysokich wynikach w którejkolwiek z tych metryk zostali umieszczeni w jednym klastrze (żółtym). Grupa ta ma wysoki odsetek mediów.

5.5.9 Klastrowanie tygodniowe dynamiczne

Przeprowadziliśmy też klastrowanie każdego z tygodni osobno, aby zobaczyć czy i jak zmieniał się rozkład społeczności w czasie. Każdy tydzień klastrowaliśmy w ten sam sposób: z użyciem metody KMeans dla sześciu klastrów i z użyciem czterech metryk: tweet_count, avg_favorite, med_retweet i avg_retweets_to_followers_count.

Wyniki tych klastrowań znajdują się na rysunku 22. Pierwszym spostrzeżeniem było to, że skoro każdy tydzień był klastrowany osobno, to tylko część klastrów powtarzała się stale w każdym tygodniu. Były to klastry charakteryzujące społeczność:
- użytkownicy zwykli,
- użytkownicy ponadprzeciętni

przy czym w jednym z tygodni nie został wyodrębniony klaster użytkowników ponadprzeciętnych.

Pozostałe cztery klastry często trudno było jednoznacznie scharakteryzować, dlatego nie przypisaliśmy im jednoznacznych nazw. Tam, gdzie było to możliwe, utrzymywaliśmy nazewnictwo z punktu 5.5.4. Często występowały klastry o następującej charakterystyce:
- użytkownicy wpływowi i skuteczni (o wysokim wyniku metryk sum_favorite i avg_retweets_to_followers_count),
- użytkownicy ekstremalni,
- użytkownicy aktywni i wpływowi (o wysokim wyniku metryk tweet_count i med_retweet)
- użytkownicy ponadponadprzeciętni - grupa charakteryzująca się względnie wysokimi wynikami wielu metryk, ale nie osiągająca najlepszych wyników w żadnej z nich. Nazwa oczywiście pochodzi od grupy ponadprzeciętnej, w celu zaznaczenia że są to jeszcze lepsi użytkownicy, ale podobnie jak tamta grupa nieosiągający wyprofilowanych wyników.

Klastry te jednak nie pojawiały się w każdym tygodniu. W niektórych klastrowaniach grupy o wysokim wyniku w dwóch metrykach były podzielone na dwie grupy o wysokim wyniku w jednej z nich.

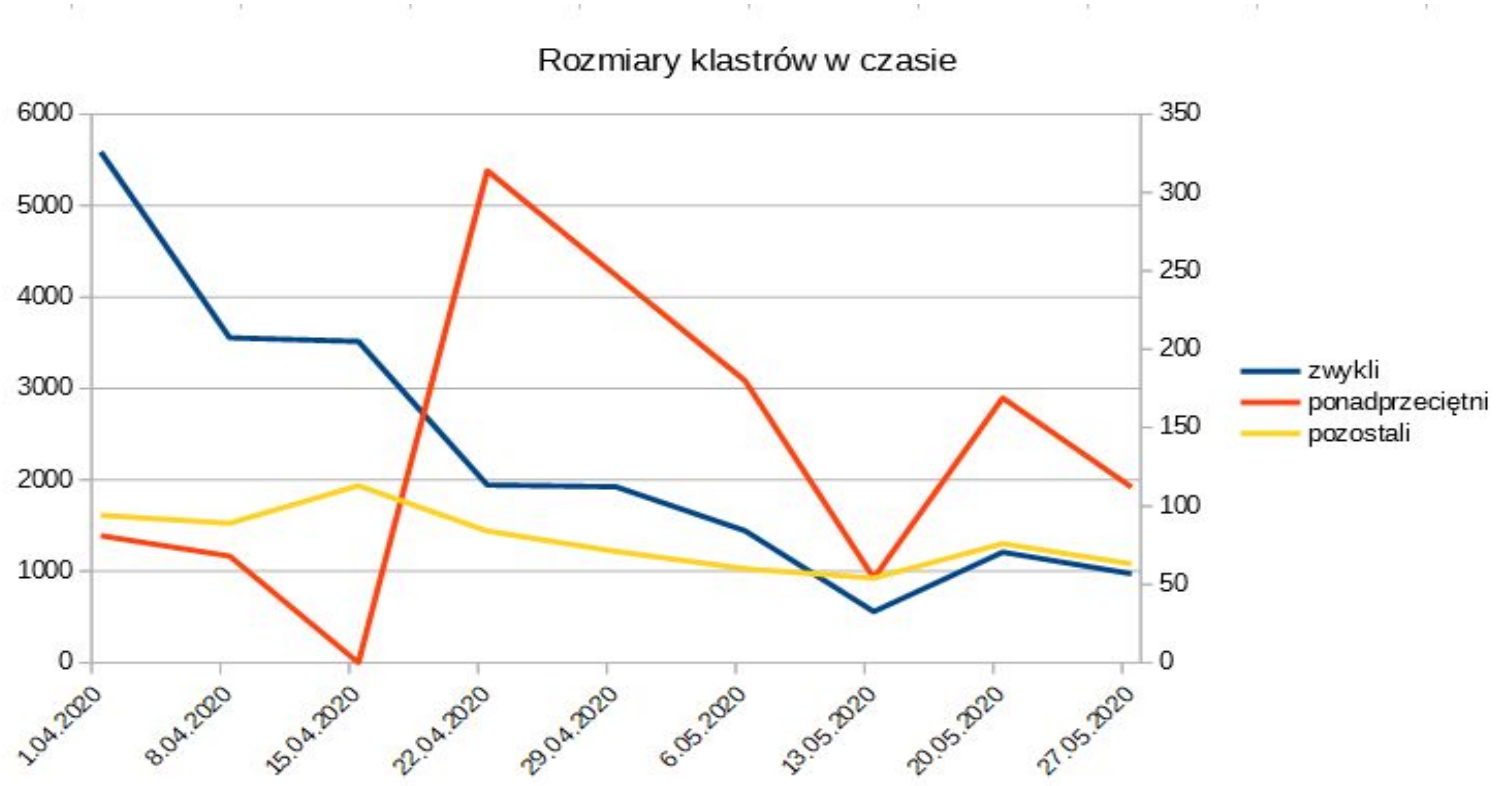
Na rysunku 23 przedstawione są rozmiary poszczególnych klastrów w kolejnych tygodniach. Przedstawiono rozmiary klastrów powtarzających się we wszystkich klastrowaniach, a pozostałe zostały połączone w grupę “pozostali”. Jest to spowodowane tym, że pozostałe klastry nie pojawiały się w każdym klastrowaniu i próba ich przedstawienia zmniejszyłaby klarowność rysunku.

Jak widać liczebność wszystkich klastrów zdecydowanie spadła w czasie. Oznacza to, że temat koronawirusa był najbardziej interesujący na początku okresu badań, a później wraz z

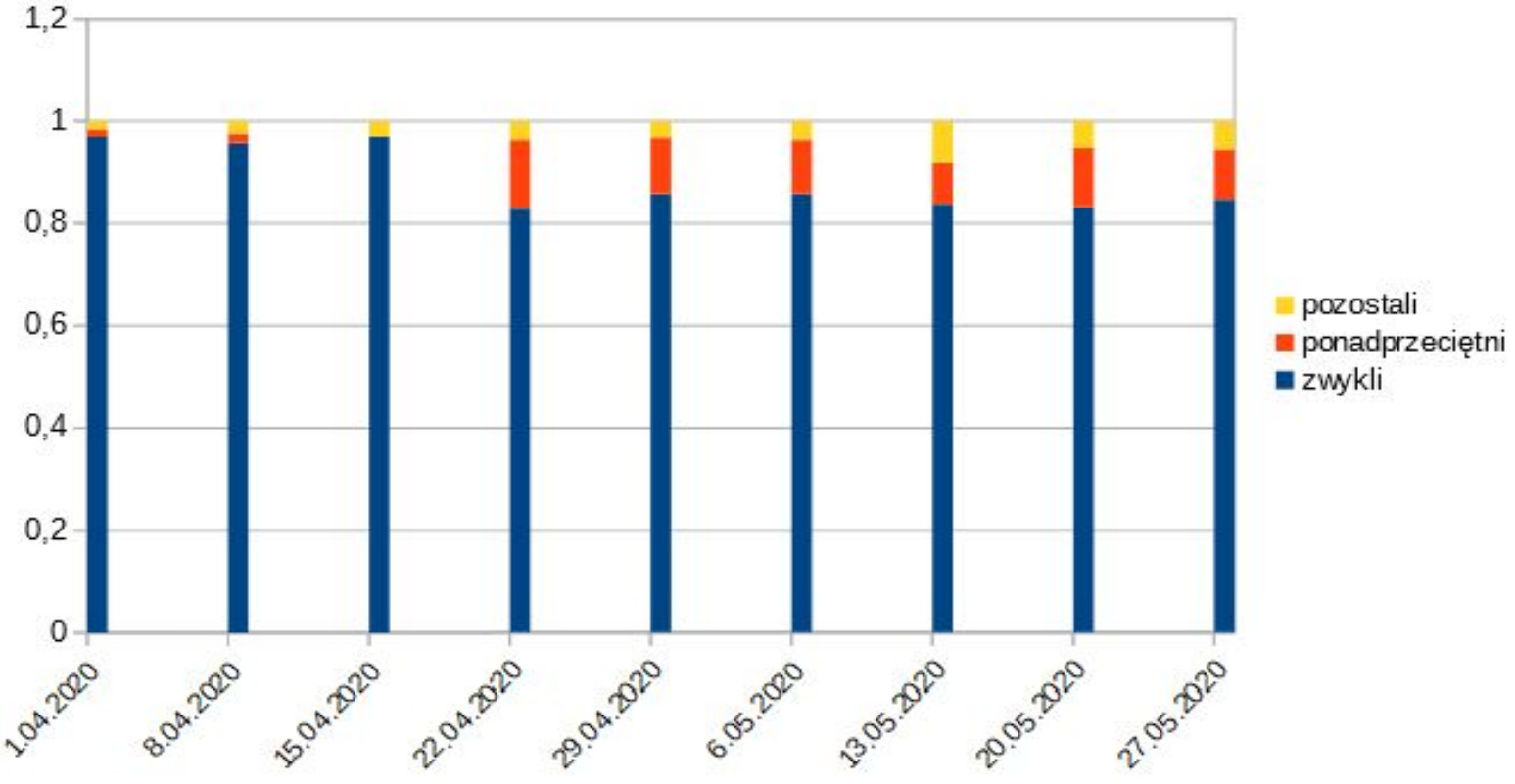
czasem jego popularność zdecydowanie spadała. Aby lepiej zobrazować proporcje rozmiarów klastrów na rysunku 24 zamieszczone są proporcjonalne rozmiary klastrów.

	count	percent	top		p	m	d	i	tweet_count				med_retweet				sum_favorite				avg_retweet_to_followers			
			50	10					mean	min	max	std	mean	min	max	std	mean	min	max	std	mean	min	max	std
01.04.2020 - 08.04.2020																								
zwykli	5589	96,96	2,7	0,4	0,2	0,1	0,1	0,2	**3118**	0	348396	14932	**14**	6	118	13	**5**	0	187	13	**1**	0	38	3
ponadprzecietni	81	1,41	48,1	11,1	13,6	3,7	7,4	14,8	**60553**	1838	376684	79469	**30**	6	145	32	**223**	69	511	108	**41**	9	127	22
ekstremalni	7	0,12	100,0	57,1	28,6	14,3	0,0	57,1	**217647**	21554	903066	290133	**164**	85	292	73	**530**	193	1061	259	**71**	23	154	41
aktywni	18	0,31	100,0	44,4	22,2	33,3	22,2	22,2	**698116**	257624	1413687	291598	**202**	13	749	207	**117**	3	354	115	**14**	0	53	16
retweet	59	1,02	66,1	22,0	6,8	18,6	6,8	30,5	**63450**	111	428796	98928	**212**	95	782	129	**17**	0	113	29	**2**	0	13	3
favorite/skuteczni	10	0,17	80,0	50,0	50,0	20,0	0,0	30,0	**97836**	18540	235961	80163	**19**	7	45	12	**757**	264	1230	258	**154**	38	347	87
08.04.2020 - 15.04.2020																								
zwykli	3555	95,77	4,3	0,7	0,4	0,1	0,3	0,4	**3582**	0	347759	16360	**13**	6	151	13	**6**	0	165	14	**1**	0	29	3
ponadprzecietni	68	1,83	30,9	10,3	10,3	1,5	4,4	13,2	**49661**	238	558583	92578	**26**	6	97	23	**214**	68	645	116	**39**	6	103	18
ekstremalni	9	0,24	100,0	44,4	11,1	22,2	0,0	66,7	**250542**	16749	903575	306254	**208**	76	677	180	**444**	45	648	178	**59**	13	121	34
aktywni/retweet	49	1,32	89,8	36,7	18,4	26,5	12,2	30,6	**254240**	515	1414710	338854	**223**	18	629	141	**36**	1	380	66	**4**	0	40	8
favorite/skuteczni	6	0,16	66,7	16,7	16,7	83,3	0,0	0,0	**115397**	20401	232262	85250	**19**	9	40	12	**1383**	1230	1598	133	**152**	66	222	53
ponadponadprzecietni	25	0,67	72,0	32,0	32,0	8,0	0,0	40,0	**137548**	2355	1052803	221237	**12**	6	45	8	**575**	312	1030	186	**100**	55	154	25
15.04.2020 - 22.04.2020																								
zwykli	3515	96,89	4,5	0,8	0,5	0,1	0,2	0,4	**4330**	0	348154	19216	**14**	6	118	12	**9**	0	468	29	**2**	0	69	6
ekstremalni	2	0,06	0,0	0,0	0,0	50,0	0,0	50,0	**50544**	20851	80236	29693	**13**	11	15	2	**3613**	2135	5090	1478	**329**	305	353	24
aktywni	14	0,39	100,0	42,9	21,4	35,7	28,6	14,3	**814015**	514907	1416065	250670	**111**	13	514	124	**126**	5	383	125	**15**	0	57	17
retweet	52	1,43	80,8	21,2	9,6	21,2	9,6	40,4	**91218**	504	429260	115351	**199**	61	660	111	**27**	1	162	37	**4**	0	20	5
retweet/favorite	7	0,19	100,0	71,4	28,6	28,6	0,0	42,9	**161972**	16916	693337	227175	**277**	63	767	240	**481**	48	1121	314	**54**	17	145	42
skuteczni/favorite	38	1,05	73,7	34,2	26,3	7,9	10,5	31,6	**112580**	446	485915	115428	**28**	6	121	29	**522**	170	1261	292	**91**	31	199	46
22.04.2020 - 29.04.2020																								
zwykli	1944	83,01	7,8	1,3	1,2	0,4	0,6	1,3	**5968**	0	376853	26857	**15**	6	265	20	**6**	0	291	19	**1**	0	38	4
ponadprzecietni	314	13,41	9,9	2,5	0,3	0,6	0,6	1,0	**5645**	271	223862	14259	**16**	6	320	26	**11**	0	190	23	**3**	0	36	5
aktywni/retweet	24	1,02	100,0	37,5	16,7	37,5	20,8	25,0	**553095**	1806	1417547	343897	**164**	6	584	167	**82**	3	381	96	**11**	0	54	14
favorite/skuteczni	10	0,43	80,0	30,0	20,0	40,0	0,0	30,0	**145650**	15657	486818	143575	**15**	6	43	11	**1135**	662	1834	299	**136**	41	211	51
ponadponadprzecietni	40	1,71	50,0	20,0	17,5	2,5	5,0	25,0	**67499**	810	339143	80544	**25**	6	105	25	**315**	68	658	146	**55**	21	112	23
ponadponadprzecietni	10	0,43	90,0	70,0	10,0	20,0	10,0	60,0	**213135**	14171	905549	303483	**150**	52	632	165	**438**	49	632	167	**46**	12	71	22
29.04.2020 - 06.05.2020																								
zwykli	1924	85,82	7,9	1,5	0,9	0,5	0,7	1,6	**5928**	0	348028	24748	**16**	6	371	24	**7**	0	346	20	**1**	0	51	4
ponadprzecietni	247	11,02	13,4	2,8	1,2	0,4	0,4	1,2	**7345**	272	223848	17736	**17**	6	277	29	**15**	0	277	31	**3**	0	48	7
aktywni	22	0,98	100,0	36,4	27,3	31,8	22,7	18,2	**555544**	163801	1417981	322060	**161**	10	530	142	**93**	2	465	115	**11**	0	48	12
favorite/skuteczni	10	0,45	70,0	50,0	20,0	30,0	0,0	30,0	**128410**	7166	238326	85852	**23**	10	49	13	**933**	602	1593	273	**137**	40	188	39
retweet	6	0,27	100,0	66,7	16,7	33,3	0,0	50,0	**287426**	14490	905951	367974	**227**	87	705	217	**479**	50	912	265	**55**	15	113	38
ponadponadprzecietni	33	1,47	72,7	24,2	27,3	9,1	6,1	30,3	**140443**	2167	1054431	200702	**30**	6	120	30	**400**	157	952	189	**56**	16	106	22
06.05.2020 - 13.05.2020																								
zwykli	1444	85,75	11,1	2,2	1,5	0,6	0,8	2,1	**7294**	0	347954	30132	**15**	6	251	20	**7**	0	341	24	**1**	0	48	4
ponadprzecietni	180	10,69	13,9	2,8	1,7	0,0	0,6	2,8	**9119**	276	223781	22998	**14**	6	188	20	**23**	0	353	50	**5**	0	59	10
aktywni/retweet	18	1,07	100,0	38,9	16,7	22,2	38,9	22,2	**616803**	179163	1418101	331251	**156**	10	516	152	**75**	3	443	119	**8**	0	51	13
favorite/skuteczni	10	0,59	70,0	10,0	20,0	40,0	0,0	40,0	**130437**	37275	238644	78561	**18**	6	42	11	**1215**	745	2204	392	**146**	38	235	53
ponadponadprzecietni	8	0,48	87,5	62,5	25,0	37,5	0,0	37,5	**369999**	15120	1002999	398929	**149**	41	577	165	**469**	47	902	275	**67**	14	197	60
ponadponadprzecietni	24	1,43	87,5	45,8	29,2	29,2	4,2	29,2	**141944**	6154	487910	136832	**81**	6	247	66	**294**	38	760	187	**38**	5	107	30
13.05.2020 - 20.05.2020																								
zwykli	558	83,78	22,4	6,5	3,8	1,8	2,2	6,1	**13933**	0	347839	43852	**16**	6	227	21	**10**	0	432	34	**2**	0	47	5
ponadprzecietni	54	8,11	37,0	9,3	7,4	1,9	3,7	5,6	**16721**	1193	183437	34618	**18**	6	160	28	**45**	0	448	86	**8**	0	85	16
ekstremalni	5	0,75	80,0	40,0	20,0	40,0	0,0	40,0	**601284**	55646	1005660	355347	**162**	41	418	137	**437**	64	753	279	**51**	19	87	28
aktywni/retweet	18	2,70	100,0	33,3	27,8	27,8	33,3	11,1	**594454**	1786	1418842	352100	**120**	7	407	117	**80**	1	436	110	**9**	0	69	16
ekstremalny	1	0,15	100,0	100,0	100,0	0,0	0,0	0,0	**253850**	253850	253850	0	**16**	16	16	0	**2428**	2428	2428	0	**388**	388	388	0
favorite/skuteczni	30	4,50	83,3	33,3	23,3	26,7	3,3	40,0	**128212**	7487	489502	115659	**32**	6	99	24	**518**	158	1513	304	**67**	12	181	41
20.05.2020 - 27.05.2020																								
zwykli	1209	83,15	11,4	2,4	1,2	0,5	0,9	2,3	**6736**	0	347837	26157	**16**	6	248	21	**8**	0	389	28	**2**	0	51	4
ponadprzecietni	169	11,62	14,8	1,2	2,4	0,6	0,0	1,8	**8732**	298	223590	23913	**15**	6	147	20	**15**	0	211	30	**3**	0	32	6
ekstremalni	2	0,14	100,0	100,0	50,0	0,0	0,0	50,0	**694809**	256485	1133132	438324	**40**	16	64	24	**1902**	1454	2350	448	**329**	248	411	82
aktywni/retweet	37	2,54	100,0	35,1	29,7	24,3	16,2	27,0	**417582**	484	1418855	344406	**167**	11	572	150	**125**	1	653	143	**15**	0	69	17
favorite	13	0,89	100,0	76,9	23,1	38,5	7,7	30,8	**184878**	16729	908066	274101	**147**	27	732	176	**652**	57	1526	415	**69**	16	202	55
skuteczni	24	1,65	70,8	20,8	20,8	20,8	12,5	29,2	**112117**	7003	354156	108821	**23**	6	70	20	**497**	182	1063	216	**88**	23	182	40
27.05.2020 - 03.06.2020																								
zwykli	972	84,74	15,7	3,3	2,3	0,6	1,2	3,6	**8898**	0	347696	31268	**15**	6	232	20	**8**	0	318	25	**2**	0	43	5
ponadprzecietni	112	9,76	12,5	3,6	1,8	0,9	0,9	0,9	**6050**	288	34490	5983	**16**	6	135	23	**23**	0	318	49	**5**	0	47	8
aktywni/retweet	21	1,83	100,0	38,1	19,0	33,3	23,8	23,8	**588328**	179677	1418705	343354	**144**	10	441	119	**96**	2	393	110	**11**	0	57	15
favorite/skuteczni	7	0,61	100,0	28,6	28,6	42,9	14,3	14,3	**318972**	42571	1134498	342415	**22**	9	46	12	**1217**	666	1678	299	**178**	36	319	85
ponadponadprzecietni	26	2,27	80,8	42,3	19,2	23,1	3,8	38,5	**128237**	7619	489674	122704	**30**	6	79	23	**394**	200	1020	206	**59**	16	114	26
retweet	9	0,78	77,8	33,3	44,4	33,3	11,1	11,1	**262477**	17210	908013	308311	**187**	80	621	171	**420**	53	707	220	**51**	13	99	30

Rys. 22 - Statystyki klastrów w czasie



Rys. 23 - Rozmiary klastrów w czasie.
Grupa zwykła jest narysowana według lewej, a pozostałe według prawej osi.



Rys. 24 - Proporcje rozmiarów klastrów w czasie

Biorąc pod uwagę dane z powyższych wykresów oraz z **[kalendarium]** można próbować wyjaśnić trendy zachodzące w okresie badań. Najbardziej gwałtowny spadek zainteresowania miał miejsce w kwietniu, mimo że to wtedy epidemia w Polsce osiągnęła swój najwyższy poziom. 19 kwietnia odnotowano bardzo wysoką liczbę 545 zakażeń, a 24 kwietnia zmarło 40 osób, co stanowi rekord w okresie badań. Powstaje więc pytanie o powód spadku liczby publikacji w tym okresie.

Wyjaśnienia można dopatrywać się we względnych proporcjach rozmiarów klastrów. Widać, że spadek liczebności dotyczył głównie klastra użytkowników zwykłych. Można to tłumaczyć tym, że kiedy temat koronawirusa był świeży, zajmował uwagę całej społeczności. Dużo zwykłych użytkowników było skłonnych wypowiadać się na jego temat, być może zaliczali się do tej grupy nawet ci użytkownicy, którzy poza pandemią w ogóle mało publikują. Było ich w początkowej fazie nawet 97%. Pod koniec kwietnia, kiedy temat już był obecny w mediach przez dwa miesiące, użytkownicy ci mogli doświadczyć przesytu tematu i z tego powodu przestać publikować tweety. Nie miało znaczenia to, że notowano kolejne rekordy, a epidemia weszła na najwyższy poziom.

Jednakże grupa użytkowników ponadprzeciętnychi z pozostałych klastrów powiększyła się. Zarówno w tym okresie jak i w maju nadal występowało dużo zdarzeń związanych z wirusem - przykładowo znoszenie obostrzeń w pierwszej połowie maja, temat drugiej fali czy temat wyborów przełożonych ze względu na epidemię. Można wnioskować, że użytkownicy ponadprzeciętni i ci zaliczający się do najlepszych klastrów, którzy świadomie próbują budować popularność, dalej publikowali treści. Poskutkowało to zwiększeniem procentowego udziału ich klastrów.

Co więcej, jeśli epidemia będzie się utrzymywała przez dłuższy czas bądź nadejdzie druga fala zachorowań, to można spodziewać się, że udział użytkowników zwykłych dalej będzie się zmniejszał. Użytkownicy ci nie będą już tak żywo zainteresowani tematem jak na początku, kiedy był on nowością. Tymczasem media i komentatorzy dalej będą publikowali w tym temacie, informacyjnie lub w kontekście innych wydarzeń.

6. Podsumowanie

6.1 Wnioski

Różne metody analizy danych pokazały, że nasz zbiór zdecydowanie ma strukturę odpowiadającą power law. Na histogramach metryk widać było ogon populacji zajmujący zdecydowaną większość histogramu oraz pik najlepszych użytkowników, wąski i wysoki. Obliczenia także potwierdzają to spostrzeżenie - średni wynik dla poszczególnych metryk był osiągany przez użytkowników znajdujących się w 90-95 centyla.

W celu dokładniejszej analizy zbioru próbowaliśmy klastrować według różnych metryk, dla różnej liczby klastrów i uwzględniając czas lub nie. W tym przypadku także udało się odróżnić bardzo dużą grupę użytkowników o słabych wynikach od użytkowników wpływowych. Największa grupa miała nawet 99,6% użytkowników, i byli to użytkownicy o najniższych wynikach metryk. Pokazało to już podstawowe klastrowanie na 3 grupy.

Zwiększenie liczby grup dało dodatkowy podział największej grupy na użytkowników zwykłych i ponadprzeciętnych, oraz podział najlepszych użytkowników na takich wyróżniających się pod względem konkretnej metryki. Jedna z grup była w dużym stopniu złożona z mediów i influencerów.

Próbowaliśmy dobierać różne zestawy metryk do klastrowania. Spodziewając się lepszego odseparowania użytkowników w zależności od reprezentowanej kategorii obliczyliśmy średnie wyniki metryk uzyskiwane przez użytkowników z danej kategorii. Na ich podstawie wybraliśmy te metryki, które dawały najbardziej wybijające się wyniki dla którejś z kategorii użytkowników. Nie dało to jednak dobrych wyników, połowa klastrów zawierała 4 użytkowników lub mniej i nie stanowiła użytecznej reprezentacji społeczności.

Spróbowaliśmy też klasteryzacji bez eliminowania żadnych metryk (to znaczy korzystając ze wszystkich jednocześnie). Co ciekawe poskutkowało to odwróceniem zależności zachodzącej przy wszystkich pozostałych próbach klastrowania. To grupa użytkowników zwykłych została podzielona na kilka, a za to użytkownicy o dobrych wynikach zostali skategoryzowani razem, niezależnie od charakteru osiąganych wyników. Nie jest to efekt pożądany gdyż interesuje nas kategoryzacja użytkowników, którzy osiągają dobre wyniki, ale było interesującym rezultatem.

W ostatnim kroku podzieliliśmy dane według tygodni i rezultaty dla każdego użytkownika z każdego tygodnia potraktowaliśmy jako oddzielne próbki przy klastrowaniu. Uzyskaliśmy w ten sposób więcej punktów o wysokich wynikach. Klastry rozłożyły się w nieco inny sposób, udało się wyodrębnić klaster o wysokiej zawartości polityków oraz drugi z wieloma profilami medialnymi.

Klastrowanie dynamiczne miało jednak ważniejszy cel, jakim była analiza stabilności klastrów. Ważnym spostrzeżeniem był zdecydowany spadek rozmiaru całego zbioru danych w czasie. Nastąpił on mimo tego, że pandemia nadal trwała pod koniec okresu badań. Można z tego można wywnioskować, że tematy w mediach społecznościowych utrzymują się tylko przez pewien czas, a okres uwagi odbiorców jest ograniczony. Społeczność była najbardziej zainteresowana tematem wtedy, kiedy wiedza na temat wirusa była niska, a obawy wysokie. W późniejszym etapie nastąpiło odwrócenie zależności: ilość nowych informacji, które mogły pobudzić społeczność spadała wraz z większym przyzwyczajeniem do nowej rzeczywistości.

Wraz ze zmniejszeniem zbioru danych zmniejszył się udział grupy zwykłych użytkowników. Może to sugerować, że tacy użytkownicy publikowali treści pod wpływem nowego trendu, jednak nie było to spowodowane chęcią długotrwałego budowania popularności. W późniejszym okresie pozostali użytkownicy publikujący regularnie, którzy niezależnie od okresu próbują zbudować wpływowość.

Podsumowując, klastrowanie pozwala wyodrębnić użytkowników z wynikami o konkretnej charakterystyce. Jednak dość trudno jest wyodrębnić samym klastrowaniem grupy zawierające wyłącznie użytkowników z konkretnych kategorii. Jest to jednak zrozumiałe, ponieważ każdy użytkownik mimo swojej kategorii ma swój własny sposób publikowania i indywidualną publiczność, niekoniecznie możliwą do łatwego zgeneralizowania.

6.2 Dalsze prace

Praca ta pozostawia wiele możliwości dalszego rozwoju w celu uzyskania bardziej szczegółowych informacji o strukturze społeczności i budowaniu wpływowości.

Jednym z naturalnych rozszerzeń pracy byłoby zwykłe powiększenie zbioru poprzez przedłużenie okresu badań. Pozwoliłoby to dostrzeżenie bardziej długoterminowych trendów. Jeśli wystąpiłaby druga fala zachorowań, można by porównać zachowanie użytkowników z pierwszej i drugiej fali. Wymagałoby to jedynie czasu, gdyż zaimplementowane przez nas narzędzia na bieżąco uzupełniają bazę o nowe tweety.

Zebrane dane można przeanalizować na wiele innych sposobów niż te zaprezentowane w pracy. Można użyć zarówno innych algorytmów i parametrów klastrowania, jak i innych narzędzi. Być może udałoby się osiągnąć takie klastrowania, w których wyodrębnione klastry w większym stopniu pokrywałyby się ze zdefiniowanymi kategoriami użytkowników.

Pomysłem nieporównywalnie bardziej wymagającym byłoby analizowanie treści tweetów. Stworzyłoby to wiele możliwości, takich jak analiza zależności wpływu od typu publikacji lub analiza sentymentu. Jest to możliwe dzięki narzędziom z dziedziny NLP (Natural Language Processing).

7. Instrukcja użytkowania

7.1 Instalacja

Aby zainstalować projekt, trzeba pobrać go z GitHuba[**GitHub**], a następnie doinstalować brakujące biblioteki. W tym repozytorium znajdują się również wszystkie dane, wykresy czy dokumentacje, które powstawały w wyniku projektu. Można to zrobić, używając następującej komendy:

```
git clone https://github.com/Wojtos/TwitterAnalisys
cd TwitterAnalisys
pip install -r requirements.txt
```

Aby uruchomić projekt trzeba najpierw stworzyć plik .env w roocie projektu i wypełnić go następującymi zmiennymi.

```
MONGO_DB_HOST=
SEARCHTWEETS_ACCESS_TOKEN=
SEARCHTWEETS_ACCESS_TOKEN_SECRET=
SEARCHTWEETS_CONSUMER_KEY=
SEARCHTWEETS_CONSUMER_SECRET=
SEARCHTWEETS_ENDPOINT=
SEARCHTWEETS_ACCOUNT_TYPE='premium'
LOG_FILE='./local.log'
```

Należy je uzupełnić podając adres hosta do bazy MongoDB oraz dane dostępowe do API pobierania twittów z Twittera. Klucze dostępu oraz endpoint Twittera można uzyskać po założeniu konta deweloperskiego [**Twitter**].

7.1 Użytkowanie

Program składa się z kilku podprogramów, które służą do pobrania danych z Twittera, wyliczeniu metryk, klastrowaniu czy generowaniu raportów.

7.1.1 Dodawanie wyszukiwania

```
python main.py --action add_search --query <query> --until <until_time>
--since_id <since_id> --lang <lang>

python main.py --action add_search --file <file.txt>  --until
<until_time> --since_id <since_id> --lang <lang>
```

Wyszukiwanie w naszym projekcie odbywa się przy użyciu fraz oraz kwantyfikatorów. Dzięki kwantyfikatorom, które są udostępniane przez Twittera, możemy wyszukiwać np. tweety konkretnych użytkowników.

W obu przykładach mamy parametry, które mówią:

- *until* - do kiedy wyszukujemy
- *since_id* - od jakiego id tweeta wyszukujemy
- *lang* - język wyszukiwanych tweetów

W pierwszym przykładzie bezpośrednio wpisujemy parametr *query,* który oznacza pojedynczą frazę, po której będziemy wyszukiwać. Natomiast, *query_file,* który jest ścieżką do pliku tekstowego, w którym znajduje się lista fraz do wyszukiwania.

7.1.2 Wyszukiwanie

```
python main.py --action search --search_id <search_id>

python main.py --action run_search
```

Wyszukiwanie polega na pobieraniu danych z API Twittera i zapisywaniu ich w bazie MongoDB. Pierwszy przykład oznacza zapisywanie tweetów na podstawie rekordu wyszukiwania, który został wcześniej zapisany w bazie, do momentu, w którym ściągnie wszystkie istniejące do tej pory tweety. Kolejny przykład, iteracyjnie wyszukuje dane na podstawie wszystkich rekordów wyszukiwania w bazie. W tym przypadku program działa do momentu wyłączenia go odpowiednim sygnałem.

7.1.3 Analiza danych, zapisanie metryk

```
python main.py --action analyse

python main.py --action analyse --period weeks|two_weeks|months
```

Tymi komendami możemy analizować metryki i zapisać te metryki w bazie. Pierwszy przykład opowiada metrykom na podstawie wszystkich tweetów użytkowników, a drugi na podstawie danych podzielonych na okresy: tygodniowe, dwutygodniowe oraz miesięczne.

7.1.4 Wygenerowanie raportu

```
python main.py --action save_reports
```

Ta komenda generuje raporty oraz wykresy na podstawie metryk zapisanych już w bazie.

7.1.5 Klastrowanie

```
python main.py --action cluster --amount <clusters_amount> --method
k-means|em
```

W klastrowaniu możemy wybrać metody: k-średnich lub EM. Kolejnym parametrem, który możemy uzupełnić jest liczba klastrów. W wyniku tej komendy generowany jest wykres klastrowania oraz użytkownikom przypisywane są klastry w bazie danych.

Bibliografia

- **EM** - Opis algorytmu Expectation–maximization - Expectation–maximization algorithm, Wikipedia, 2020
- **GitHub** - Link do naszego projektu na githubie, Antoni Pięta, Wojciech Gruszka, 2020
- **Twitter -** Link do uzyskania dostępu do konta deweloperskiego na Twitterze, 2020
- **sklearn -** biblioteka, z której użyliśmy metod do klastrowania, 2020
- **TwitterSearchAPI -** dokumentacja opisująca pobieranie tweetów, 2020
- **numpy** - biblioteka obliczeniowa, która specjalizuje się w operowaniu na złożonych strukturach danych, 2020
- **statistics** - dokumentacja biblioteki, która służy do wyliczenia prostych statystyk, 2020
- **kalendarium -** kalendarium koronawirusa w Polsce, Redakcja Onetu, 2020
- **kshellPaper -** praca na temat analizy wpływowości na twitterze, używajac dekompozycji K-shell - Measuring User Influence on Twitter Using Modified K-Shell Decomposition, Philip E. Brown, Junlan Feng, 2011
- **timeInfluencePaper -** praca skupiająca się na zdobywaniu wpływowości. - Measuring User Influence in Twitter: The Million Follower Fallacy, Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi, 2010
- **clusterAlgorithms** - opis algorytmów do klastrowania - Clustering Algorithms, Google, 2020
- **coronavirusImpact** - Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-Based Study and Research Directions, A. Mourad, A. Srour, H. Harmanani, C. Jenainati, M. Arafeh, 2020