

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Informatyki, Elektroniki i Telekomunikacji
Katedra Informatyki



Inżynieria oprogramowania 2019/2020

Środowisko do pozyskiwania spójnych tematycznych zbiorów danych
(wg. zbiorów użytkowników lub według zbiorów słów kluczowych) z
platformy Twitter, "Wybory prezydenckie w Polsce", wersja końcowa

Przygotowali:

Dominik Guz (dominikguz1@gmail.com)

Mateusz Nabywaniec

Dominika Mlynarczyk

Grzegorz Niedziela

Spis treści

1. Wstęp i opis problemu	3
2. Cel i rezultat projektu	5
2.1 Wymagania funkcjonalne	6
2.1.1 Wymagania dotyczące danych	6
2.1.2 Wymagania dotyczące analiz	6
2.2 Wymagania niefunkcjonalne	7
2.2.1 Wymagania produktowe	7
2.2.2 Wymagania organizacyjne	7
2.2.3 Wymagania zewnętrzne	7
2.3 Analiza ryzyka	8
3. Wizja rozwiązania	10
3.1 Technologie	10
3.2 Analizowane dane	11
3.3 Propozycje analiz danych	16
3.4 Schemat architektury	19
4. Wyniki analiz statystycznych	26
5. Wyniki analiz tekstu	45
6. Wyniki analiz społeczności	51
7. Przykłady użycia	95
7.1 Uruchomienie aplikacji	95
7.2 Przykłady korzystania z aplikacji	95
8. Podział prac	100
8.1 Opis procesu wytwarzania	100
9. Podsumowanie	101
9.1 Wnioski końcowe z projektu	101
9.2 Ewaluacja oceny ryzyka	102
9.3 Wnioski z projektu dotyczące procesu wytwarzania produktu	103
9.4 Ocena użyteczności i kierunek rozwoju projektu	104
9.5 Wnioski dotyczące używanych technologii	105
9.6 Co byśmy zmienili, gdybyśmy wykonali projekt od początku	105
10. Raporty ze spotkań	106
11. Źródła	115

1. Wstęp i opis problemu

Twitter jest serwisem społecznościowym pozwalającym na prowadzenie własnego profilu i publikowanie krótkich wiadomości (maks. 280 znaków). Dzięki temu możliwa jest szybka wymiana myśli między użytkownikami. Z tego powodu zyskał popularność m.in. wśród polityków i partii politycznych.

Wybory odgrywają kluczową rolę we wszystkich krajach demokratycznych, a media społecznościowe są ważnym aspektem tego procesu. Obecnie partie polityczne coraz częściej polegają na platformach społecznościowych, takich jak Twitter i Facebook, do komunikacji z elektoratem, reklamowania się i agitacji przedwyborczej. Wykorzystanie mediów społecznościowych w politycznych kampaniach dramatycznie wzrosło w ciągu ostatnich kilku lat. Przewiduje się, że stanie się ono jeszcze bardziej krytyczne dla przyszłych kampanii politycznych, ponieważ tworzy się doskonały kanał obustronnej komunikacji na linii polityk-społeczeństwo oraz sprzyjają tworzeniu więzi kandydatów z ich zwolennikami. Popularne serwisy mogą być także użyte w celach analitycznych. Dzięki zbieraniu danych i wykorzystywaniu metod statystycznych oraz uczenia maszynowego można dostrzec trendy, takie jak popularność pewnych hasł/osób, co może być skorelowane z popularnością w świecie realnym, czego przykładem mogą być właśnie wybory.

Wybory prezydenckie odbywają się w Polsce co 5 lat. Zgodnie z art. 128 Konstytucji RP wybory prezydenckie muszą odbyć się w dzień wolny od pracy między 75. a 100. dniem przed końcem kadencji ustępującego prezydenta. Obecnym prezydentem jest Andrzej Duda, którego kadencja kończy się 6 sierpnia 2020 roku. Marszałek Sejmu Elżbieta Witek ogłosiła termin wyborów 5 lutego na 10 maja 2020 roku. Ewentualna druga tura została zaplanowana na 24 maja 2020 roku (odbędzie się jeśli w I turze żaden z kandydatów nie uzyska połowy prawidłowych głosów). Z datą ogłoszenia wyborów rozpoczęła się kampania wyborcza.

Kandydaci na urząd prezydenta zgodnie z Konstytucją muszą spełniać różnego rodzaju warunki m.in. mieć co najmniej 35 lat. Najistotniejszym wymaganiem jest uzyskanie 100 tysięcy podpisów osób uprawnionych pod listami poparcia kandydata.

Postanowieniem Marszałek Sejmu 26 marca minął termin składania podpisów w Państwowej Komisji Wyborczej.

Obecnie PKW nie zweryfikowała wszystkich kandydatów. W naszym projekcie zajmiemy się analizą danych odnośnie kandydatów którzy są najpopularniejsi, mają największe szanse według sondaży. Są to:

- Andrzej Duda (PiS)
- Małgorzata Kidawa-Błońska (PO)

- Władysław Kosiniak-Kamysz (PSL)
- Robert Biedroń (Lewica)
- Szymon Hołownia (Bezpartyjny)
- Krzysztof Bosak (Konfederacja)

Chociaż w Polsce Twitter nie jest tak popularny jak np. Facebook (dane z 2018 roku mówią o ok. 4 mln realnych użytkowników^[6]) to warto pamiętać o tym, że w obecnie najważniejsze tweety publikowane na tym portalu są szeroko komentowane w programach telewizyjnych, portalach internetowych. Z tego powodu faktyczny zasięg tych informacji wykracza szeroko poza grupę użytkowników Twittera.

O tym, że Twitter odgrywa ważną rolę w dzisiejszej polityce świadczy praca naukowców z Uniwersytetu Duisburg-Essen "Trump versus Clinton – Twitter Communication During the US Primaries"^[4]. Naukowcy przeanalizowali tweety tworzone przez Donalda Trumpa i Hillary Clinton w czasie wyborów prezydenckich w USA w 2016 roku. Analiza ta pokazała zwiększoną aktywność kandydatów w czasie ważnych wydarzeń na świecie. Naukowcy zbadali również najczęściej używane słowa - na tym polu ujawniły się również różnice świadczące o tym że kandydaci kierowali tweety do różnych grup głosujących. Dane potwierdziły częściowo tezę, że Donald Trump porusza kwestie które wcześniej zainicjowała Hillary Clinton. Możemy snuć domysły, że ta taktyka pozwoliła zyskać Trumpowi dodatkowych zwolenników, ponieważ łatwiej jest odnosić się do tweetów kontrkandydata. Sam Donald Trump w wywiadzie dla Financial Times^[2] stwierdził że bez Twittera nie wygrałby wyborów.

Powyższe przykłady pokazują jak dużą rolę odgrywają media społecznościowe w wyborach. Dzięki analizie danych możemy stwierdzić jakie kwestie były najczęściej poruszane przez kandydatów a także jak były odbierane przez użytkowników np. przez ilość polubień. Popularność kandydata w internecie znajdowała też odzwierciedlenie w jego wynikach, była wprost proporcjonalna do ilości zdobytych głosów.

2. Cel i rezultat projektu

Naszym celem jest stworzenie aplikacji umożliwiającej zbieranie danych z twittera na podstawie słów kluczowych a także odpowiednich użytkowników związanych z wyborami prezydenckimi. Do zbierania tweetów będziemy używali API Twittera (konieczna jest rejestracja na stronie developer.twitter.com w celu uzyskania kluczów dostępu).

Do zbierania tweetów będziemy używali API Twittera (konieczna jest rejestracja na stronie developer.twitter.com w celu uzyskania kluczów dostępu).

API Twittera pozwala nam na dostęp do tweetów opublikowanych przez konkretnego użytkownika (poprzez metodę *api.user_timeline*) z ostatniego miesiąca, a dostęp tweetów zawierających np. konkretny hashtag (poprzez funkcję *api.search*) z ostatnich 7 dni. Zgodnie z polityką Twittera nie możemy przechowywać pełnych tweetów możemy jednak przechowywać id tweeta dlatego będziemy mieli do nich dostęp nawet po wielu dniach (przy użyciu funkcji *api.get_status*). W każdej z tabel znajdują się id tweeta a także dane które wykorzystamy przy analizie danych: *created_date* - data utworzenia tweeta, *favorite_count* - ilość polubień tweeta, *retweet_count* - ilość polubień.

Zbierane dane zostały opisane w punkcie 3.2. ID tweetów zbieramy do różnych tabel:

- **election_tweets** zawiera tweety zawierających ogólne hasztagi wyborcze np: #wybory2020
- dla każdego kandydata "x" mamy tabelę o nazwie **x_tweets** (np. **duda_tweets**) zawierającą tweety opublikowane przez niego lub jego sztab (nie filtryjemy tych tweetów, aby móc później wykonać analizę słów używanych przez danego kandydata)
- dla każdego kandydata "x" mamy tabelę o nazwie **x_hashtags** (np. **duda_hashtags**) zawierającą tweety zawierające hasztagi dotyczące danego kandydata np. #duda2020.
- **journalist_tweets** zawiera tweety z kont dziennikarzy opisanych w punkcie 3.2. Tweety te zawierają słowa kluczowe dotyczące wyborów.

Dzięki zebranym tweetom możemy dokonać wielu różnych analiz. Pierwszą z nich jest analiza popularności kandydatów. Można tego dokonać poprzez zliczanie ilości polubień i retweetów. Pozwoli to sprawdzić który kandydat jest najbardziej zaangażowany w kampanię. Można także badać tweety z hasztagami poparcia - liczba tych tweetów będzie świadczyła o poziomie poparcia dla danego kandydata. Kolejną z analiz którą chcemy przeprowadzić jest analiza tekstu tweetów. Chcemy znaleźć najczęściej poruszane tematy i najczęściej używane przez danego kandydata słowa. Pozwoli nam to na wyciągnięcie wniosków co jest ważne dla kandydatów i do jakich grup odbiorców kierowane są ich tweety. Ostatnią grupą analiz jest stworzenie grafów followersów oraz społeczności dla każdego kandydata.

Dokładniejsze analizy zostały opisane w punkcie 3.5.

2.1 Wymagania funkcjonalne

2.1.1 Wymagania dotyczące danych

- Aplikacja powinna umożliwiać analizę danych pochodzących z platformy Twitter
- Aplikacja powinna pobierać informacje o tweetach i użytkownikach z użyciem API Twittera
 - Informacje o tweetach:
 - id tweetu
 - autor tweetu
 - czas opublikowania tweetu
 - liczba retweetów
 - liczba polubień tweetu
 - hashtag/kandydat, z którym jest związany
 - Informacje o użytkownikach - kandydatach, członkach sztabów, kontach partii oraz dziennikarzach:
 - nazwa konta
 - czas utworzenia konta
 - liczba obserwujących konto
 - liczba przyjaciół
 - liczba tweetów
 - liczba polubień
- Aplikacja powinna przechowywać wymienione wyżej pobrane dane w bazie danych, aby skrócić czas dostępu do nich
- (Opcjonalnie) Aplikacja powinna pobierać dane cyklicznie samoistnie tj. w określonych odstępach czasu powinna sama aktualizować bazę danych tweetów na podstawie podanych hashtagów

2.1.2 Wymagania dotyczące analiz

- Aplikacja powinna wykonywać wiele analiz danych:
 - statystyczne
 - analizy aktywności kandydatów
 - analizy popularności kandydatów
 - eksplorację danych
 - analizy popularności tematów poruszanych przez kandydatów

- analizy stosunku użytkowników do kandydatów
 - badanie sieci społecznych (SNA)
 - analizy powiązań między wyborcami
- Aplikacja powinna udostępniać wyniki wykonanych analiz użytkownikowi w postaci wykresów słupkowych, map słów oraz grafów w przeglądarce
- Aplikacja powinna pozwalać użytkownikowi wybrać do analizy dane, które go interesują

2.2 Wymagania niefunkcjonalne

2.2.1 Wymagania produktowe

- Aplikacja powinna uruchamiać się możliwie jak najszybciej
- Aplikacja powinna działać zarówno pod urządzeniami z systemem Windows, jak i Linux/Unix
- Baza danych powinna być lekka i przenośna z uwagi jej lokalny charakter - baza znajduje się na maszynie
- Wyniki analiz powinny być zrozumiałe dla zwykłego użytkownika
- Interfejs graficzny aplikacji powinien być czytelny i intuicyjny

2.2.2 Wymagania organizacyjne

- Aplikacja powinna być zaimplementowana w języku Python
- Tworzenie i działanie aplikacji powinno być dokładnie udokumentowane - dokumentacja powinna zawierać opis użytych technologii, architektury oraz oferowanych funkcjonalności

2.2.3 Wymagania zewnętrzne

- Aplikacja nie może udostępniać kluczy do API Twittera ani innych danych uwierzytelniających dostęp do portalu
- Aplikacja nie może przekraczać limitów pobierania informacji z Twittera zdefiniowanych w zasadach użytkowania API
- Aplikacja powinna zapewniać bezpieczeństwo przechowywanych danych o użytkownikach Twittera

2.3 Analiza ryzyka

Ryzyko wynikające z czynników zewnętrznych:

lp	ryzyko	ocena ryzyka (1-10)	możliwe rozwiązanie
1.	możliwość zachorowania związana z obecną epidemiczną	3	praca zdalna, stosowanie się do zaleceń WHO i Ministerstwa Zdrowia
2.	trudności w komunikacji związane ze słabą jakością internetu	5	rozsądne dzielenie pracy między członków zespołu
3.	ograniczona ilość czasu związana z innymi zajęciami na uczelni, pracą itp.	5	rozsądne dzielenie pracy między członków zespołu

Ryzyko wynikające z czynników wewnętrznych

lp	ryzyko	ocena ryzyka (0 -10)	możliwe rozwiązanie
1.	ograniczenia związane z korzystaniem z API Twittera: jest ograniczona ilość tweetów które możemy pobierać - następnie musimy odczekać określony czas	8	filtrowanie tweetów które są interesujące, aktualizowanie bazy co parę dni
2.	zbyt mała ilość tweetów ponieważ mamy dostęp do tweetów z ostatniego miesiąca	6	regularne aktualizowanie danych, ewentualne analizy tweetów pod innym kątem
3.	mała możliwość analizy tweetów z powodu zawieszenia kampanii/odwołania wyborów	5	analiza tweetów pod innym kątem np. opinie kandydatów o działaniach związanych z epidemią, obecnie wybory nie zostały przeniesione także takie ryzyko jest mało prawdopodobne
4.	mała ilość danych do analizy np. najczęściej używanych słów gdyż głównym tematem jest epidemia koronawirusa	6	odpowiednie filtrowanie słów podczas analizy, analiza tweetów odnośnie sytuacji epidemicznej

5.	Dane i wyniki analiz mogą być przedstawione mało zrozumiałe	5	dopasowanie odpowiedniego widoku aplikacji
6.	Dane mogą być przechowywane w sposób niewygodny i utrudniający analizę, zbyt wolne zapytanie o tweet przez id	9	przemyślany sposób przechowywania danych, zmiana koncepcji - umieszczenie nie tylko id ale także innych informacji w bazie (daty, ilości lajków itd)

Największe zagrożenie w tym momencie stanowi dla nas epidemia, która powoduje, że pochłania ona bardzo dużą część uwagi użytkowników Twittera. W aktualnej sytuacji wybory schodzą nieco na drugi plan, a nawet na tym drugim planie króluje to, czy powinny się odbyć czy nie, a nie to, co chcieliśmy pierwotnie zaobserwować, czyli kampanie polityczne. Nie mamy na to żadnego wpływu, jednakże analiza pod tym kątem również będzie ciekawa. Punkty 3 i 4 są zatem na ten moment największym zagrożeniem dla naszego projektu. Analiza będzie jednak nieco utrudniona. Źle przeprowadzona analiza to nasze kolejne ryzyko, ale postaramy się wyciągnąć te dane, które będą znaczące i przedstawić je w przejrzysty i czytelny sposób, mając na uwadze ryzyko nr 5 - mało zrozumiała wizualizacja danych. Zagrożenia nr 1 i nr 6 nie stanowią dla nas dużego ryzyka - problem zbyt małej ilości danych można łatwo rozwiązać wykorzystując grupowy klucz do Api Twittera. Natomiast przechowywanie odpowiednich danych o tweecie umożliwi wykonanie zapytania tylko raz, co sprawi, że operacja ta będzie szybka.

3. Wizja rozwiązania

3.1 Technologie

Pierwszy etapem prac jest wybór pomocnych nam technologii oraz środowiska. Językiem programowania, który wykorzystamy do zrealizowania projektu będzie Python. Jest łatwy w zrozumieniu, dobrze udokumentowany i niezależny od systemu operacyjnego, co czyni go idealnym wyborem do stworzenia rozwiązania takiego typu. Jest też dostępnych wiele bibliotek do komunikacji z API

Twittera np. tweepy, co umożliwi nam wysokopoziomową komunikację z serwisem. Dane planujemy przechowywać w relacyjnej bazie danych SQLite3. Dostarcza ona wystarczającą ilość funkcjonalności, które będą nam potrzebne, a ponadto jest lekka, niezawodna i przenośna, więc dobrze nadaje się do naszych zastosowań.

Wyniki będą prezentowane w formie aplikacji webowej napisanej we frameworku do pythona - Flasku. Umożliwi nam on napisanie prostego interfejsu użytkownika oglądającego wyniki naszych analiz z poziomu przeglądarki internetowej. Część front-endową będą stanowić strony stworzone za pomocą Bootstrapa - frameworka CSS wypełniane przez back-end za pomocą silnika szablonów Jinja2, który umożliwia ewaluację zmiennych wewnętrz przygotowanego wcześniej szkieletu HTML. Jako, że wbudowany serwer WSGI Flaska nie jest rekomendowany, do uruchomienia aplikacji będziemy wykorzystywać serwer WSGI gunicorn.

Ważnym aspektem technologicznym jest użycie odpowiednich narzędzi do analizy danych. Pandas jest bardzo powszechnie wykorzystywany frameworkiem do pythona w zakresie analizy danych. Wyposażony jest w wiele modułów, dzięki czemu jest niesamowicie elastyczny i prosty w użyciu. Dodatkowym plusem jest posiadanie narzędzi do przetwarzania języka naturalnego, co pozwala na analizę sentymentu tweetów. Do analizy sieci społecznych wykorzystamy bibliotekę NetworkX. Jest ona popularna i dobrze udokumentowana. Umożliwia tworzenie i badanie struktury, dynamiki oraz funkcji sieci złożonych. Udostępnia wiele standardowych algorytmów grafowych oraz funkcji służących do badania sieci za pomocą licznych miar używanych przy analizie grafów.

W przypadku, gdybyśmy zdecydowali się na upublicznenie naszej aplikacji do internetu, dodatkowo uruchomimy *reverse proxy* w postaci Nginxa, który to przekazywałby ruch z dowolnego portu dostępnego z zewnątrz na port dostępny tylko wewnątrz maszyny, na którym nasłuchiwałaby nasza aplikacja. Jako maszyny, użylibyśmy serwera VPS, zakupionego na platformie Digital Ocean z systemem Debian 4.9 i adresem publicznym. Jednak nasze chęci do publicznego uruchomienia strony będą zależeć od przebiegu i wyników całego projektu.

3.2 Analizowane dane

Początkowym etapem prac jest ustalenie sposobu, w jaki będziemy wybierać dane do późniejszej analizy. Najistotniejsze dla naszego projektu będą tweety kandydatów oraz najważniejszych członków ich sztabów wyborczych, a także tweety zawierające hasztagi dotyczące wyborów - zarówno te ogólne, jak i oparte na hasłach wyborczych poszczególnych kandydatów. Ponadto weźmiemy pod uwagę związane z wyborami tweety dziennikarzy wypowiadających się na tematy polityczne.

W przypadku hashtagów mają priorytety w kolejności malejącej (tak jak zostały wymienione) - jeśli tweet będzie zawierał więcej niż jeden hashtag to zostanie zaklasyfikowany pod hashtagiem z najwyższym priorytetem. Wyбралиśmy to rozwiązanie, aby uniknąć redundancji tweetów w obrębie tabeli election. Uznaliśmy, że pozwoli to na analizę statystyczną tweetów, a jeśli będziemy chcieli analizować te tweety pod względem tekstu to będziemy musieli pobierać ich zawartość do pliku json. Kolejność hashtagów ustaliliśmy w ten sposób, że na początku sprawdzamy czy tweet zawiera dany hashtag - dlatego pierwsze hashtagi jak #GłosowanieKorespondencyjne zawierają pewną informację czego może dotyczyć dany tweet. Hashtagi które są niżej w tej liście dotyczą prawdopodobnie ogólnych wiadomości dotyczących wyborów.

Hashtagi ogólne:

- [#GłosowanieKorespondencyjne](#)
- [#WyboryKorespondencyjne](#)
- [#PrzełożyćWybory](#)
- [#IdziemyNaWybory](#)
- [#wybory2020](#)
- [#WyboryPrezydenckie2020](#)
- [#wybory](#)
- [#wyPAD2020](#)

W przypadku kandydatów z danej partii hashtagi mają priorytety w kolejności malejącej (tak jak zostały wymienione) - jeśli tweet będzie zawierał więcej niż jeden hashtag to zostanie zaklasyfikowany pod hashtagiem z najwyższym priorytetem. Wybraлиśmy to rozwiązanie, aby uniknąć redundancji tweetów w obrębie danej tabeli. Uznaliśmy, że pozwoli to na analizę statystyczną tweetów, a jeśli będziemy chcieli analizować te tweety pod względem tekstu to będziemy musieli pobierać ich zawartość do pliku json.

PIS

o Konta

- <https://twitter.com/AndrzejDuda>
- <https://twitter.com/AndrzejDuda2020>
- <https://twitter.com/AdamBielan> - rzecznik sztabu
- <https://twitter.com/mecenasJTK> - była szefowa sztabu (do 14 marca)

- <https://twitter.com/jbrudzinski> - obecny szef sztabu
- <https://twitter.com/pisorgpl>

- o Hashtagi

- [#Duda2020](#)
- [#AndrzejDuda2020](#)
- [#PAD2020](#)
- [#NiechŻyjePolska](#)

KO

- o Konta

- https://twitter.com/M_K_Blonska
- <https://twitter.com/adamSzlakpa> - rzecznik sztabu
- <https://twitter.com/Arlukowicz> - szef sztabu
- https://twitter.com/Platforma_org

- o Hashtagi

- [#Kidawa2020](#)
- [#KidawaTeam](#)
- [#MuremzaKidawą](#)
- [#PrawdziwaPrezydent](#)

Lewica

- o Konta

- <https://twitter.com/RobertBiedron>
- <https://twitter.com/poseiTTrala> - szef sztabu
- https://twitter.com/B_Maciejewska - rzeczniczka prasowa
- https://twitter.com/_Lewica

- o Hashtagi

- [#Biedron2020](#)
- [#Polska2020](#)
- [#StudioBiedron](#)

PSL

- o Konta

- <https://twitter.com/KosiniakKamysz>
- <https://twitter.com/magdasobkowiak> - szefowa sztabu
- <https://twitter.com/DariuszKlimczak> - rzecznik sztabu

- <https://twitter.com/nowePSL>

- o Hashtagi

- [#Kosiniak2020](#)
- [#NadziejaDlaPolski](#)

Konfederacja

- o Konta

- <https://twitter.com/krzysztofbosak>
- <https://twitter.com/Bosak2020>
- <https://twitter.com/PLusiadek> - szef sztabu
- <https://twitter.com/annabrylka> - rzecznik sztabu
- <https://twitter.com/Konfederacja>

- o Hashtagi

- [#Bosak2020](#)
- [#PrezydentBosak](#)
- [#NaprzódPolsko](#)

Bez partii

- o Konta

- https://twitter.com/szymon_holownia
- <https://twitter.com/michalkobosko> - pełnomocnik wyborczy

- o Hashtagi

- [#Hołownia2020](#)
- [#ekipaSzymona](#)
- [#BezpartyjnyBezpiecznik](#)

Dodaliśmy również nową tabelę accounts zawierającą dane o kontach kandydatów i ich sztabów. Zawiera nazwę konta, datę utworzenia, ilość statusów(tweetów i retweetów), ilość followersów, ilość polubień (wykonanych przez konto) oraz ilość przyjaciół. Te dane będą mogły posłużyć do sprawdzenia zasięgów konta danego użytkownika.

Tabela: accounts

	account	created_at	followers_count	statuses_count	favourites_count	friends_count
	Filtr	Filtr	Filtr	Filtr	Filtr	Filtr
1	AndrzejDuda	2010-10-13 07:58:45	1119990	7677	28	764
2	RobertBiedron	2012-01-17 19:49:58	224131	17405	28278	1896
3	pisorgpl	2008-08-16 23:26:19	205312	55920	795	691
4	krzysztofbosak	2009-10-07 20:47:37	173223	99231	50297	2495
5	Arlukowicz	2010-01-03 20:10:37	133427	19231	30809	744
6	Platforma_org	2009-07-02 07:46:15	130927	83891	18741	826
7	KosiniakKamysz	2012-11-18 10:30:44	107849	7903	10958	1260
8	jbrudzinski	2010-02-14 10:26:00	105490	20725	12524	720
9	M_K_Blonska	2014-01-07 09:28:49	103566	3754	13596	519
10	__Lewica	2009-08-17 10:17:40	59117	31252	10979	808
11	AdamBielan	2009-05-14 16:01:53	49971	7842	1463	1310
12	Konfederacja_	2019-02-07 14:01:26	39415	17717	12300	549
13	szymon_holownia	2019-11-11 17:43:07	35816	893	269	204
14	adamSzlapka	2012-02-28 22:11:31	28904	10436	29286	1332
15	nowePSL	2014-04-04 09:05:19	22493	29490	23550	305
16	AndrzejDuda2020	2018-02-22 13:31:55	15151	787	28	146

Rys.1 Tabela zawierająca interesujące nas konta.

W wyborach prezydenckich bierze udział jeszcze czterech kandydatów: Stanisław Żółtek (Kongres Nowej Prawicy), Marek Jakubiak (Federacja dla Rzeczypospolitej), Mirosław Piotrowski (Ruch Prawdziwa Europa) oraz Paweł Tanajno (bezpartyjny). Zdecydowaliśmy jednak, że do analizy wybierzemy tylko tych najbardziej znaczących, pojawiających się w sondażach przedwyborczych.

Dziennikarze

- Tomasz Lis - https://twitter.com/lis_tomasz
- Eliza Michalik - <https://twitter.com/EMichalik>
- Konrad Piasecki - <https://twitter.com/KonradPiasecki>
- Dominika Wielowieyska - <https://twitter.com/DWielowieyska>
- Stanisław Janecki - https://twitter.com/St_Janecki
- Bartosz Węglarczyk - <https://twitter.com/bweglarczyk>
- Katarzyna Kolenda-Zaleska - <https://twitter.com/KolendaK>
- Tomasz Sekierski - <https://twitter.com/sekielski>
- Michał Karnowski - <https://twitter.com/michalkarnowski>
- Mariusz Kolonko - <https://twitter.com/maxkolonko>
- Cezary Gmyz - <https://twitter.com/cezarygmyz>
- Rafał Ziemkiewicz - https://twitter.com/R_A_Ziemkiewicz

Dziennikarze zostali wybrani na podstawie rankingu opublikowanego na stronie [wirtualnemedia.pl](#)^[9], z dodatkowym zastrzeżeniem na ilość followersów

powyżej 100 tysięcy. Przedstawiony w źródłach ranking jest ogólny, dlatego wybrano z niego tylko dziennikarzy politycznych.

Analogicznie do tabeli z kontami kandydatów stworzyliśmy tabelę journalist_accounts. Posiada ona taką samą strukturę jak accounts.

Tabela: journalist_accounts											
	account	created_at	followers_count	statuses_count	favourites_count	friends_count	Filtr	Filtr	Filtr	Filtr	Filtr
1	lis_tomasz	2012-09-03 1...	903837	28825	28133	1135					
2	EMichalik	2011-07-12 0...	177217	8888	16561	439					
3	KonradPiasecki	2009-08-28 0...	485231	46179	7542	433					
4	DWielowieyska	2013-03-12 1...	136484	22021	7447	460					
5	St_Janecki	2012-08-31 1...	168479	56857	1791	549					
6	bweglarczyk	2010-04-18 0...	339307	117932	28530	3229					
7	sekierski	2010-09-15 1...	347944	6500	4944	464					
8	michalkarnow...	2009-05-19 1...	159992	16105	2229	2262					
9	maxkolonko	2009-04-25 0...	145710	1300	3284	0					
10	cezarygmzyz	2010-07-21 1...	163665	40950	857	1388					
11	R_A_Ziemkie...	2013-04-29 0...	192122	46349	1993	252					

Rys.2 Tabela zawierająca interesujące nas konta dziennikarzy.

API Twittera pozwala na zbieranie tweetów opublikowanych przez konkretnych użytkowników (poprzez metodę `api.user_timeline`) z ostatniego miesiąca, a zbieranie tweetów zawierających np. konkretny hashtag (poprzez funkcję `api.search`) z ostatnich 7 dni. Ważną kwestią jest konieczność filtrowania retweetów - są to tweety pochodzące od innego użytkownika które zostały uznane za warte uwagi przez udostępniającego. Rozpoznajemy jest przez zawieranie "RT @" w zawartości tekstu tweeta.

Planujemy zbierać id tweetów zawierających ogólne hasztagi wyborcze do osobnej tabeli, podobnie jak id tweetów z hashtagami odnoszącymi się do konkretnych kandydatów - jest to, aby uniknąć sytuacji w której tweet zawierający więcej niż jeden hashtag zostanie zaklasyfikowany do konkretnej grupy. W tabeli z tweetami kandydatów będą przechowywane id tweetów, które zostały opublikowane przez konta związane z danym kandydatem (konto kandydatów, jego partii, szefa sztabu, rzecznika), a także nazwa autora. W analogiczny sposób planujemy pobierać tweety dziennikarzy.

3.3 Propozycje analiz danych

Analizy statystyczne:

I p	Analiza	Sposób przedstawienia	Używane dane	Cel
1	Porównanie łącznej liczby tweetów danych kandydatów i ich sztabów	Wykres słupkowy	Tweety z tabeli candidates_tweets	Porównanie aktywności kandydatów
2	Porównanie łącznej liczby polubień które otrzymały tweety kandydatów i ich sztabów	Wykres słupkowy	Tweety z tabeli candidates_tweets	Porównanie popularności tweetów kandydatów
3	Porównanie łącznej liczby retweetów które otrzymały tweety kandydatów i ich sztabów	Wykres słupkowy	Tweety z tabeli candidates_tweets	Porównanie popularności tweetów kandydatów
4	Liczba tweetów danego kandydata i jego sztabu w zależności od czasu	Wykres liczby tweetów w zależności od czasu	Tweety z tabeli candidates_tweets	Dowiedzenie się z czego wynikała aktywność (np. wydarzenia krajowe)
5	Najpopularniejszy tweet kandydata	Tekst	Tweety z tabeli candidates_tweets	Dowiedzenie się czego dotyczył najpopularniejszy tweet

Analizy tekstu:

lp	Analiza	Sposób przedstawienia	Używane dane	Cel
1	Ranking hashtagów używanych przez kandydata	Mapa słów, wykres słupkowy	Treść tweetów z tabeli candidates_tweet s pobrana do pliku json	Określenie najpopularniejszych hashtagów dla poszczególnych kandydatów (co miało też wpływ na zasięgi tweetów)
2	Najpopularniejsze słowa używane przez kandydatów	Mapa słów,wykres słupkowy	Treść tweetów z tabeli candidates_tweet s pobrana do pliku json	Zobaczenie najpopularniejszych słów używanych przez kandydata, sprawdzenie czy ma to związek z wydarzeniami społecznymi, politycznymi
3	Użytkownicy, z którymi kandydaci wchodzili w interakcję poprzez mechanizm odpowiedzi (reply)	tabela	Treść tweetów z tabel *_hashtags pobrana do pliku json	Zobaczenie jak często kandydaci reagowali na innych użytkowników twittera oraz jak bardzo aktywny był każdy z nich względem innych osobistości przez mechanizm odpowiedzi
4	Miara podobieństwa między wypowiedziami kandydatów mierzone za pomocą odległości między wektorami słów, z których składały się ich tweety	tabela	Treść tweetów z tabel *_hashtags pobrana do pliku json	Wskazanie na to, jak podobni są kandydaci i ich wypowiedzi, pomimo teoretycznych różnic między programami wyborczymi

Analizy społeczności:

I p	Analiza	Sposób przedstawienia	Używane dane	Cel
1	Analiza społeczności, z którymi kandydaci wchodzą w interakcje	Graf	Tweety z tabel candidates_tweet s pobrane do pliku json	Prezentacja kont, z którymi kandydat wchodzi w interakcje w celu pokazania powiązań
2	Analiza społeczności kandydatów	Graf	Tweety z tabel *_hashtags pobrane do pliku json	Prezentacja kont które tweetują o kandydatach, aby stwierdzić czy użytkownicy tworzą osobne obozy poparcia czy obserwują wielu kandydatów

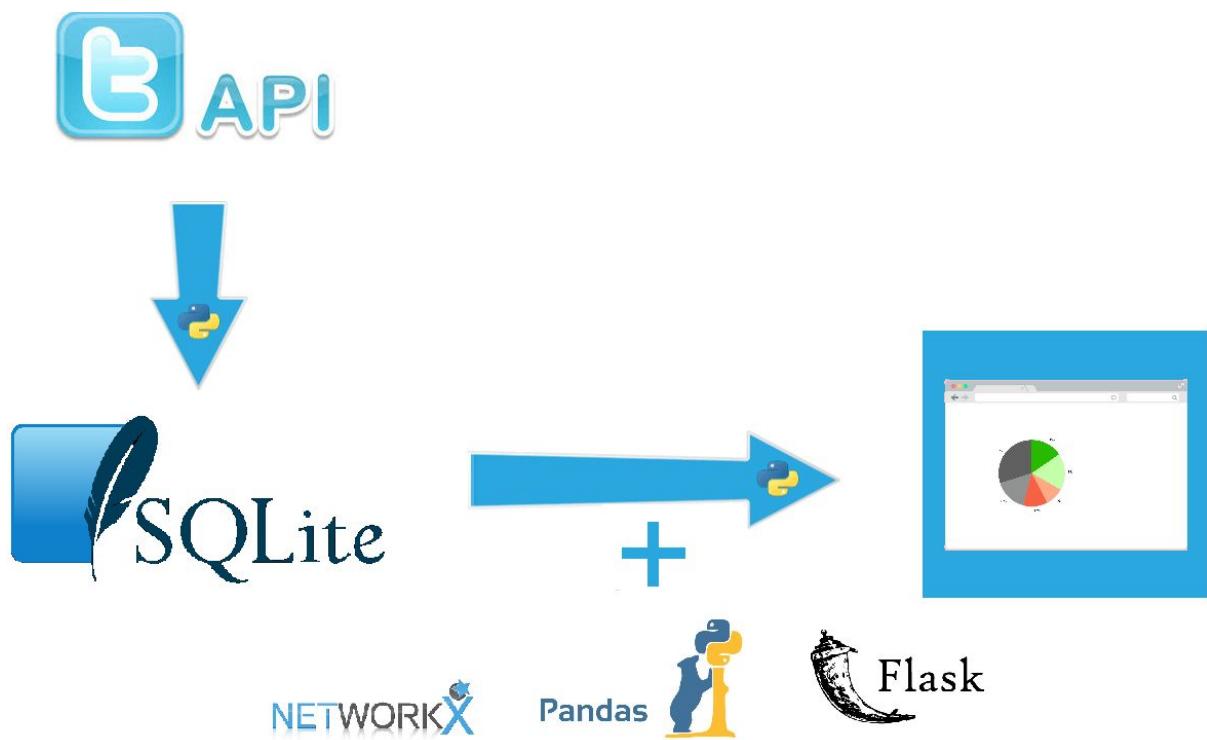
Po pobraniu danych należy ustalić sposób ich analizy. Na początek zostanie przeprowadzona prosta analiza popularności kandydatów tj. średnia i maksymalna ilość reakcji (polubień) oraz retweetów postów od każdego kandydata, średnia ilość postów związanych z kandydatem, wybranie najpopularniejszego postu danego kandydata. Ponadto wykres ilości tweetów publikowanych każdego dnia pozwoli zauważać dni, w których doszło do ważnych wydarzeń (w te dni i w kolejnych kilka więcej tweetów będzie publikowanych).

Innym ciekawym pomysłem jest wykorzystanie analizy sentymentu. Polega ona na analizie tekstu pod kątem nacechowania emocjonalnego. Dzięki wykorzystaniu tej formy badania można wywnioskować, jak społeczeństwo widzi kandydata (kandydat może się wydawać popularny przez dużą ilość retweetów lub postów z jego hashtagiem, jednak większość zawartości może być nacechowana negatywnie). Wymagać to będzie oczywiście ustawienia filtra na słowa takie jak np zaimki, które mogą zaciemniać obraz analizy. Analiza tekstowa tweetów kandydatów

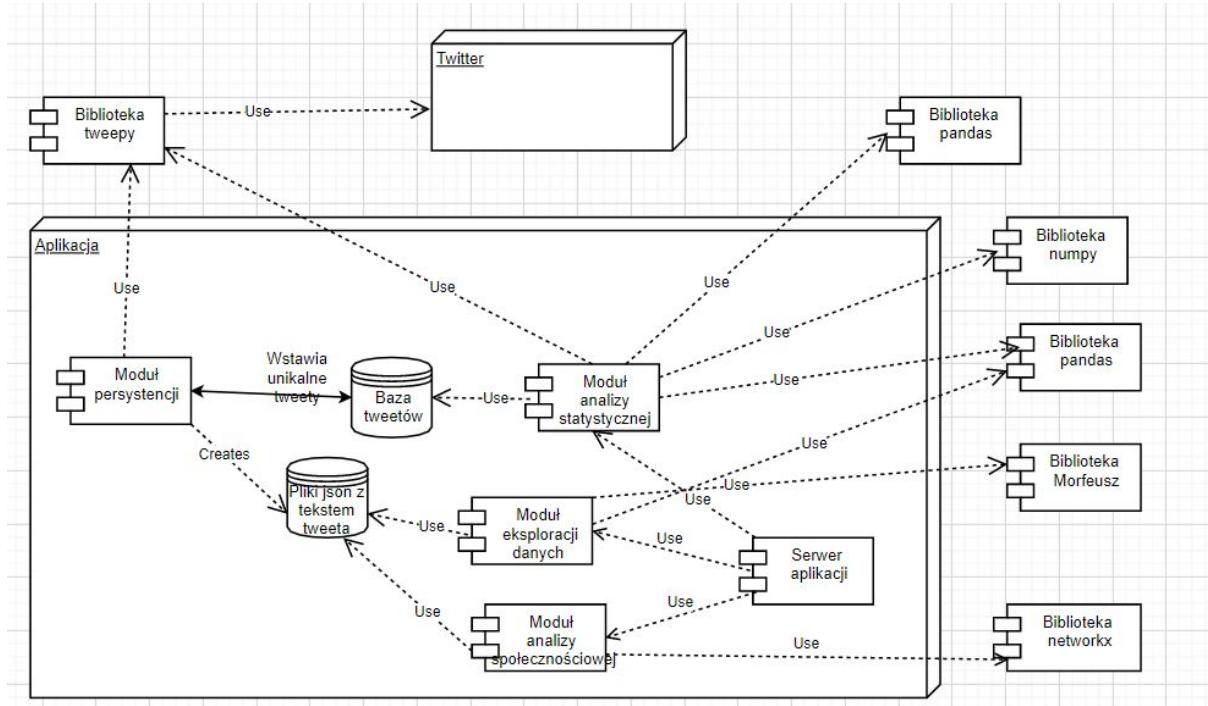
pozwoli też przybliżyć główne problemy czy zagadnienia poruszane przez poszczególnych polityków.

Trzecim rodzajem planowanych analiz są SNA czyli analizy sieci społecznych. Na twitterze są 3 możliwe formy interakcji między użytkownikami: mention (wspomnienie użytkownika w tweecie, jego nazwa znajduje się w tweecie), retweet (publikacja czegoś tweeta u siebie), reply (odpowiedź). SNA pozwala na tworzenie grafów między użytkownikami: kto kogo retweetuje, komu odpowiada, kto jest wspominany.

3.4 Schemat architektury



Rys. 3 Idea aplikacji.



Rys. 4 Architektura aplikacji.

Moduł pobierania tweetów - jest to moduł który pozwala na pobieranie i umieszczenie tweetów w bazie. Składa się z następujących metod:

- **get_tweets_by_users** - zdobywa tweety stworzone przez konta powiązane z danym kandydatem i wstawia unikalne tweety (sprawdzając czy tego tweetu nie ma już w bazie) do tabeli **candidates_tweets**. Do pobierania tweetów używamy metody `api.user_timeline` z biblioteki tweepy która pozwala na dostęp do tweetów z ostatniego miesiąca.
- **get_tweets_by_journalist_accounts** - zdobywa tweety których autorami są dziennikarze. Umieszcza unikalne tweety w tabeli **journalist_tweets**. Do pobierania tweetów używamy również metody `api.user_timeline` z biblioteki tweepy.
- **get_tweets_by_hashtag** - zdobywa tweety zawierające podany hashtag i wstawia unikalne tweety do tabeli podanej jako argument. Metoda ta jest używana do wstawiania tweetów do tabeli **election_tweets** a także do wstawiania tweetów do tabel z tweetami zawierającymi hasztagi odnoszące się do danego kandydata np. **duda_hashtags**. Do pobierania tweetów zawierających dany hashtag używamy metody `api.search` z biblioteki tweepy która pozwala na dostęp do tweetów z ostatniego tygodnia.

Baza danych:

Struktura bazy:

Baza składa się z 9 tabel:

Nazwa
▼ Tabele (9)
> biedron_hashtags
> bosak_hashtags
> candidates_tweets
> duda_hashtags
> election_tweets
> holownia_hashtags
> journalist_tweets
> kidawa_hashtags
> kosiniak_hashtags

Rys. 5 Tabele w bazie danych.

Może zastanawiać to że mamy wspólną tabelę candidates_tweets dla wszystkich kandydatów i kont ich sztabów a osobne tabele związane z hashtagami np. duda_hashtags, biedron_hashtags itp. Było to spowodowane tym że o ile wiadomo kto jest autorem tweetów w tabeli candidates_tweets to tweety zawierające hashtagi odnoszące się do danego kandydata mogą odnosić się do więcej niż jednego ubiegającego się o urząd. W ten sposób, aby uniknąć utraty informacji związanej z zaklasyfikowaniem tweeta do jednego kandydata zdecydowaliśmy się utworzyć osobną tabelę *_hashtags. Wadę tego rozwiązań jest mniejsza skalowalność - dla każdego nowego kandydata musielibyśmy stworzyć nową tabelę. Uznaliśmy jednak, że w przypadku projektu mamy jasno zidentyfikowanych kandydatów a pobieranie danych jest ograniczone przez API więc mało prawdopodobna jest dodanie nowego kandydata.

Dla każdego z kandydatów mamy tabelę z hashtagami dotyczącymi tego kandydata np. duda_hashtags zawiera tweety zawierające hashtagi dotyczące Andrzeja Dudy. Autorami tych tweetów mogą być różni użytkownicy którzy często wyrażają swoje poparcie dla kandydata. W tabeli przechowujemy id, hashtag, datę utworzenia, nazwę autora oraz liczby mówiące o popularności tweeta, reakcji na tweet - liczbę polubień i liczbę retweetów. Hashtag jest to ten hashtag który jako pierwszy w został "złapany" przy analizowaniu tekstu. Oznacza to, że jeśli tweet zawiera hashtagi odnoszące się do Małgorzaty Kidawy-Błońskiej '#Kidawa2020' i '#PrawdziwaPrezydent' to ten tweet trafi do tabeli kidawa_hashtags i zostanie w polu hashtag znajdzie się '#Kidawa2020'.

Przykład dla tabeli duda_hashtags:

Tabela: duda_hashtags						
	tweet_id	hashtag	created_at	favorite_count	retweet_count	author_name
1	1246219326069321728	#Duda2020	2020-04-03 23:33:35	693	164	Bart_Wielinski
2	1245682172930994176	#Duda2020	2020-04-02 11:59:08	473	110	Tulajew
3	1245254951175143424	#Duda2020	2020-04-01 07:41:30	397	58	RobertTelus
4	1245773665582678019	#Duda2020	2020-04-02 18:02:41	392	116	Waszczykows...
5	1247200365235273729	#Duda2020	2020-04-06 16:31:53	378	69	AndrzejDuda...
6	1245089036584325126	#Duda2020	2020-03-31 20:42:13	257	44	MosinskiJan
7	1246025601418485760	#Duda2020	2020-04-03 10:43:48	230	7	KrisZdaniuk
8	1245023959554445312	#Duda2020	2020-03-31 16:23:38	220	45	Ivanka833
9	1247201320290914307	#Duda2020	2020-04-06 16:35:41	199	58	AndrzejDuda...

Rys. 6 Dane zawarte w tabeli duda_hashtags.

W bazie mamy także tabelę candidates_tweets.

Zawiera ona tweety których autorami są konta związane z danym kandydatem - on sam, jego rzecznik, sztab itd (dokładniej zostali opisani w punkcie 3.2). W tej tabeli przechowujemy id tweeta, datę utworzenia, nazwę autora, ilość polubień i retweetów oraz nazwisko kandydata do którego sztabu należy autor.

Przykład:

Tabela: candidates_tweets						
	tweet_id	candidate_name	author_name	favorite_count	retweet_count	created_at
1	1241026460988715...	Kidawa	Arlukowicz	8332	466	2020-03-20 15:39:00
2	1239227299226058...	Duda	AndrzejDuda	7648	1271	2020-03-15 16:29:46
3	1249271650471890...	Duda	AndrzejDuda	7314	742	2020-04-12 09:42:26
4	1236050727128231...	Biedron	RobertBiedron	6939	1146	2020-03-06 22:07:12
5	1245798659771564...	Duda	AndrzejDuda	6665	798	2020-04-02 19:42:00
6	1236050810116681...	Kidawa	M_K_Blonska	6360	1266	2020-03-06 22:07:32
7	1240607231940460...	Duda	AndrzejDuda	5943	800	2020-03-19 11:53:08
8	1244598402077413...	Duda	AndrzejDuda	5581	921	2020-03-30 12:12:37
9	1248505235435212...	Duda	AndrzejDuda	5401	920	2020-04-10 06:56:58
10	1249774476302209...	Biedron	RobertBiedron	5316	1381	2020-04-13 19:00:29

Rys. 7 Dane zawarte w tabeli candidates_tweets.

Kolejną tabelą w bazie jest tabela election_tweets.

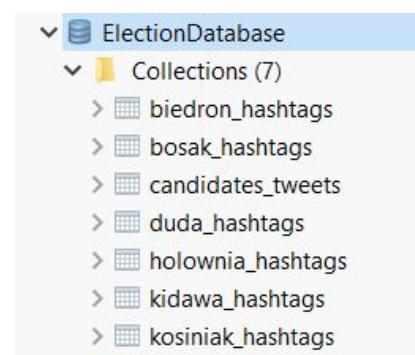
Zawiera ona tweety które zawierają hasztagi związane ogólnie z wyborami (opisane w punkcie 3.2). Tweety zawierają zarówno ogólne hasztagi jak #WyboryPrezydenckie jak i bardziej specyficzne które zyskały popularność w ostatnich dniach np. #GłosowanieKorespondencyjne. W tej tabeli przechowujemy id tweeta, hasztag, datę utworzenia, nazwę autora, ilość polubień i retweetów.

Przykładowe rekordy:

Tabela: election_tweets						
	tweet_id	hashtag	created_at	favorite_count	retweet_count	author_name
1	12446148532...	#PrzełożyćWybory	2020-03-30 13:17:59	2475	251	szymon_holownia
2	12475945515...	#WyboryKorespondencyjne	2020-04-07 18:38:14	1171	260	ZGryglas
3	12465256526...	#PrzełożyćWybory	2020-04-04 19:50:49	1081	174	szymon_holownia
4	12472614008...	#wybory	2020-04-06 20:34:25	1055	168	SutrykJacek
5	12456300967...	#Wybory2020	2020-04-02 08:32:12	975	201	rafalhubert
6	12511868211...	#WyboryKorespondencyjne	2020-04-17 16:32:38	963	117	RobertBiedron
7	12512216459...	#Wybory2020	2020-04-17 18:51:01	912	249	przepisnazwierz
8	12471349562...	#WyboryKorespondencyjne	2020-04-06 12:11:58	898	164	MichałSzczerba
9	12474108966...	#WyboryPrezydenckie2020	2020-04-07 06:28:28	887	307	tvpikorea
10	12468185064...	#Wybory2020	2020-04-05 15:14:31	881	506	OloCzarny

Rys. 8 Dane zawarte w tabeli election_tweets.

Struktura bazy danych stwarza problemy przy analizie tweetów pod względem tekstu. Aby móc to wykonać musimy pobierać tweety korzystając z API twittera do jsona na podstawie id wyciągniętego z bazy. Dodatkowym problemem może być pewna redundancja danych - w naszej bazie nie mamy powiązań przez co do różnych tabel mogą trafić te same tweety w zależności od autora oraz zawartości. Rozwiązanie problemu może być zmiana podejścia. Możemy przechowywać dane w bazie dokumentowej. Taką bazą jest MongoDB.



Rys. 9 Przykładowe kolekcje tweetów.

Każda z tabel jest kolekcją dokumentów w formacie JSON. Każdy dokument dotyczy jednego tweeta. Zawiera pełne informacje o tweecie pobrane z API.

Key	Value	Type
✓ (1) ObjectId("5ec2c6c787c4925728dd6cdd")	{ 25 fields }	Object
└ _id	ObjectId("5ec2c6c787c4925728dd6cdd")	ObjectId
└ created_at	Tue Mar 03 14:26:11 +0000 2020	String
└ id	1234847543747653633	Int64
└ id_str	1234847543747653633	String
└ full_text	Premier @MorawieckiM w #Katowice: w Polsce nie ma potwierzonego przypa...	String
└ truncated	false	Boolean
└ display_text_range	[2 elements]	Array
└ entities	{ 4 fields }	Object
└ source	Twitter Web App	String
└ in_reply_to_status_id	null	Null
└ in_reply_to_status_id_str	null	Null
└ in_reply_to_user_id	null	Null
└ in_reply_to_user_id_str	null	Null
└ in_reply_to_screen_name	null	Null
└ user	{ 42 fields }	Object
└ geo	null	Null
└ coordinates	null	Null
└ place	null	Null
└ contributors	null	Null
└ is_quote_status	false	Boolean
└ retweet_count	14	Int32
└ favorite_count	60	Int32
└ favorited	false	Boolean
└ retweeted	false	Boolean
└ lang	pl	String

Rys.10 Przykładowa reprezentacja tweeta w MongoDB.

Dzięki temu podejściu, nie musielibyśmy pobierać tweetów do bazy SQLite a później tekst do JSONa. Dalej dochodziłoby jednak do redundancji danych z czym można się pogodzić ponieważ cechą charakterystyczną baz dokumentowych jest redundancja. Zaletą z kolei byłaby szybkość zapytań - jest to główny plus baz noSQL.

Ustaliliśmy jednak że w projekcie będziemy używali bazy SQLite. Podjęliśmy taką decyzję, ponieważ baza ta pozwala na łatwe pisanie zapytań SQL co jest ważne przy analizach statystycznych. Dodatkowo przy tych analizach nie potrzebujemy wszystkich danych - konieczne są dane o polubieniach, retweetach a także o dacie stworzenia. W przypadku analiz tekstu oraz społeczności używamy pełnych tweetów ściagniętych do plików json. Dzięki temu mamy możliwość analizy treści tweeta a także powiązań między użytkownikami.

Moduł analizy tweetów.

Moduł ten składa się z kilku skryptów które pobierają interesujące nas dane z bazy. Przy obliczaniu np. ilości polubień i retweetów dla tweetów zawierających dany hashtag związany z kandydatem stosujemy zapytanie zliczające ilość polubień (favorite_count) i retweetów (retweet_count) z danej tabeli związanej z danym kandydatem. Otrzymane dane wyświetlamy na wykresie słupkowym wykorzystując biblioteki pandas i matplotlib.

W podobny sposób przeprowadzamy inne analizy statystyczne.

W przypadku analiz tekstu konieczne okazało się pobieranie treści tweetów. Dokonujemy tego przy użyciu skryptu, który przyjmuje id tweetów z bazy i pobiera tweety przy pomocy zapisuje dane w pliku json przy pomocy metody api.get_status z funkcji.

Dane z bazy zostaną wykorzystane do analizy statystycznej z użyciem pandas. Analogicznie zawartość tweetów będzie pobrana w pliku json, które potem będą analizowane z użyciem takich bibliotek jak pandas i networkx do wykonania grafów kandydatów oraz po odpowiedniej tokenizacji i filtrowaniu słów.

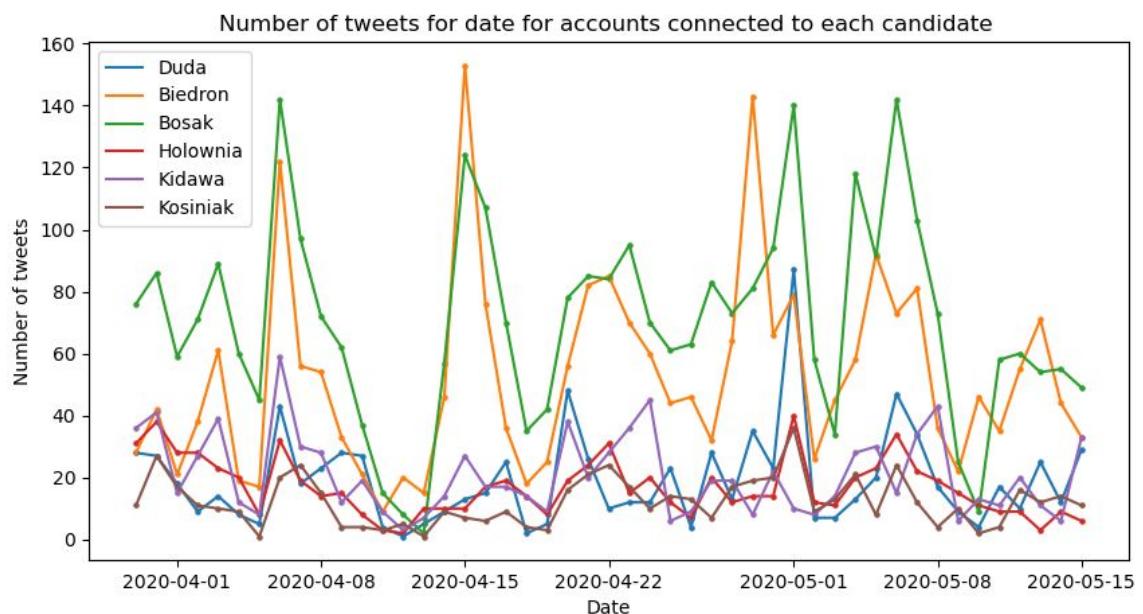
4. Wyniki analiz statystycznych

W tym rozdziale prezentujemy wyniki analiz statystycznych, które to skupiały się na pokazaniu oraz porównaniu kandydatów pod względem ich aktywności. Wyniki analiz zostały zaprezentowane w formie wykresów słupkowych oraz liniowych.

- Średnia liczba publikowanych tweetów, polubień oraz retweetów dla kandydatów - mając w bazie daty możemy zliczyć tweety opublikowane

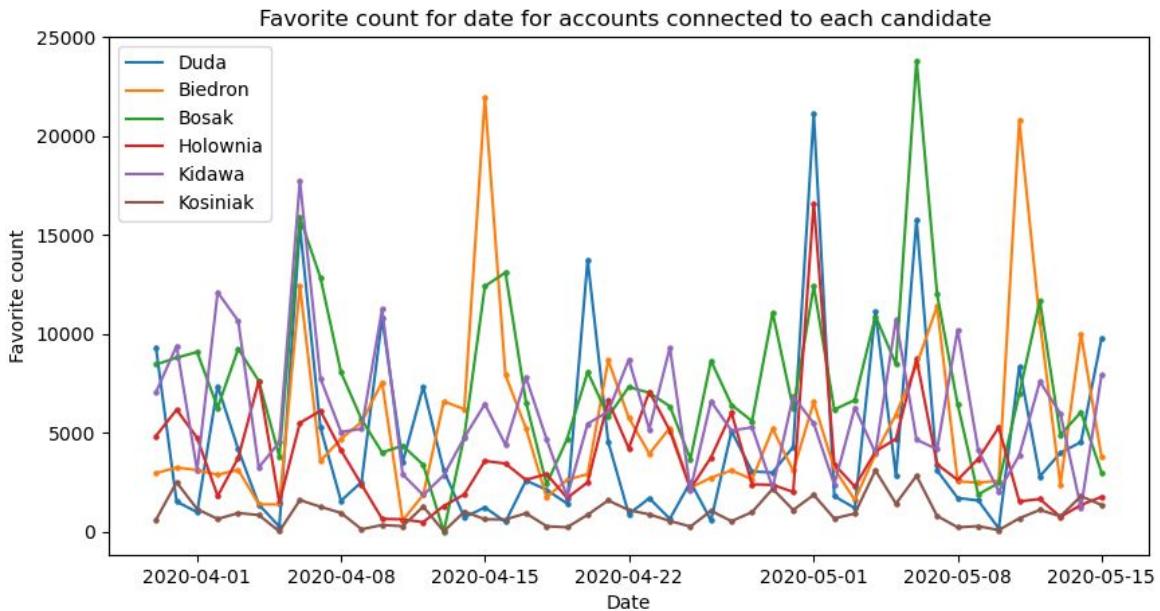
w podanym okresie czasu przez danego kandydata. Zliczamy także ilość lajków oraz retweetów, które dostały tweety w danym dniu.

Otrzymujemy w ten sposób wykresy mówiące wiele o aktywności kandydatów a także o odbiorze tweetów przez użytkowników



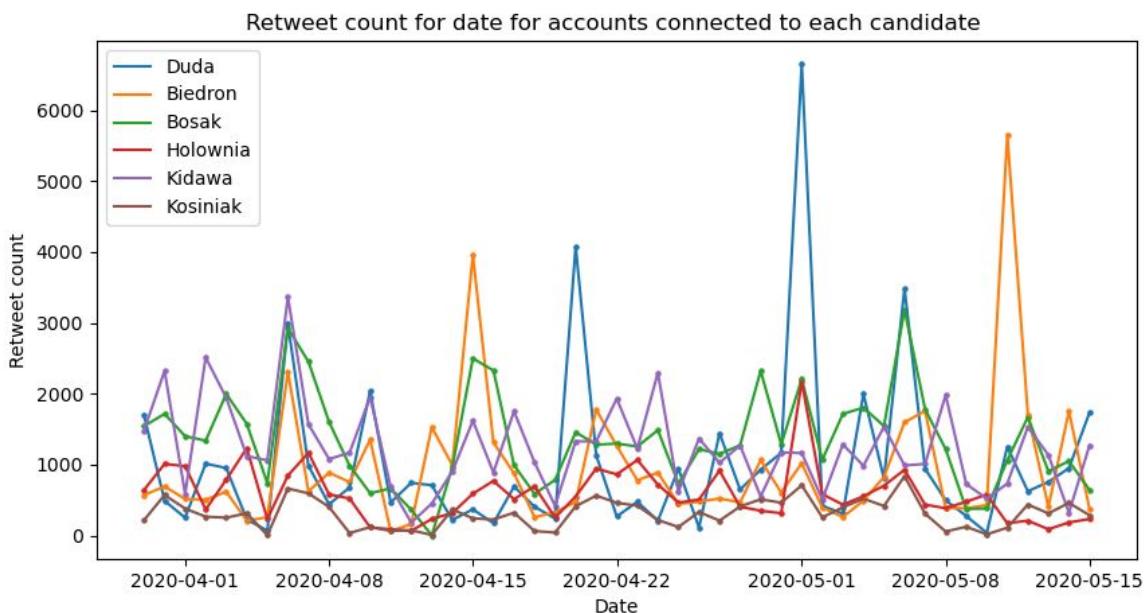
Rys.11 Wykres liczby tweetów tworzonych przez konta związane z kandydatem w zależności od czasu.

Analizując rysunek 11 łatwo, że najaktywniejszymi na Twitterze są Robert Biedroń i Krzysztof Bosak. Pozostali kandydaci są mniej aktywni. Widać również, że w pewnych dniach aktywność kandydatów wzrosła. Jest to np. 6 kwietnia kiedy przygotowywana była ustanowiona o wyborach korespondencyjnych. Kolejnym takim dniem jest 15 kwietnia - był to dzień po przerwie Wielkanocnej kiedy wznowiono obrady nad sposobem przeprowadzenia wyborów. Dodatkowo miał wtedy wchodzić nakaz zasłaniania twarzy w miejscach publicznych. Następnymi dniami ze wzmożoną aktywnością był 1 maja - Święto Pracy, początek długiego weekendu a także 6 maja wtedy podjęto ostatecznie decyzję, że wybory się nie odbędą. Było to porozumienie między Jarosławem Kaczyńskim a Jarosławem Gowinem.



Rys.12 Wykres liczby polubień tweetów tworzonych przez konta związane z kandydatem w zależności od czasu.

Analizując rysunek 12 widać że nie ma wyraźnej dominacji pewnych kandydatów tak jak w przypadku wykresu ilości tweetów. Może być to spowodowane tym, że np. Andrzej Duda ma większą ilość obserwujących przez co jego tweety trafiają do większej liczby użytkowników. Można zauważyć piki ilości polubień w dniach w których było tweetów. Było to prawdopodobne ponieważ większa ilość tweetów daje więcej możliwości do reagowania na nie.



Rys.13 Wykres liczby retweetów tweetów tworzonych przez konta związane z kandydatem w zależności od czasu.

Na wykresie z rysunku 13 również zauważalne piki w tych samych dniach w których była największa ilość tweetów i lajków. Także na tym wykresie nie widać wyraźnej dominacji pewnego kandydata nad pozostałymi.

Stworzyliśmy także tabelę pokazującą średnie ilości tweetów, polubień i retweetów dla danego kandydata.

Candidate	Avg tweets/day	Avg favorites/day	Avg retweets/day	Avg favorites/tweet	Avg retweets/tweet
Duda	19	4520	1019	237	53
Biedron	52	5273	955	101	18
Bosak	70	7469	1349	106	19
Holownia	16	3687	578	230	36
Kidawa	20	5904	1231	295	61
Kosiniak	12	952	309	79	25

Rys. 14 Tabela zawierająca średnie liczby dotyczące tweetów kandydatów.

Analizując wykresy z rysunków 11,12,13 oraz tabelę z rysunku 14 możemy wysnuć pewne wnioski dotyczące aktywności kandydatów. Widać, że najbardziej aktywne pod względem średniej liczby tweetów w ciągu dnia były sztab Krzysztofa Bosaka (70 tweetów) i Roberta Biedronia (50) tweetów. Pozostali kandydaci utrzymywali aktywność w zakresie 20-12 tweetów dziennie. Jeśli chodzi o średnią liczbę polubień to liczba tweetów sztabu Bosaka przeniosła się na największą liczbę (ponad 7000).

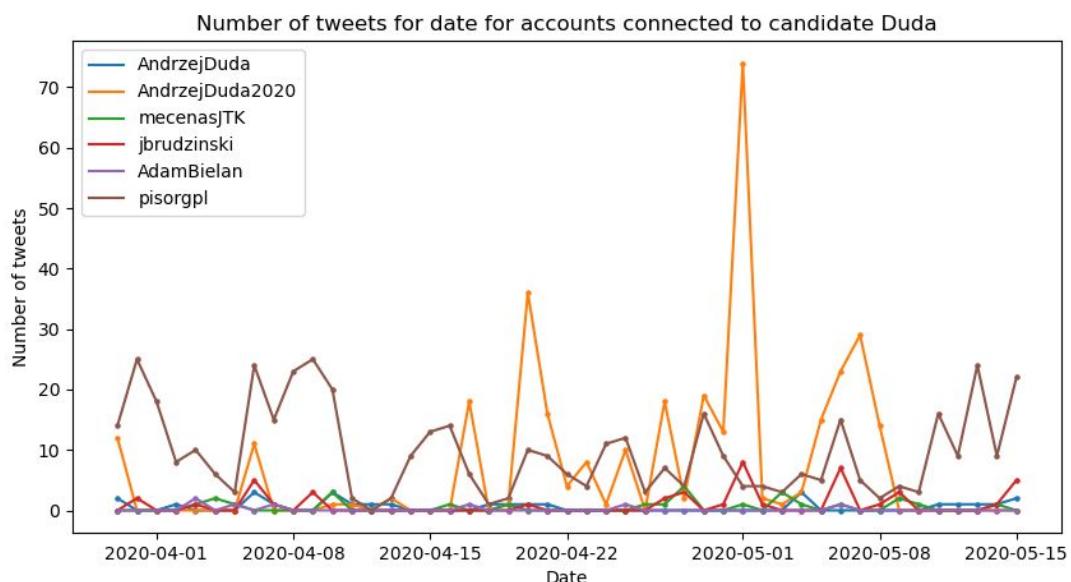
Dużo polubień otrzymywały tweety z obozu Małgorzaty Kidawy-Błońskiejj (prawie 6000) oraz Roberta Biedronia (ponad 5000).

Średnia ilość polubień na tweet pokazuje że średnio najwięcej polubień na tweet uzyskali Andrzej Duda, Małgorzata Kidawa-Błońska i Szymon Hołownia (około 200). Kandydaci którzy tworzyli najwięcej tweetów osiągnęli mniejszą średnią liczbę polubień (około 100). Mogło to być spowodowane tym że użytkownicy nie nadążali z reagowaniem na wszystkie tweety tworzone przez sztab tych kandydatów.

Jeśli chodzi o retweety, to najczęściej otrzymywały tweety pochodzące z kont Krzysztofa Bosaka, Małgorzaty Kidawy-Błońskiejj i Andrzeja Dudy. Retweety pokazują często pokazują że użytkownicy zgadzają się z danym kandydatem, uważając że treść jest interesująca. W porównaniu z ilością tweetów najlepiej wychodzą kandydaci popierani przez KO i PiS (ich tweety otrzymywały średnio 50-60 retweetów). Tweety Krzysztofa Bosaka i Roberta Biedronia podobnie jak przy polubieniach dostawały niewiele retweetów na 1 tweet.

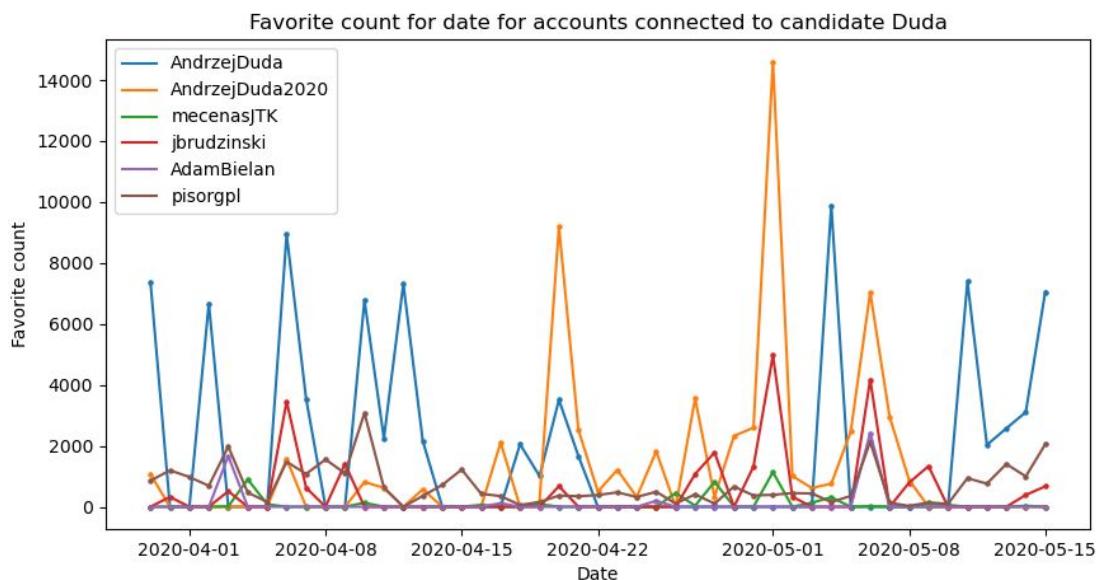
Ciekawą obserwacją jest niska aktywność kont związanych z Władysławem Kosiniakiem-Kamyszem. Generowały one najmniej tweetów, które otrzymały najmniej retweetów i polubień.

Ponadto stworzyliśmy wykresy liczby tweetów, polubień oraz retweetów dla kont związanych z poszczególnym kandydatem. Wykresy te pozwolą na analizę tego, które konto w sztabie danego kandydata było najaktywniejsze.



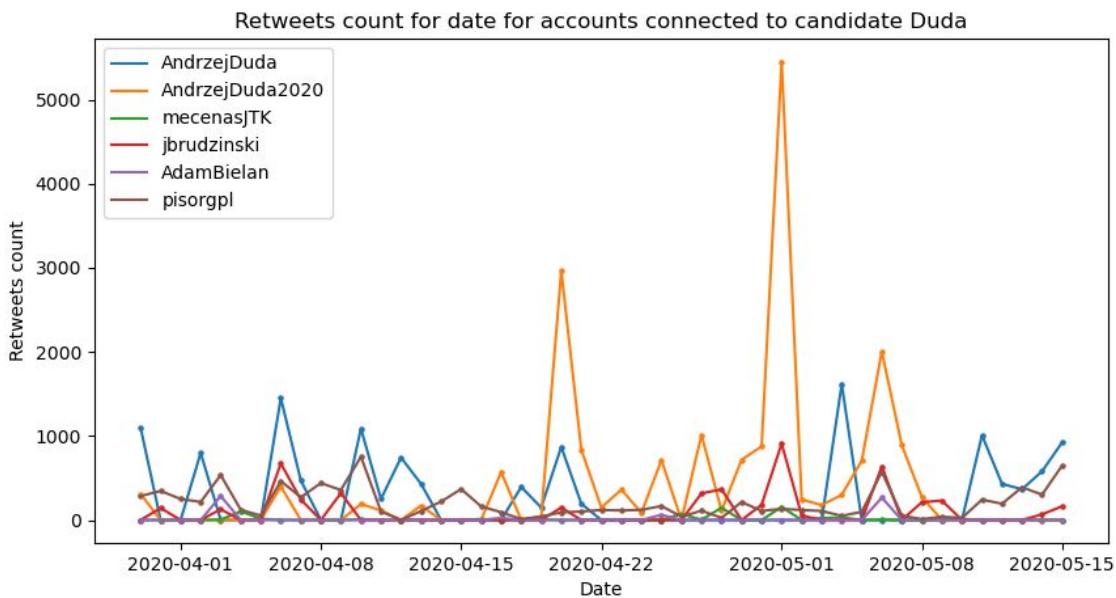
Rys.15 Wykres liczby tweetów tworzonych przez konta związanych z Andrzejem Dudą w zależności od czasu.

Analizując wykres z rys. 15, można zauważyc, że konta, które tworzyły najwięcej tweetów to konta "AndrzejDuda2020" - oficjalne konto kampanii, oraz "pisorgpl" - konto partii PiS popierającej Andrzeja Dudę. Widać również, jak w przypadku wykresu na rys. 11 wzrosty aktywności kont związanych z ubiegającym się o reelekcję prezydentem w pewnych dniach co jest związane z wydarzeniami politycznymi. Zastanawiająca jest niska aktywność oficjalnego konta Andrzeja Dudy - "AndrzejDuda". Liczba tweetów tworzonych przez to konto stanowi ułamek wszystkich tweetów tworzonych przez obóz kandydata. Pozostałe konta notowały raczej średnią aktywność.



Rys.16 Wykres polubień tweetów tworzonych przez konta związanych z Andrzejem Dudą w zależności od czasu.

Analizując wykres z rys. 16, można stwierdzić, że tweety, które otrzymywały najwięcej polubień, pochodziły najczęściej z konta "AndrzejDuda". Można było się spodziewać, ponieważ zasięgi tego konta są jednymi z największych w Polsce, przez co było prawdopodobne, że tweety będą cieszyły się sporą popularnością. W dniach w których najwięcej tweetów pochodziło z konta "AndrzejDuda2020" widać, że wtedy liczba polubień była również wysoka. Jest to zrozumiałe, ponieważ duża liczba tweetów daje możliwości reakcji na nie. Mało polubień otrzymywały tweety z konta "pisorgpl" szczególnie, gdy porównamy to z liczbą tworzonych tweetów na podstawie wykresu z rys. 15.



Rys.17 Wykres retweetów tweetów tworzących przez konta związanych z Andrzejem Dudą w zależności od czasu.

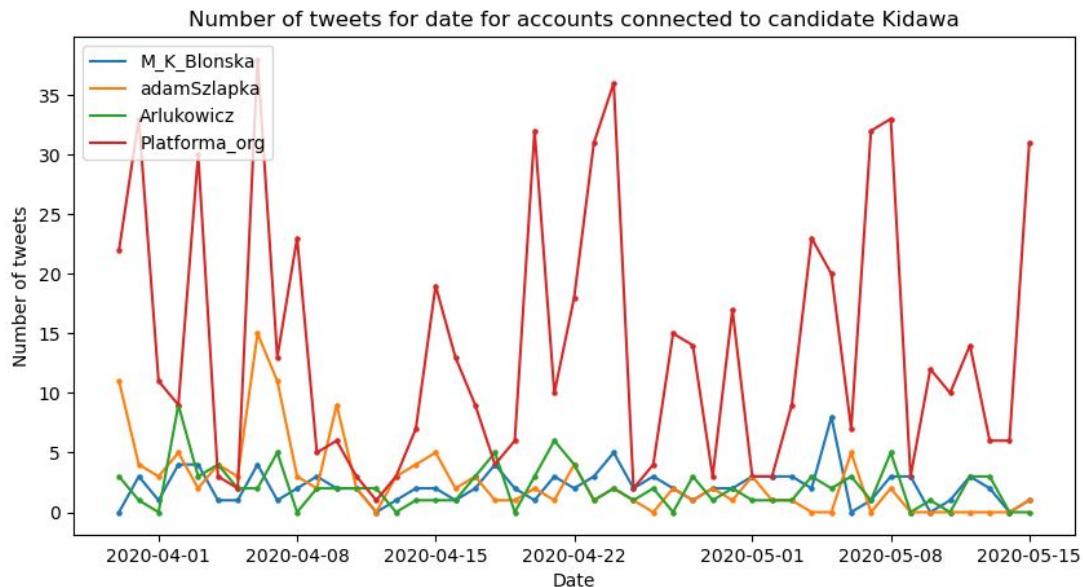
Analizując wykres z rys. 16, można stwierdzić, że tweety pochodzące z kont “AndrzejDuda2020” i “AndrzejDuda” otrzymywały najczęściej retweetów. Widać również, że w dniach gdy powstawało najwięcej tweetów, było najwięcej retweetów. Wykres potwierdza to co można wysnuć na podstawie wykresu z rys. 15, że konto kandydata, było tym, na którego tweety użytkownicy reagowali najchętniej.

Account	Avg tweets/day	Avg favorites/day	Avg retweets/day	Avg favorites/tweet	Avg retweets/tweet
AndrzejDuda	0.6	1814	273	3023	455
AndrzejDuda2020	7.1	1310	418	184	58
mecenasJTK	0.5	93	12	186	24
jbrudzinski	1.0	507	101	507	101
AdamBielan	0.1	93	13	930	130
pisorgpl	9.8	702	199	71	20

Rys. 18 Tabela zawierająca średnie liczby dotyczące tweetów związanych z Andrzejem Dudą.

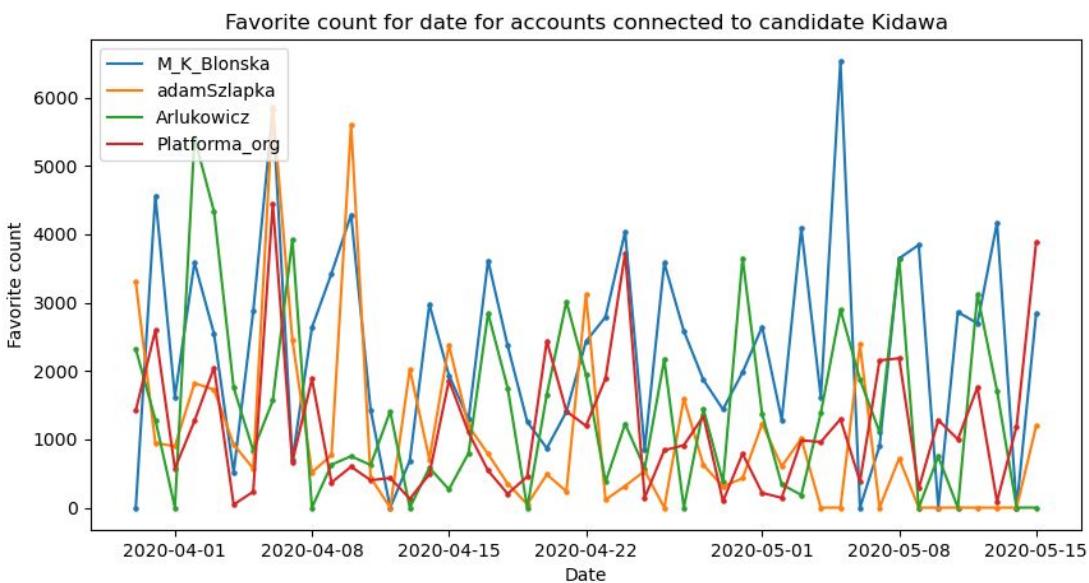
Analizując tabelę z rysunku 18 można stwierdzić, że kontem które tworzyło najczęściej tweetów było konto “pisorgpl”. Jest to konto partii PiS popierającej kandydaturę Andrzeja Dudy. Jednak to konto tworzyło nie tylko tweety dotyczące kampanii prezydenckiej, ale również zajmowało się innymi kwestiami. Dużo tweetów (średnio 7,1 dziennie) tworzyło konto “AndrzejDuda2020”. Jest to oficjalne konto kampanii urzędującego prezydenta. Warto zauważyć, że najczęściej polubień i retweetów

otrzymywało oficjalne konto kandydata - "AndzejDuda" mimo, że tworzyło mniej niż 1 tweet dziennie. Jest to zrozumiałe, ponieważ to konto ma największy zasięg. Można wysnuć hipotezę, że ludzie chętnie reagują na wpisy z konta kandydata.



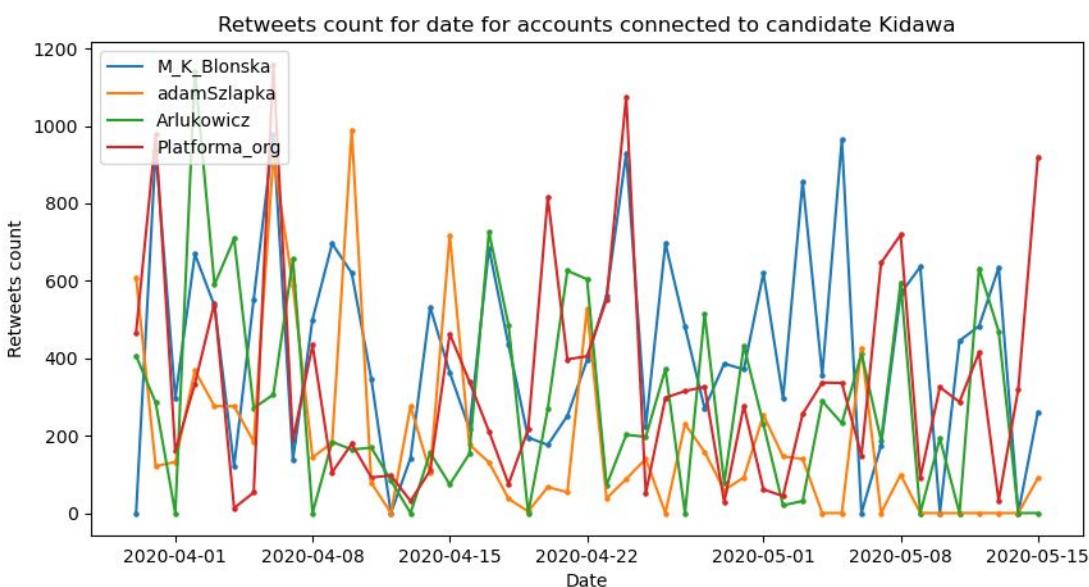
Rys.19 Wykres liczby tweetów tworzonych przez konta związkowe z Małgorzatą Kidawą-Błońską w zależności od czasu.

Analizując wykres z rys. 19, można zauważyć, że konta, które tworzyły najczęściej tweetów to konta "Platforma_org" - oficjalne konto partii PO popierającej Małgorzatę Kidawą-Błońską. Pozostałe konta związane z kandydatką tworzyły zdecydowanie mniej tweetów. Konto Małgorzaty Kidawy-Błońskiej "M_K_Blonska" tworzyło więcej tweetów niż konto Andrzeja Dudy, lecz również nie była to największa liczba w obozie związanym z kampanią kandydatki.



Rys.20 Wykres liczby polubień tweetów tworzonych przez konta związkowych z Małgorzatą Kidawą-Błońską w zależności od czasu.

Analizując wykres z rys. 20 można stwierdzić, że o ile na wykresie z rys. 19 w liczbie tweetów dominowało konto "Platforma_org", tak nie można zauważyc dominacji konta jeśli chodzi o liczbę otrzymywanych polubień. Widać, że tweety z konta "M_K_Blonska" otrzymywały dużo polubień. Jest to zrozumiałe, ponieważ prawdopodobnie użytkownicy częściej reagują na tweety z konta danego kandydata.



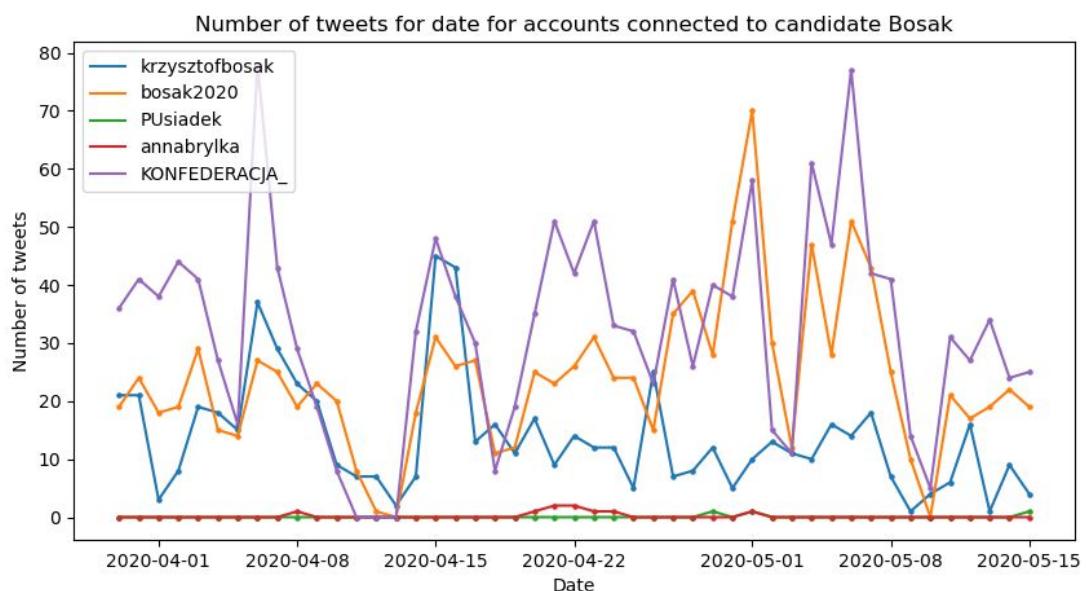
Rys. 21 Wykres liczby retweetów tweetów tworzonych przez konta związkowych z Małgorzatą Kidawą-Błońską w zależności od czasu.

Analizując wykres z rysunku 21, można stwierdzić, jak w przypadku analizy rys. 20 brak dominacji konkretnego konta jeśli chodzi o liczbę otrzymywanych retweetów przez tweety. Podobnie jak w przypadku polubień, dużo retweetów otrzymywały tweety z konta "M_K_Blonska".

Account	Avg tweets/day	Avg favorites/day	Avg retweets/day	Avg favorites/tweet	Avg retweets/tweet
M_K_Blonska	2.1	2320	425	1104	202
adamSzlapka	2.6	1026	189	394	72
Arlukowicz	2.1	1401	281	667	133
Platforma_org	13.9	1156	334	83	24

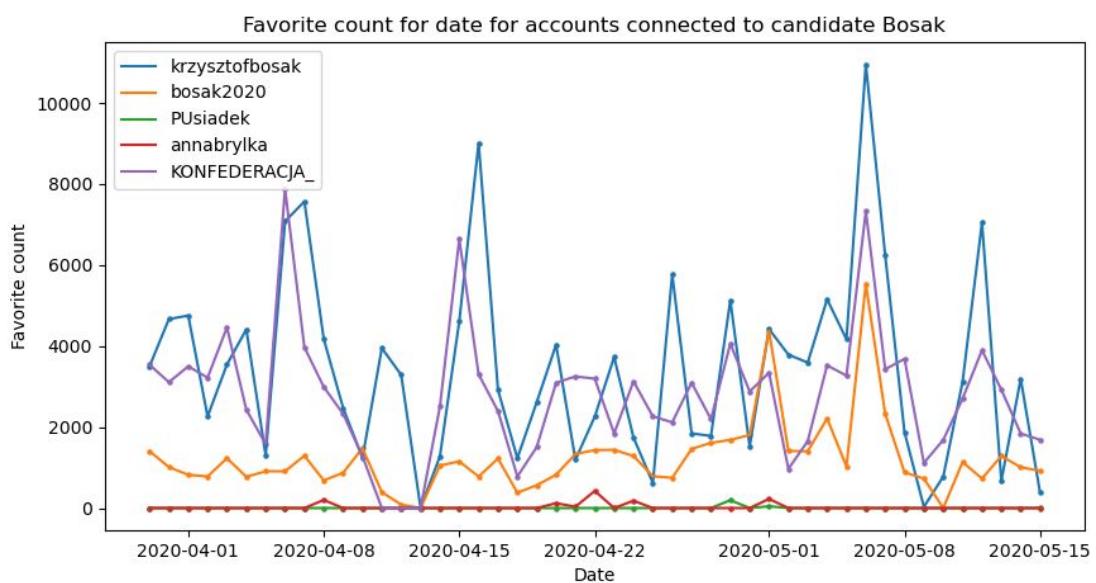
Rys. 22 Tabela zawierająca średnie liczby dotyczące tweetów związanych z Małgorzatą Kidawą-Błońską.

Analizując tabelę z rysunku 22 można stwierdzić, że kontem które tworzyło najwięcej tweetów było konto "Platforma_org". Jest to konto partii PO popierającej kandydaturę Małgorzaty Kidawy-Błońskiej. Jednak to konto tworzyło nie tylko tweety dotyczące kampanii prezydenckiej, ale również zajmowało się innymi kwestiami. Pozostałe konta związane z tą kandydaturą tworzyły średnio około 2 tweety dziennie. Najwięcej polubień i retweetów otrzymywały tweety pochodzące z konta "M_K_Blonska" - jest to oficjalne konto Małgorzaty Kidawy-Błońskiej.



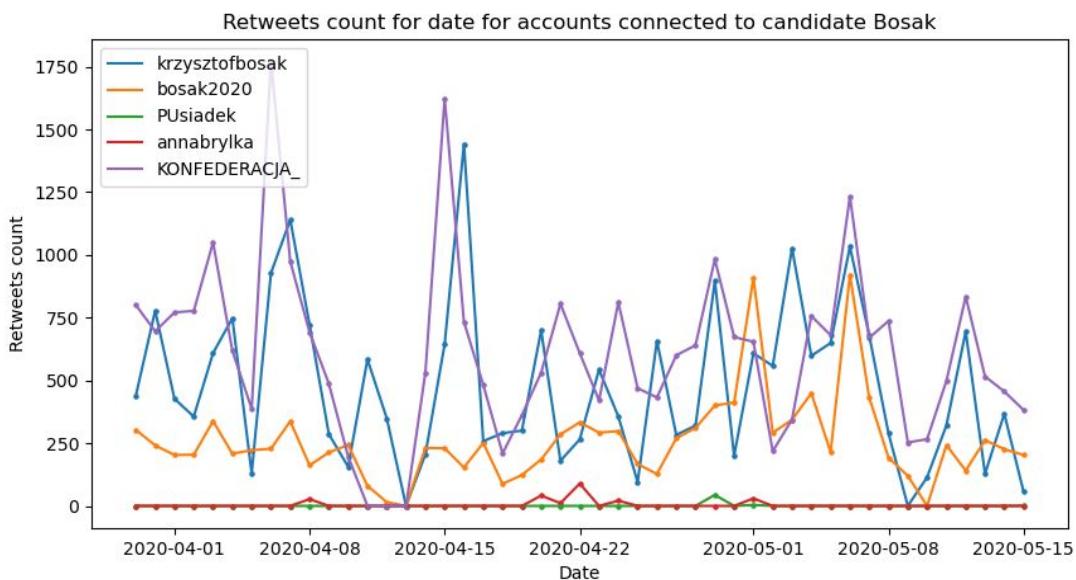
Rys. 23 Wykres liczby tweetów tworzonych przez konta związane z Krzysztofem Bosakiem w zależności od czasu.

Analizując wykres z rys. 23, można zauważyc, że konta, które tworzyły najwięcej tweetów to konta "KONFEDERACJA_" - oficjalne konto partii Konfederacja popierającej Krzysztofa Bosaka. Dużo tweetów pochodziło także z konta "bosak2020" - oficjalnego konta kampanii. Dużo tweetów szczególnie w porównaniu z Andrzejem Dudą pochodziło z konta "krzysztofbosak" - oficjalnego konta kandydata. Można domniemywać, że kandydat aktywnie zabiega o głosy.



Rys. 24 Wykres liczby polubień tweetów tworzonych przez konta związanych z Krzysztofem Bosakiem w zależności od czasu.

Analizując wykres z rys. 24, można zauważyc, że najwięcej polubień otrzymywały konta "KONFEDERACJA_" i "krzysztofbosak". Podobnie jak w przypadku Andrzeja Dudy i Małgorzaty Kidawy-Błońskiej, nie dziwi wysoka liczba reakcji na tweety kandydata, jednak zaskakująca jest duża liczba polubień tweetów z konta partii.



Rys. 25 Wykres liczby retweetów tweetów tworzonych przez konta związańych z Krzysztofem Bosakiem w zależności od czasu.

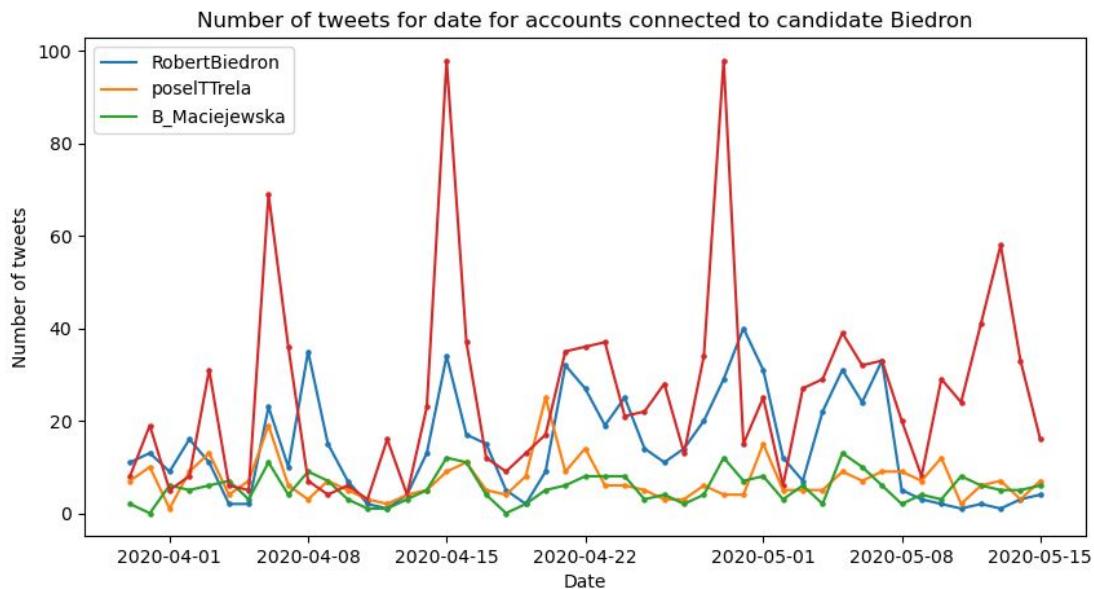
Analizując wykres z rys. 25, można zauważać, podobnie jak na wykresie z rys. 24, że najczęściej reakcji otrzymywały konta "KONFEDERACJA_" i "krzysztofbosak".

Account	Avg tweets/day	Avg favorites/day	Avg retweets/day	Avg favorites/tweet	Avg retweets/tweet
krzysztofbosak	13.6	3423	476	251	35
bosak2020	23.9	1216	257	50	10
PUsiadek	0.1	5	1	50	10
annabrylka	0.2	25	4	125	20
KONFEDERACJA_	32.3	2799	609	86	18

Rys. 26 Tabela zawierająca średnie liczby dotyczące tweetów związanych z Krzysztofem Bosakiem.

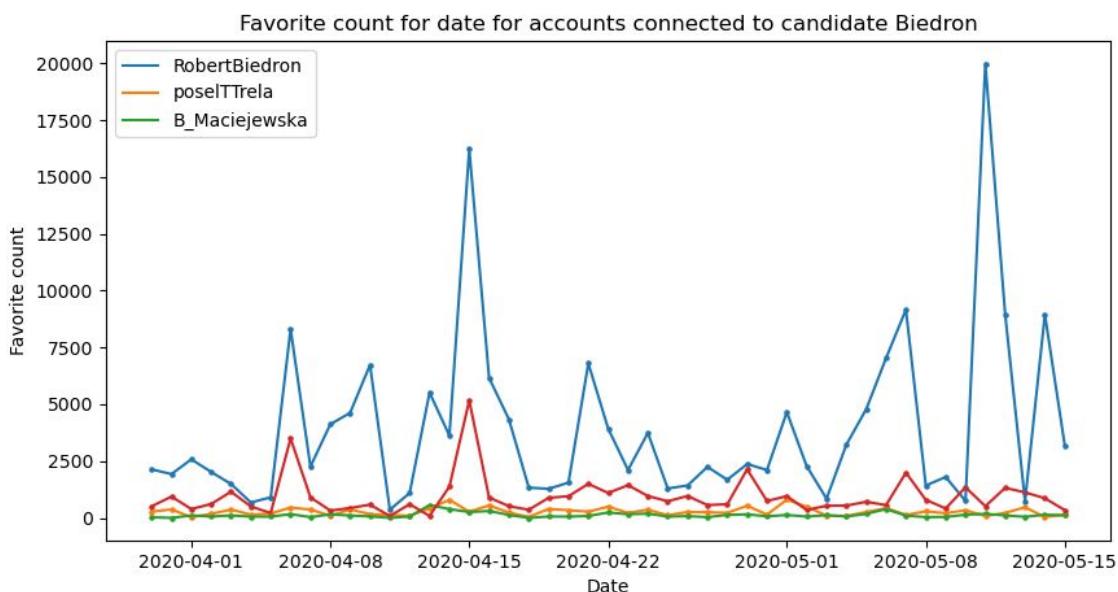
Analizując tabelę z rysunku 26 można stwierdzić, że kontem które tworzyło najczęściej tweetów było konto "KONFEDERACJA_" (około 32 tweety dziennie). Jest to konto partii Konfederacja popierającej kandydaturę Krzysztofa Bosaka. Jednak to konto tworzyło nie tylko tweety dotyczące kampanii prezydenckiej, ale również zajmowało się innymi kwestiami. Dużo tweetów (prawie 24 dziennie) tworzyło oficjalne konto kampanii K. Bosaka - "bosak2020". Można zauważać również bardzo dużą aktywność oficjalnego konta kandydata "krzysztofbosak" - ponad 13 tweetów dziennie. Tweety pochodzące z tego konta dostawały dużo polubień i retweetów w przeliczeniu na dzień. Jednak w przeliczeniu na tweet, widać, że reakcja na treści

nie była imponująca (średnio 251 polubień i 35 retweetów). Są to liczby znacznie mniejsze od tych które notowali Andrzej Duda czy Małgorzata Kidawa-Błońska.



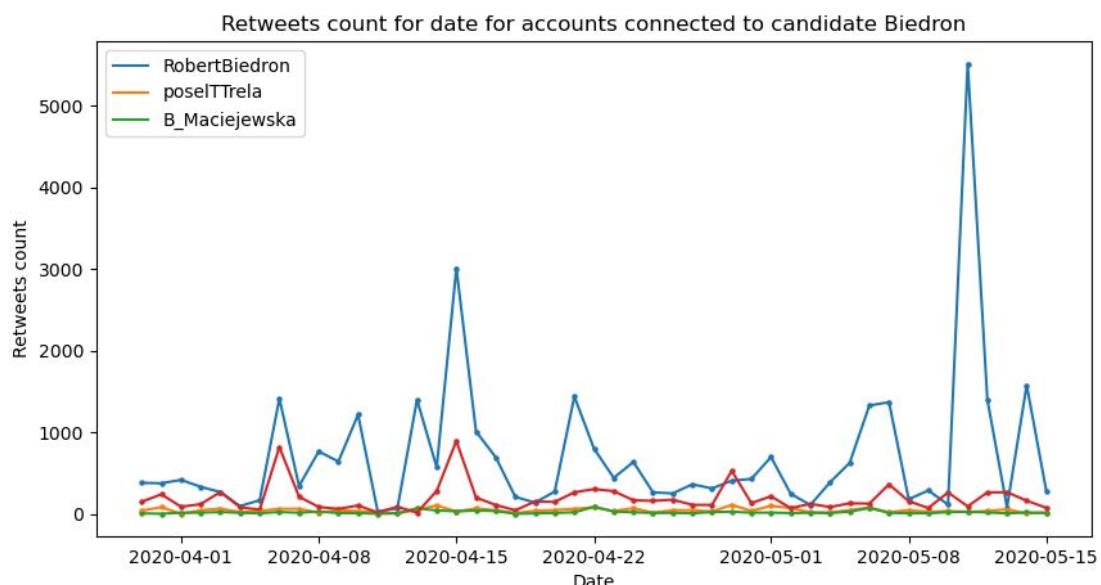
Rys. 27 Wykres liczby tweetów tworzonych przez konta związkowych z Robertem Biedroniem w zależności od czasu.

Analizując wykres z rys. 27. można zauważyć, że najczęściej tweetów pochodziło z konta “__Lewica” - jest to konto partii Lewica popierającej Roberta Biedronia. Aktywny był także sam kandydat - jego konto “RobertBiedron” stworzyło do 40 tweetów dziennie. Można zauważać pewne pikи aktywności w niektórych dniach co było zapewne spowodowane wydarzeniami politycznymi.



Rys. 28 Wykres liczby polubień tweetów tworzonych przez konta związań Robertem Biedroniem w zależności od czasu.

Analizując wykres z rys. 25, można zauważyć, że najwięcej polubień otrzymywały konta tweety z konta "RobertBiedron". Zaskakująca jest dominacja tego konta chociażby nad kontem "__Lewica", które jak wynika z rys. 24 tworzyło najczęściej tweetów. Może to świadczyć o tym, że konto partii ma niewielkie zasięgi, przez co tweety nie trafiają do użytkowników zainteresowanych kandydaturą Roberta Biedronia.



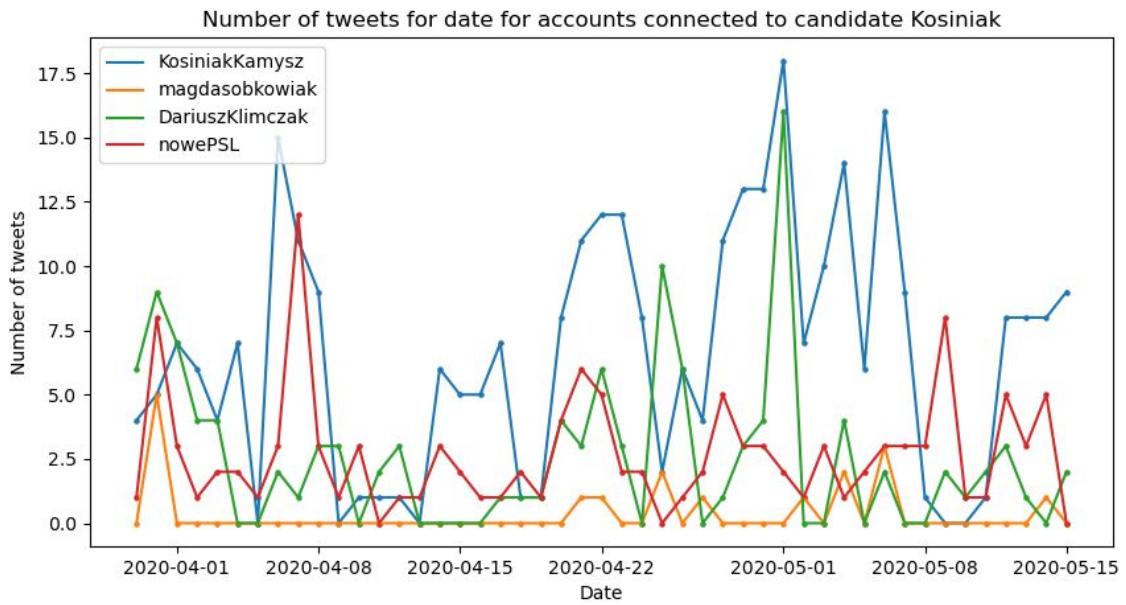
Rys. 29 Wykres liczby retweetów tweetów tworzonych przez konta związań Robertem Biedroniem w zależności od czasu.

Analizując wykres z rys. 29, można zauważyc, że podobnie jak na wykresie polubień z rys. 28, najwięcej retweetów otrzymywały tweety z konta "RobertBiedron". Pozostałe konta otrzymywały zdecydowanie mniej reakcji. Zdecydowanie najsilniejszym punktem kampanii Roberta Biedronia na Twitterze pod względem reakcji jest zatem sam kandydat.

Account	Avg tweets/day	Avg favorites/day	Avg retweets/day	Avg favorites/tweet	Avg retweets/tweet
RobertBiedron	14.2	3931	707	276	49
poselITrela	7.1	280	41	39	5
B_Maciejewska	5.4	123	18	22	3
_Lewica	25.4	937	187	36	7

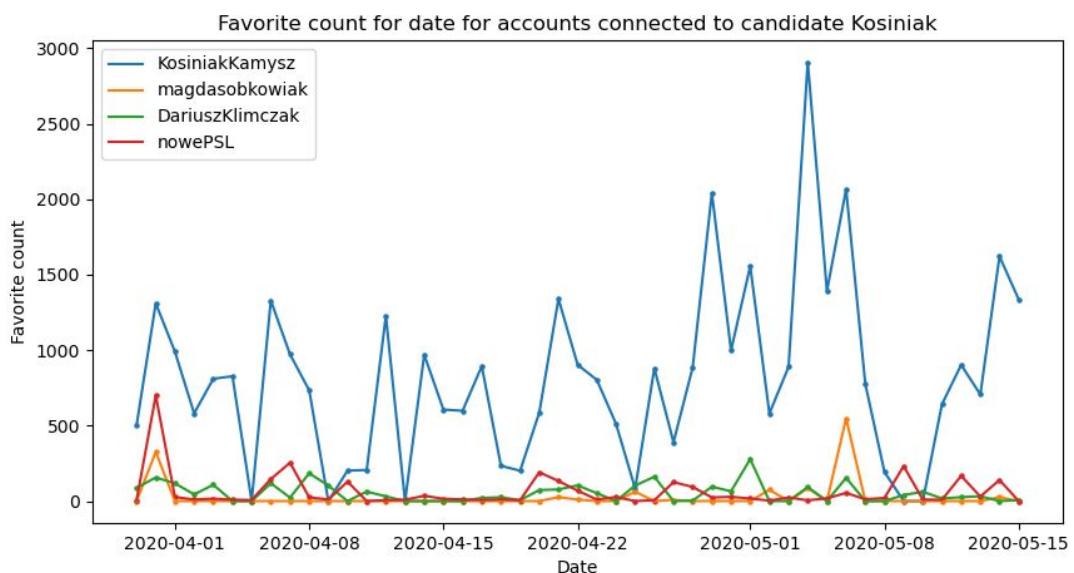
Rys. 30 Tabela zawierająca średnie liczby dotyczące tweetów związanych z Robertem Biedroniem.

Analizując tabelę z rysunku 30 można stwierdzić, że kontem które tworzyło najwięcej tweetów było konto "__Lewica" (około 25 tweety dziennie). Jest to konto partii Lewica popierającej kandydaturę Roberta Biedronia. Jednak to konto tworzyło nie tylko tweety dotyczące kampanii prezydenckiej, ale również zajmowało się innymi kwestiami. Dużo tweetów (prawie 24 dziennie) tworzyło oficjalne konto Roberta Biedronia - "RobertBiedron" (ponad 14 tweetów dziennie). Widać, że tweety pochodzące z tego konta dostawały najwięcej polubień i retweetów w porównaniu z innymi kontami związanymi z kampanią Roberta Biedronia. Podobnie jak przy tweetach związanych z Krzysztofem Bosaka, jest to mniejsza ilość polubień i retweetów niż w przypadku Andrzeja Dudy i Małgorzaty Kidawy-Błońskiej.



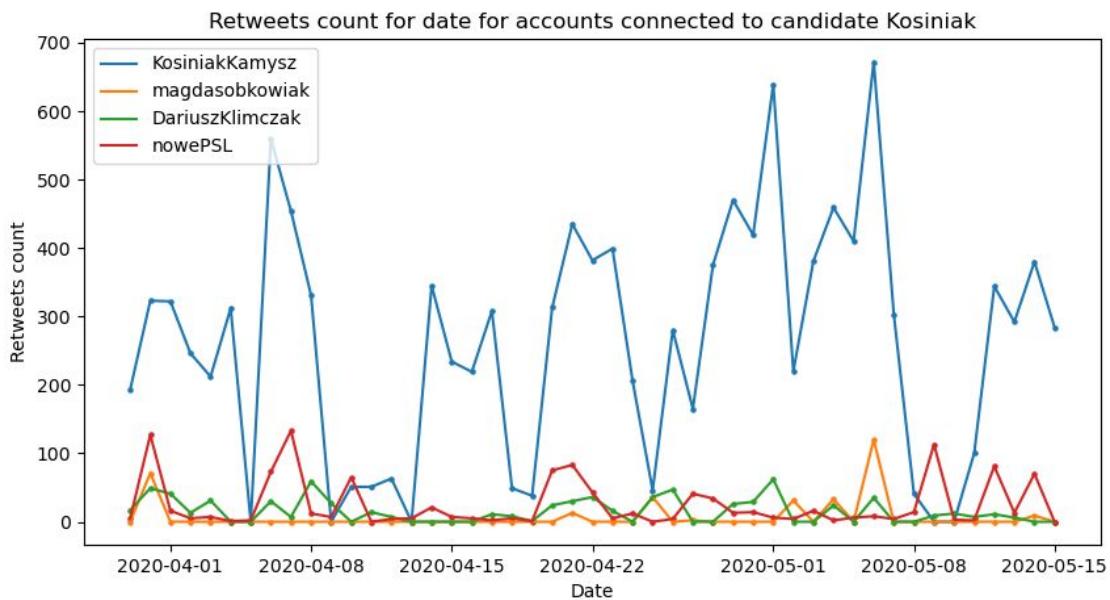
Rys. 31 Wykres liczby tweetów tworzonych przez konta związkowych z Władysławem Kosiniakiem-Kamyszem w zależności od czasu.

Analizując wykres z rys. 31, można zauważyć, że najwięcej tweetów pochodziło z konta "KosiniakKamysz" - oficjalnego konta Władysława Kosiniaka-Kamysza. Zastanawiająca jest niska aktywność konta "nowePSL" - konta partii PSL popierającej tego kandydata. W porównaniu z pozostałymi kontami partii, liczba tweetów jest bardzo niska.



Rys. 32 Wykres liczby polubień tweetów tworzonych przez konta związkowych z Władysławem Kosiniakiem-Kamyszem w zależności od czasu.

Analizując wykres z rys. 32, można zauważyć, że najwięcej polubień otrzymywały tweety z konta "KosiniakKamysz". W porównaniu z poprzednimi kandydatami, te liczby nie są jednak imponujące. Widać również, że konto "nowePSL" nie jest chętnie odbierane. Mała liczba tweetów w przypadku tego konta oznacza bardzo małą liczbę polubień.



Rys. 33 Wykres liczby retweetów tweetów tworzonych przez konta związanych z Władysławem Kosiniakiem-Kamyszem w zależności od czasu.

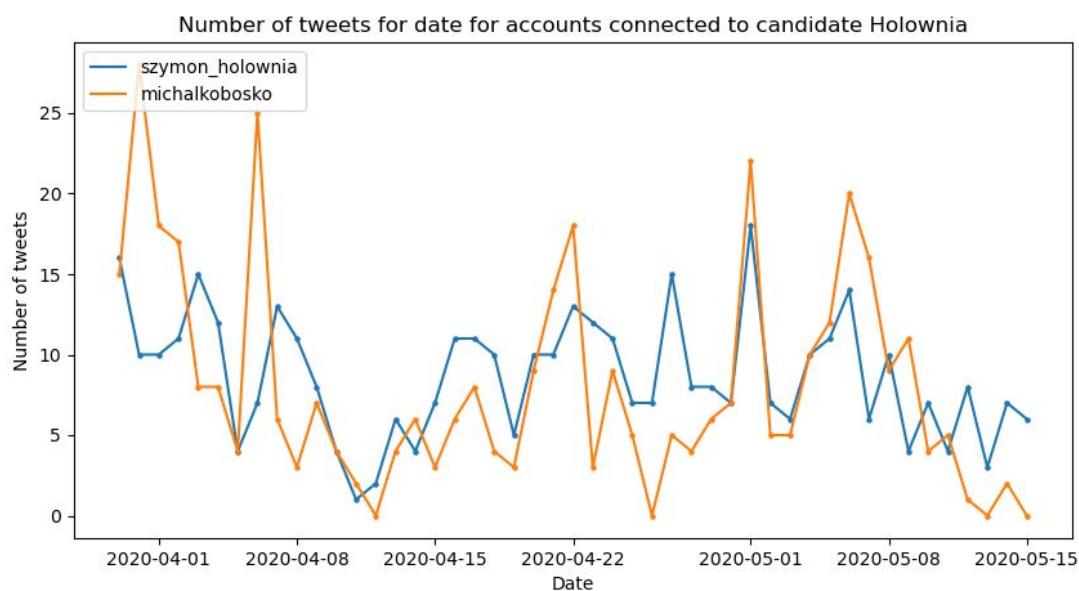
Analizując wykres z rys. 33, można zauważyć, że najwięcej retweetów otrzymywały tweety z konta "KosiniakKamysz". Podobnie jak na wykresie z rys. 32, pozostałe konta otrzymywały mało reakcji.

Account	Avg tweets/day	Avg favorites/day	Avg retweets/day	Avg favorites/tweet	Avg retweets/tweet
KosiniakKamysz	6.6	812	262	123	39
magdasobkowiak	0.4	24	6	60	15
DariuszKlimczak	2.6	53	15	20	5
nowePSL	2.7	61	24	22	8

Rys. 34 Tabela zawierająca średnie liczby dotyczące tweetów związanych z Władysławem Kosiniakiem-Kamyszem.

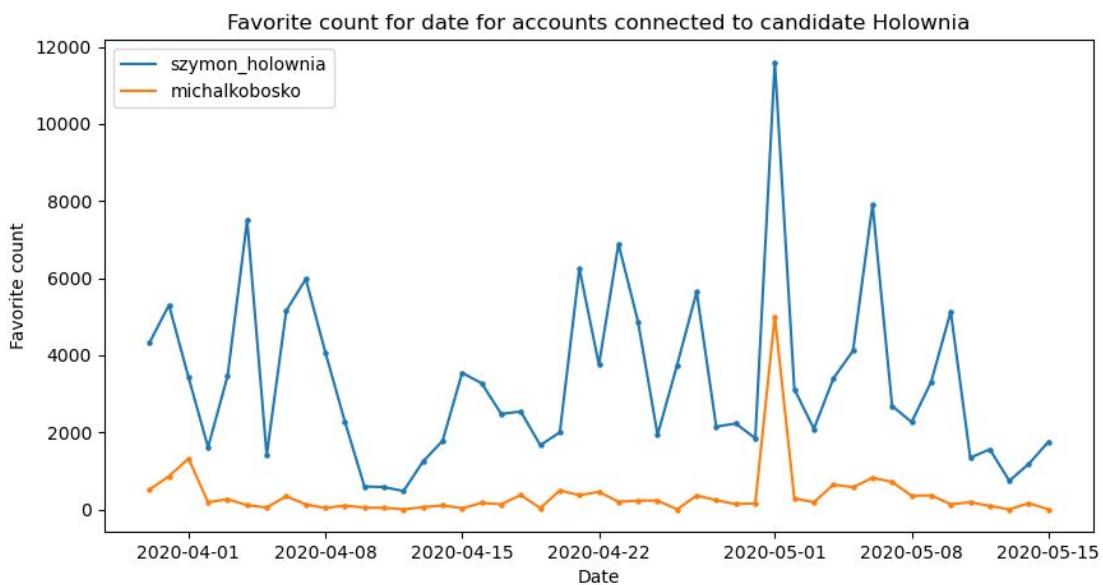
Analizując tabelę z rysunku 34 można stwierdzić, że kontem które tworzyło najwięcej tweetów było konto "KosiniakKamysz" (około 6 tweetów dziennie). Jest to oficjalne konto Władysława Kosiniaka-Kamysza. Zastanawiająca jest niska aktywność konta "nowePSL" - jest to konto partii PSL która popiera kandydata. Konto to tworzyło

średnio mniej niż 3 tweety dziennie. Mała liczba tweetów nie oznaczała lepszego odbioru - tweety dostawały średnio 22 polubienia i 8 retweetów co jest wartością marginalną. Konto "KosiniakKamysz" jest najsielniejszym ogniwem Władysława Kosiniaka-Kamysza na Twitterze, jednak w porównaniu z opisanymi pozostałymi kandydatami nie imponuje liczbą polubień i retweetów (średnio 123 polubienia i 39 retweetów na tweet). Pozostałe konta związane z tym kandydatem stanowią niewielką siłę na Twitterze. Można wysnuć hipotezę, że w kampanii Władysława Kosiniaka-Kamysza nie postawiono na aktywność na tym medium społecznościowym.



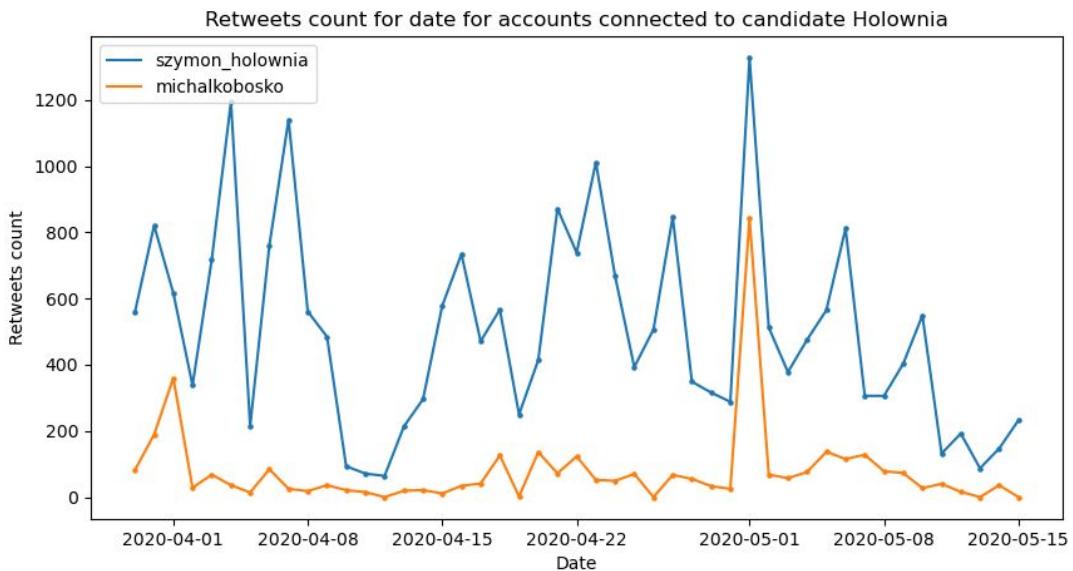
Rys. 35 Wykres liczby tweetów tworzonych przez konta związane z Szymonem Hołownią w zależności od czasu.

Analizując wykres z rys. 35 można stwierdzić, że zarówno konto "szymon_holownia" - oficjalne konto kandydata jak i "michalkobosko" - konto pełnomocnika wyborczego tworzyły podobne liczby tweetów. W porównaniu z innymi kandydatami konto "michalkobosko" to pierwsze konto ze sztabu kandydata, które tak aktywnie włączyło się w kampanię na Twitterze.



Rys. 36 Wykres liczby polubień tweetów tworzonych przez konta związkanych z Szymonem Hołownią w zależności od czasu.

Analizując wykres z rys. 36 można stwierdzić, że widać dysproporcję w liczbie polubień w między kontami w porównaniu z wykresem z rys. 35. Zdecydowanie więcej polubień otrzymały tweety Szymona Hołowni, co nie dziwi, ponieważ jest to kandydat obecny w mediach. Michał Kobosko jest prawdopodobnie osobą anonimową dla większości osób.



Rys. 37 Wykres liczby retweetów tweetów tworzonych przez konta związkanych z Szymonem Hołownią w zależności od czasu.

Analizując wykres z rys. 37 można stwierdzić, że podobnie jak na wykresie z rys. 36 najwięcej retweetów otrzymywały tweety z konta “szymon_holownia”, choć różnica między kontem “michalkobosko” wydaje się zdecydowanie mniejsza.

Account	Avg tweets/day	Avg favorites/day	Avg retweets/day	Avg favorites/tweet	Avg retweets/tweet
szymon_holownia	8.7	3321	501	381	57
michalkobosko	8.1	366	76	45	9

Rys. 38 Tabela zawierająca średnie liczby dotyczące tweetów związanych z Szymonem Hołownią.

Analizując tabelę z rysunku 38 można stwierdzić, że kontem które tworzyło najwięcej tweetów było konto “szymon_holownia” (około 9 tweetów dziennie). Jest to oficjalne konto Szymona Hołowni. Podobną aktywnością wykazuje się również “michalkobosko” - konto pełnomocnika wyborczego. Widać również, że tweety tworzone przez Szymona Hołownię mają dość dobry odbiór - średnio tweet dostawał około 380 polubień i prawie 60 retweetów co przy tej liczbie tweetów składa się na ponad 3000 polubień i 500 retweetów dziennie. Nie są to oczywiście liczby tak duże jak w przypadku kont Andrzeja Dudy czy Małgorzaty Kidawy Błońskiej, jednak biorąc pod uwagę brak wsparcia w postaci partii politycznej, można uznać aktywność Szymona Hołowni za dość dużą.

- Najbardziej popularny tweet dla każdego kandydata(oddziennie mierzone dla ilości retweetów jak i polubień)
- Porównanie łącznej liczby tweetów między kandydatami
- Porównanie łącznej liczby retweetów postów kandydatów
- Porównanie łącznej liczby polubień postów kandydatów

Powyższe liczby (retweetów i polubień) możemy porównać z zasięgami danego użytkownika (liczba followersów), aby stwierdzić, który z członków sztabu jest najbardziej zaangażowany w kampanię.

5. Wyniki analiz tekstu

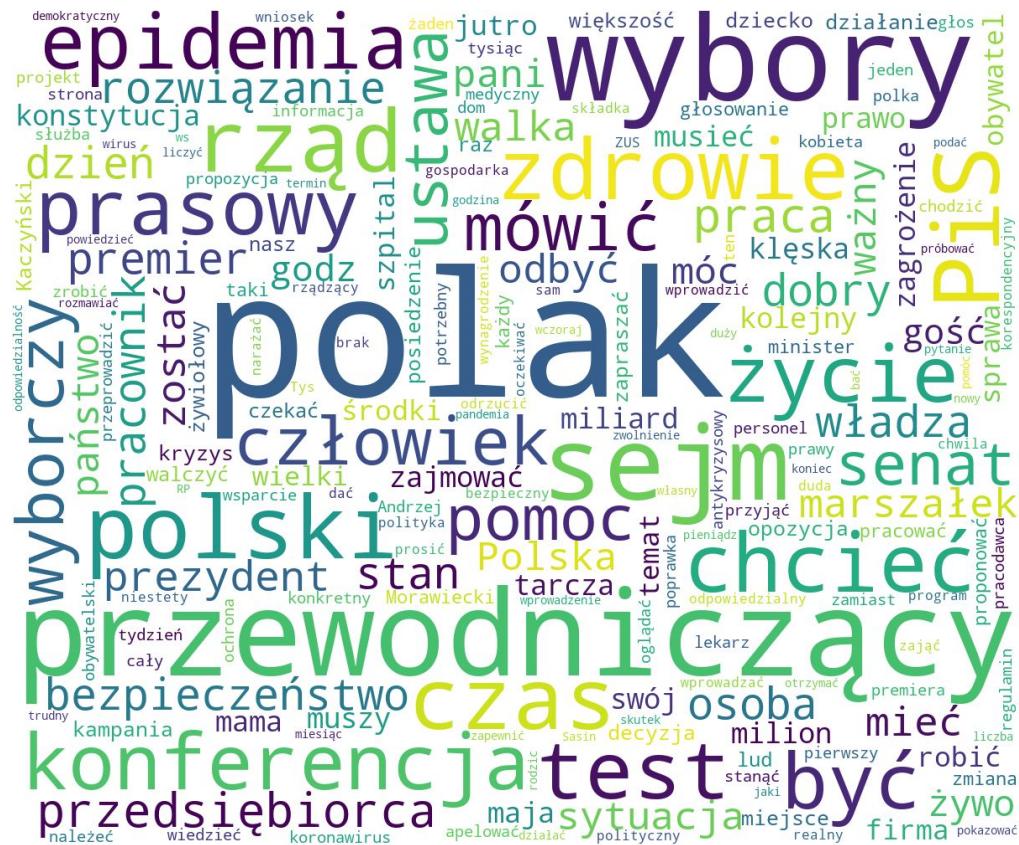
Ten rozdział poświęcony został analizie tekstu. W ramach tej analizy przetwarzaliśmy dane by uzyskać:

- Najpopularniejsze hashtagi używane przez kandydatów
 - Najpopularniejsze słowa używane przez kandydatów
 - Zmierzenie ilości odpowiedzi (reply) kandydatów do innych użytkowników Twittera
 - Porównanie odległości wektorów słów pomiędzy kandydatami za pomocą miary kosinusowej

Analizy polegają na ekstrakcji pełnej tekstuowej zawartości tweeta przez API, odfiltrowaniu emotikonów, średników, przecinków; usunięciu tzw stopwords (i, ale, jednakże itd) oraz lematyzacji, czyli doprowadzeniu słów do ich formy podstawowej. Następnie można przejść do właściwej analizy. Dzisiaj chętnie stosowaną formą przedstawienia analiz tekstowych są “wordclouds” czyli obraz zawierający słowa których wielkość jest wprost proporcjonalna do popularności słowa w zbiorze. Dla każdego kandydata została wygenerowana chmura słów:



Rys. 39 Chmura słów dla tweetów kont związanych z Andrzejem Dudą



Rys.40 Chmura słów dla tweetów kont związanych z Małgorzatą Kidawą-Błońską



Rys 41. Chmura słów dla tweetów kont związanych z Władysławem Kosiniakiem-Kamyszem



Rys. 42 Chmura słów dla tweetów kont związańych z Krzysztofem Bosakiem



Rys.43 Chmura słów dla tweetów kont związańych z Robertem Biedroniem



Rys.44 Chmura słów dla tweetów kont związanych z Szymonem Hołownią

Co zadziwiające, to mimo iż tweety były zbierane już w momencie wystąpienia epidemii, to nie jest ona nawet jednym z popularniejszych tematów co widać na obrazkach nr 39-44. Wordcloudy zostały wykonane także dla popularnych tagów. Ponadto dla każdej analizy tekstu zostały także wykonane wykresy słupkowe a także dane zostały przedstawione w tabelach

Word cloud nie jest dokładny dlatego w celu poznania realnego liczbowego podobieństwa wypowiedzi między kandydatami wykorzystana jest miara kosinusowa odległości wektorów słów, które to słowa są inną formą tweetów. Wszystkie tweety są zamieniane do listy słów, a następnie łączone w wektor zawierający w sobie słowo wraz z liczbą wystąpień.

```

import operator
sorted_sim = sorted(similarity.items(), key=operator.itemgetter(1), reverse=True)
sorted_sim

[('Kidawa-Biedron', 0.8099719524979767),
 ('Kosiniak-Biedron', 0.7675138478990894),
 ('Kidawa-Kosiniak', 0.7563787367217715),
 ('Biedron-Holownia', 0.748066004019902),
 ('Kidawa-Holownia', 0.7281915074023474),
 ('Kosiniak-Holownia', 0.727924632174207),
 ('Bosak-Biedron', 0.6985407490115011),
 ('Kidawa-Bosak', 0.6792682827780961),
 ('Bosak-Holownia', 0.666420823644531),
 ('Bosak-Kosiniak', 0.6518630702823902),
 ('Duda-Biedron', 0.5510850378714002),
 ('Duda-Kosiniak', 0.5489735866572103),
 ('Duda-Holownia', 0.5464571510332731),
 ('Duda-Kidawa', 0.5080744982755042),
 ('Duda-Bosak', 0.4685629370099766)]

```

Rys. 45 Podobieństwo wypowiedzi między kandydatami

Rysunek 45 prezentuje efekt porównania tekstów tweetów kandydatów w postaci wektorów słów miarą kosinusową. Jest to wartość z przedziału [0-1] Im większa wartość, tym wybrane wektory są bardziej podobne. Chmury słów pozwalają w sposób empiryczny stwierdzić podobieństwa między tweetami kandydatów (największe słowa między chmurami się powtarzają), jednak miara kosinusowa między wektorami słów tweetów kandydatów pokazuje jasno znaczące (w przypadku kandydatów Kidawy i Biedronia to aż 0.8) podobieństwa między ich wypowiedziami. Najbardziej wyróżniają się Duda i Bosak, mają oni najmniejsze zbieżności wektorów z resztą. Wzór na podobieństwo wektorów za pomocą miary kosinusowej:

$$sim(v1, v2) = \frac{v1 \cdot v2}{|v1| * |v2|}$$

jest to wzór na znormalizowany iloczyn skalarny wektorów. Współrzędne wektora tekstopewego w przestrzeni są zdefiniowane jako częstość występowania słów, a długość to ilość występujących w wektorze słów.

Narzędzia możliwe do wykorzystania do analizy tekstu:

- Morfeusz - lemmatizer języka polskiego, sprowadza on słowa do ich formy podstawowej np. premier ->premier, by potem móc je zgrupować i ułatwić dalsze przetwarzanie. Program Morfeusz wykonuje analizę morfologiczną dla języka polskiego. W obecnej wersji nie zawiera modułu zgadującego nieznane słowa (można więc powiedzieć, że jest słownikiem morfologicznym). Wykorzystuje słownik fleksyjny Polimorf stanowiący połączenie danych SGJP z tworzonymi społecznościowo danymi Morfologika/sjp.pl. Jest on jednym z najpopularniejszych do tej pory narzędzi, choć powoli popularne biblioteki (np Scapy) pracują nad obsługą polskiego

języka, którego trudność wynika z fleksyjności (wiele wersji słowa z uwagi przez wiele form odmiany każdego z nich)^[12].

- NLTK - biblioteka służąca zgodnie z nazwą do przetwarzania języka naturalnego (natural language toolkit), w tym przypadku służy do odfiltrowania tzw stopwords. Jest bardzo popularna z uwagi na mnogość zastosowań. Pozwala na analizy statystyczne, graficzne przedstawianie wyników, analizę tekstu. NLTK zawiera demonstracje graficzne i przykładowe dane. Autorzy przygotowali również książkę opisującą pojęcia wykorzystywane w NLTK. Niestety na ten moment jej zastosowanie w projekcie jest minimalne z uwagi na brak wsparcia dla bardziej złożonych operacji w języku polskim^[13].
- Wordcloud python - narzędzie do tworzenia chmury słów w języku Python. Pozwala na tworzenie chmury w różnych kształtach (maska o dowolnym kształcie np. logo firmy) i wersjach kolorystycznych. Choć zastosowanie w celach badawczych jest znikome z uwagi na brak dokładności, tak graficzne przedstawienie danych dla zwykłego użytkownika jest ciekawsze i przyjemniejsze. Projekt jest na licencji open source a autor Andreas Mueller pozwala go edytować do własnych celów^[14].

6. Wyniki analiz społeczności

Analiza relacji/sieci społecznej (SNA - Social Network Analysis) to badanie zbiorowości poprzez analizowanie relacji zachodzących pomiędzy jej elementami. Badana sieć jest przedstawiana w postaci grafu, a więc sieci wierzchołków, reprezentujących elementy zbiorowości, i sieci krawędzi, czyli relacji między nimi^[11].

Podstawowymi miarami branymi pod uwagę przy analizie sieci społecznej są^[15]:

- rozmiar sieci - liczba jej elementów
- liczba, ukierunkowanie, odwzajemnianie oraz przechodniość powiązań pomiędzy elementami sieci
- centralność - zbiór metryk używanych do mierzenia istotności i wpływu elementów sieci:
 - stopień - liczba bezpośrednich powiązań danego elementu (badanie istotności i aktywności w sieci)
 - bliskość - średnia długość najkrótszych ścieżek od danego elementu do wszystkich innych w sieci (badanie efektywności)
 - wektor własny - średnia odległość od innych elementów w sieci, ale brana pod uwagę jest istotność elementów, do których istnieją połączenia poprzez nadanie im wag (badanie wpływu w sieci)
 - pośrednictwo - liczba połączeń między elementami sieci przebiegających przez dany obiekt (badanie kontroli w sieci)
- gęstość powiązań - stosunek liczby bezpośrednich połączeń istniejących w sieci do liczby wszystkich możliwych
- pomosty - elementy zapewniające jedyne połączenie pomiędzy dwoma innymi elementami czy grupami

Narzędzia możliwe do wykorzystania do analizy sieci społecznej:

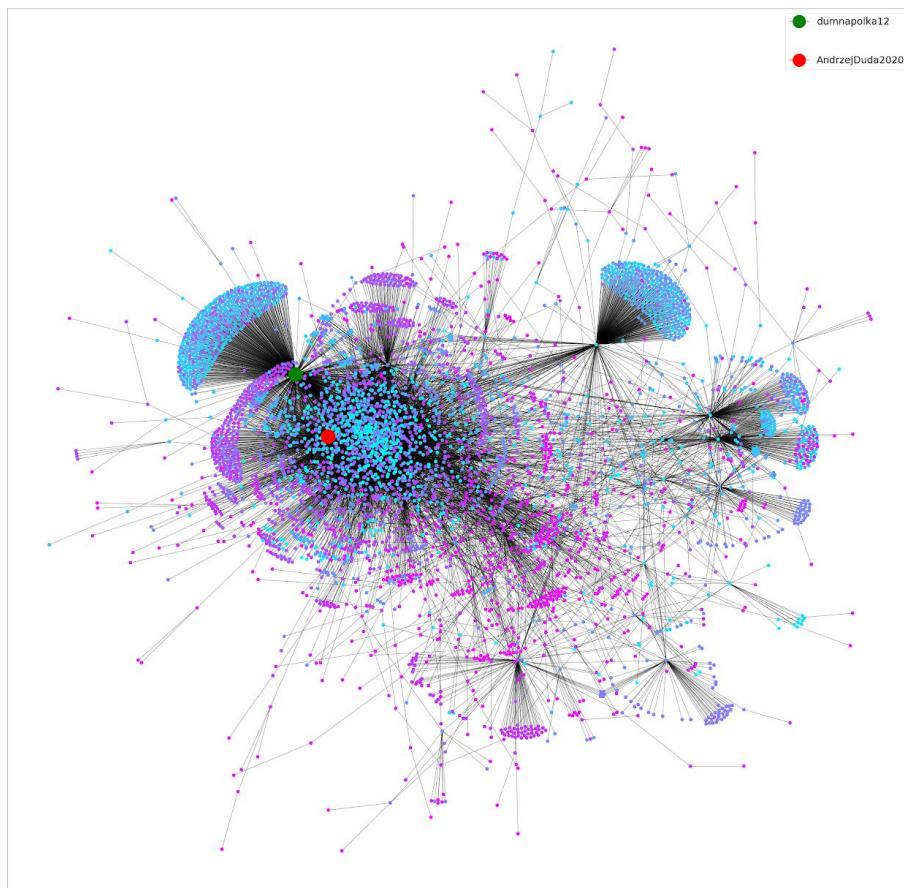
- NetworkX - jest to czysto Pythonowa biblioteka służąca do tworzenia i badania struktury, dynamiki oraz funkcji sieci złożonych. Umożliwia użycie wielu standardowych algorytmów grafowych oraz badanie sieci za pomocą różnych metryk. Dodatkowo umożliwia bezpośrednie importy z pandas Dataframe oraz działa z matplotlib. Jest dobrze udokumentowana^[5].

W ramach SNA wykonane zostały następujące analizy:

- Grafy "społeczności" dla każdego kandydata
- Graf społeczności

6.1 Sieć Andrzeja Dudy

Sieć społeczna stworzona przez użytkowników Twittera, którzy w swoich wypowiedziach używali hashtagów dotyczących Andrzeja Dudy:



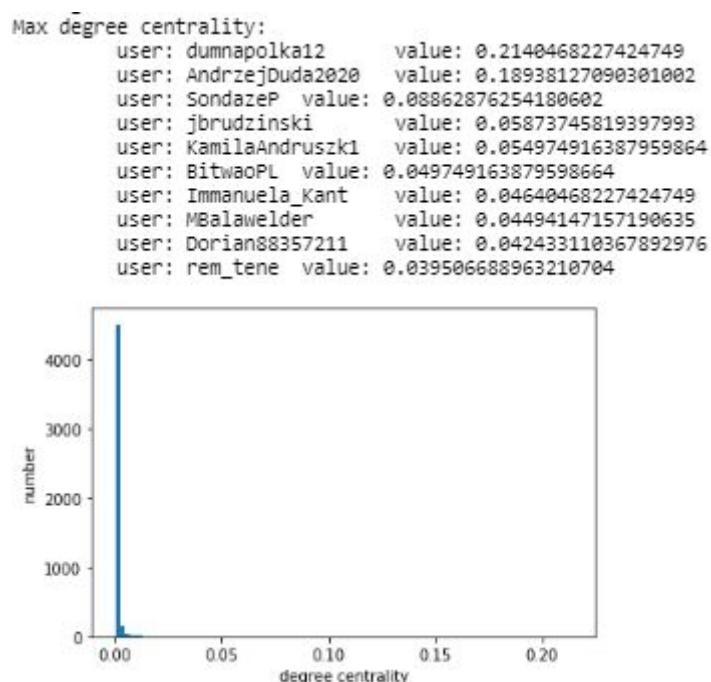
Rys. 46 Sieć ludzi tweetujących o Andrzeju Dudzie
Podstawowe własności grafu:

```
Number of nodes: 5040
Number of edges: 11113
Max degree: 1024
Min degree: 1
Density: 0.000875157895400007
Graph connected: False
Number of connected components: 115
Largest connected component:
Number of nodes: 4785
Number of edges: 10972
Max degree: 1024
Min degree: 1
Density: 0.0009586116032892644
Average clustering coefficient: 0.07261295458020314
Transitivity: 0.010038099255636127
Diameter: 10
Average distance between two nodes: 3.86
```

Rys. 47 Podstawowe własności grafu ludzi tweetujących o Andrzeju Dudzie

Jak widać na rys. 23 użytkownicy, których tweety pobraliśmy, nie stworzyli jednego spójnego grafu, jednakże jego największa wspólna składowa zawiera w sobie większość jego wierzchołków i krawędzi i ona została wybrana do dalszej analizy. Graf, który powstał nie jest gęsty, ma również niski współczynnik grupowania. Średnica grafu i średnia odległość między wierzchołkami nie są wysokie.

Miary centralności w grafie:



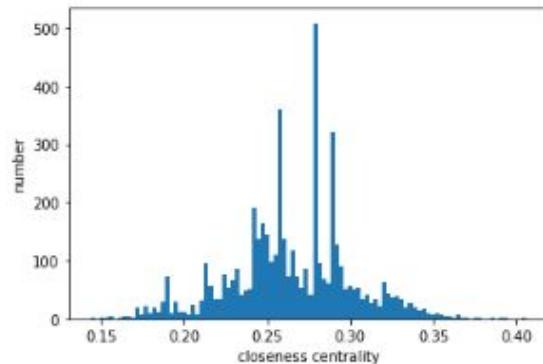
Rys. 48 Użytkownicy według stopnia w grafie Andrzeja Dudy

Na rys. 48 można zauważyć, że prawie wszyscy użytkownicy mają bardzo niski i zbliżony do siebie stopień w sieci. Wyróżniają się jedynie oficjalne konto kampanii (AndrzejDuda2020), co jest zrozumiałe, oraz konto zwolenniczki obecnego prezydenta (dumnapolka12), co jest bardziej zaskakujące. Konto kampanii ma większe zasięgi, ale użytkownik dumnapolka12 publikuje znacznie więcej tweetów o Andrzeju Dudzie.

```

Max closeness centrality:
user: AndrzejDuda2020    value: 0.40545808966861596
user: PiotrZiba11         value: 0.393615270692776
user: zdybell             value: 0.39177790516747196
user: dumnapolka12        value: 0.38593094546627943
user: Magdalena_C22       value: 0.37842113589621895
user: ptrembec            value: 0.37556916313392996
user: MosinskiJan          value: 0.3754807314967428
user: Tadeusz87434686     value: 0.3702786377708978
user: ZofiaB53             value: 0.369449378330373
user: Adam38678946         value: 0.3669837373427432

```



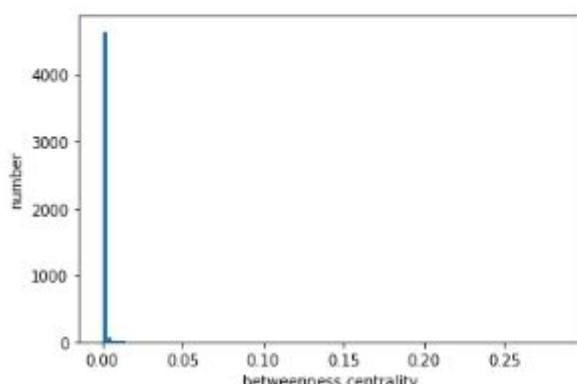
Rys. 49 Użytkownicy według bliskości w grafie Andrzeja Dudy

Na rys. 49 można zauważyc, że miara bliskości użytkowników w sieci jest już bardziej równomiernie rozłożona i żaden z nich pod tym względem się nie wyróżnia. Wynika z tego, że większość użytkowników ma dosyć krótkie średnie odległości od pozostałych, w tym przypadku większość użytkowników znajduje się w centrum sieci.

```

Max betweenness centrality:
user: dumnapolka12      value: 0.28413580428855967
user: AndrzejDuda2020    value: 0.24941387329165923
user: SondazeP            value: 0.1735689023698662
user: zdybell              value: 0.1003790041101001
user: MBalawelder         value: 0.06683389924379561
user: Dorian88357211      value: 0.0633808057723557
user: TomaszPanflic86     value: 0.05908586818976969
user: BitwaoPL             value: 0.04629091286538178
user: jbrudzinski           value: 0.04130371756098268
user: Bart_Wielinski        value: 0.037663878542392976

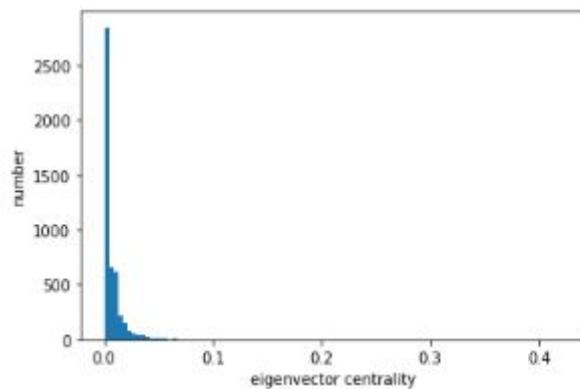
```



Rys. 50 Użytkownicy według pośrednictwa w grafie Andrzeja Dudy

Na rys. 50 można zauważyć, że podobnie jak w przypadku stopnia, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość pośrednictwa w sieci. Tutaj również wyróżniają się konta AndrzejDuda2020 i dumnapolka12, a także SondazeP - konto przeprowadzające sondaże w internecie i publikujące informacje o kampanii. Wynika z tego, że wiele połączeń w sieci przechodzi właśnie poprzez te konta.

```
Max eigenvector centrality:  
    user: AndrzejDuda2020  value: 0.42074457615216154  
    user: dumnapolka12    value: 0.33959801992989347  
    user: jbrudzinski      value: 0.14877556228611313  
    user: Tulajew         value: 0.12218145163137095  
    user: KamilAndruszki1 value: 0.11706517406053853  
    user: MosinskiJan     value: 0.11190068386907853  
    user: Immanuel_Kant   value: 0.10214153280676908  
    user: PiS_Slaskie     value: 0.09624255886872417  
    user: ForfilterP       value: 0.09428974694315859  
    user: rem_tene         value: 0.09353416800622748
```



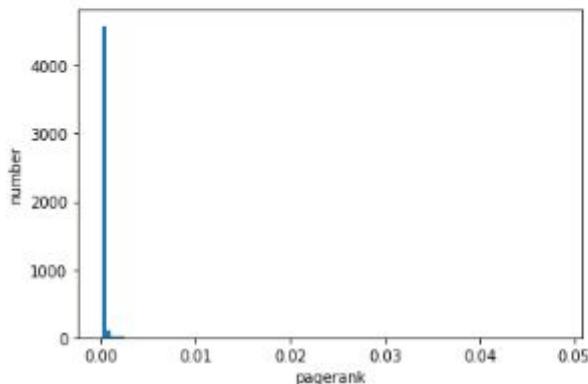
Rys. 51 Użytkownicy według wektora własnego w grafie Andrzeja Dudy

Na rys. 51 można zauważyć, że podobnie jak w przypadku stopnia i pośrednictwa, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość wektora własnego. Tutaj szczególnie wyróżnia się konto AndrzejDuda2020, co oznacza, że ma ono największy wpływ w sieci.

```

Max pagerank:
    user: dumnapolka12      value: 0.048426495947840646
    user: AndrzejDuda2020   value: 0.03454850425472104
    user: SondazeP          value: 0.02923475854985533
    user: MBalawelder       value: 0.013369375097970364
    user: Dorian88357211   value: 0.012414988517398022
    user: BitwaoPL          value: 0.009709563509903065
    user: jbrudzinski        value: 0.009515666763805357
    user: KamilaAndruszki  value: 0.009387211930825605
    user: Immanuela_Kant   value: 0.007859932375736274
    user: Bart_Wielinski    value: 0.007729813119355264

```



Rys. 52 Użytkownicy według miary Page Rank w grafie Andrzeja Dudy

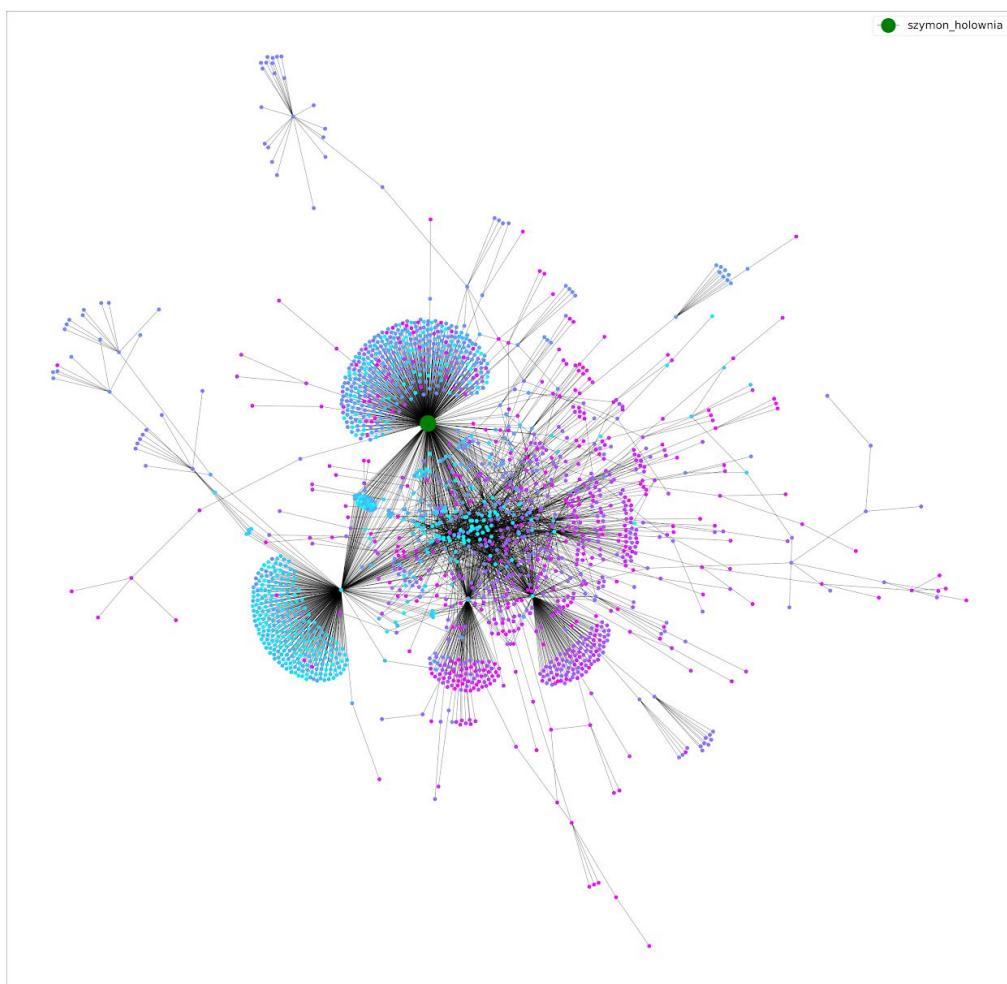
Na rys. 52 widać, że w przypadku miary PageRank u wszystkich użytkowników w sieci jest ona bardzo niska.

Analizując rys. 48-52 można zauważyc, że wśród użytkowników o największych miarach centralności w grafie społeczności Andrzeja Dudy znajdują się: AndrzejDuda2020 - konto sztabu, dumnapolka12 - użytkownik wspierający PIS i Andrzeja Dudę, SondazeP - konto tweetujące o polityce i wyborach, jbrudzinski - poseł PIS do Parlamentu Europejskiego, Bart_Wielinski - dziennikarz Gazety Wyborczej, MosinskiJan - poseł PIS. Pozostałe konta to zwykli użytkownicy Twittera, w większości wspierający obecnego prezydenta.

Wartości poszczególnych miar pokazanych na rysunkach 48-52 zostały znormalizowane. Można zauważyc, że jedynie w przypadku miary bliskości (rys. 49) wartości te są rozłożone dosyć równomiernie między wartościami 0 a 0.40 z największą liczbą użytkowników znajdujących się w przedziale 0.25-0.30, co jest związane z dosyć krótkim średnimi odległościami między wierzchołkami w grafie. W przypadku pozostałych miar (rys. 48, 50-52) prawie wszyscy użytkownicy znajdują się w przedziale 0-0.05, jedynie kilku ma miarę wyższą niż 0.1. Jednym z tych kont jest AndrzejDuda2020 - konto sztabowe kandydata. Jest to zrozumiałe, gdyż jest to konto powstałe dla promowania kandydatury Andrzeja Dudy, a więc udostępnia liczne tweety z hashtagami dotyczącymi kampanii i jest często retweetowane. Ciekawa natomiast jest obecność użytkownika dumnapolka12 na wysokich pozycjach, gdyż nie jest to konto z bardzo dużą liczbą obserwatorów.

6.2 Sieć Szymona Hołowni

Sieć społeczna stworzona przez użytkowników Twittera, którzy w swoich wypowiedziach używali hashtagów dotyczących Szymona Hołowni:



Rys. 53 Sieć ludzi tweetujących o Szymonie Hołowni

Podstawowe właściwości grafu:

```

Number of nodes: 1947
Number of edges: 3059
Max degree: 670
Min degree: 1
Density: 0.0016147328670191735
Graph connected: False
Number of connected components: 42
Largest connected component:
Number of nodes: 1843
Number of edges: 2996
Max degree: 670
Min degree: 1
Density: 0.0017650493135690228
Average clustering coefficient: 0.10114815184386433
Transitivity: 0.01756926447407091
Diameter: 11
Average distance between two nodes: 3.32

```

Rys. 54 Podstawowe własności grafu ludzi tweetujących o Szymonie Hołowni

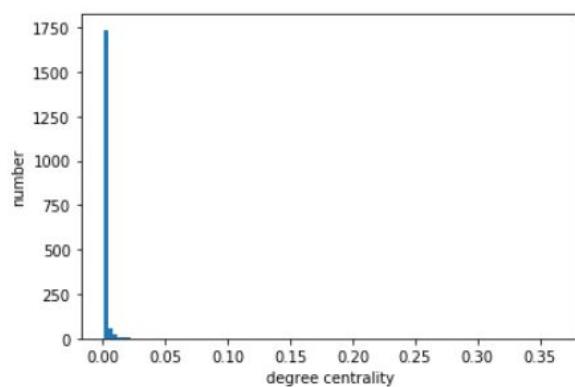
Jak widać na rys. 54 użytkownicy, których tweety pobraliśmy, nie stworzyli jednego spójnego grafu, jednakże jego największa wspólna składowa zawiera w sobie większość jego wierzchołków i krawędzi i to ona jest analizowana w dalszej części. Gęstość i współczynnik grupowania grafu są większe niż w przypadku Andrzeja Dudy, natomiast średnia odległość między wierzchołkami jest niższa. Graf jest mniejszy, a użytkownicy go tworzący bardziej połączeni.

Miary centralności w grafie:

```

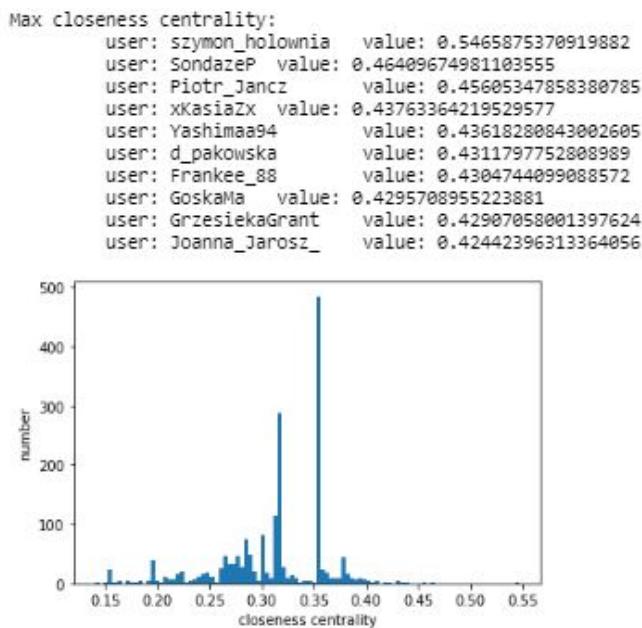
Max degree centrality:
user: szymon_holownia    value: 0.36373507057546145
user: SondazeP   value: 0.21064060803474485
user: Piotr_Jancz   value: 0.11889250814332249
user: Frankee_88   value: 0.09283387622149837
user: holownia_live   value: 0.053203040173724216
user: DarioPolo9   value: 0.0499457111834962
user: piotrasik   value: 0.04505971769815418
user: Mark84801721   value: 0.03745928338762215
user: xKasiaZx   value: 0.03420195439739414
user: michalkobosko   value: 0.03040173724212812

```



Rys. 55 Użytkownicy według stopnia w grafie Szymona Hołowni

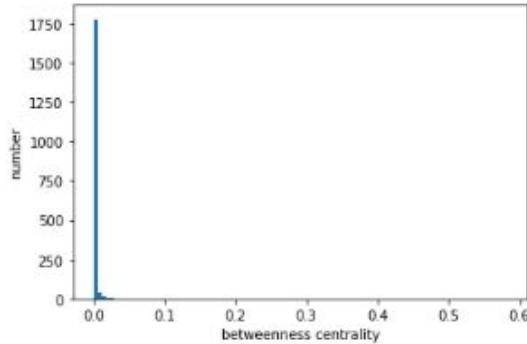
Na rys. 55 można zauważyć, że prawie wszyscy użytkownicy mają bardzo niski i zbliżony do siebie stopień w sieci. Wyróżnia się jedynie oficjalne konto kandydata - szymon_holownia, co jest zrozumiałe. Jest to najaktywniejsze konto w sieci.



Rys. 56 Użytkownicy według bliskości w grafie Szymona Hołowni

Na rys. 56 można zauważyć, że miara bliskości użytkowników w sieci jest już bardziej równomiernie rozłożona niż stopień i żaden z nich pod tym względem się nie wyróżnia. Wynika z tego, że większość użytkowników ma dosyć krótkie średnie odległości od pozostałych, wynika to zapewne z faktu, że nie jest to duża sieć.

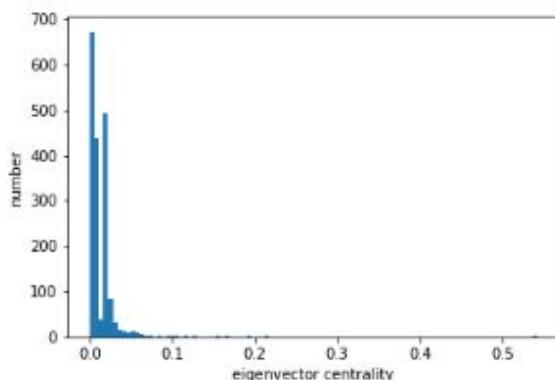
```
Max betweenness centrality:
user: szymon_holownia  value: 0.581814172922196
user: SondazeP  value: 0.3342178168449518
user: Piotr_Jancz  value: 0.16436249649619972
user: Frankee_88  value: 0.10759734926175264
user: DarioPolo9  value: 0.06138222635291534
user: piotrasik  value: 0.03403961816820038
user: holownia_live  value: 0.028702328448464674
user: AndrzejOlkiewi  value: 0.026839745453584772
user: Mark84801721  value: 0.025500746761195275
user: xKasiaZx  value: 0.024855068434744696
```



Rys. 57 Użytkownicy według pośrednictwa w grafie Szymona Hołowni

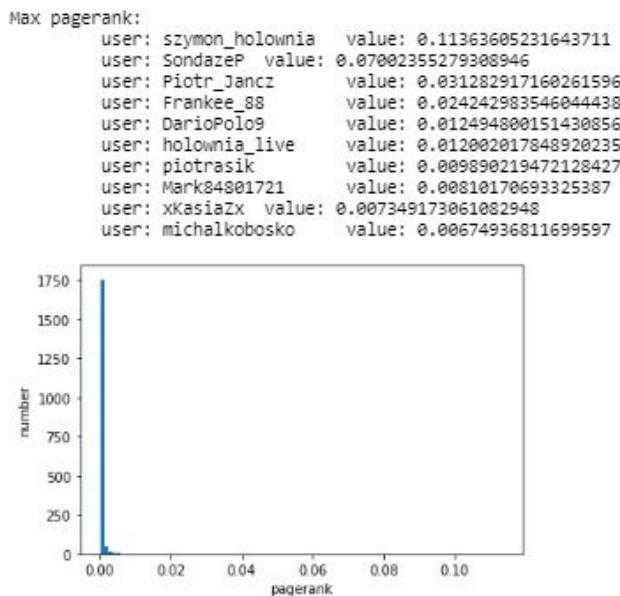
Na rys. 57 można zauważyć, że podobnie jak w przypadku stopnia, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość pośrednictwa w sieci. Wyróżnia się jedynie konto `szymon_holownia`, które ma zdecydowanie największą kontrolę w sieci.

```
Max eigenvector centrality:
user: szymon_holownia  value: 0.5420428309002481
user: SondazeP  value: 0.21567814291086718
user: Piotr_Jancz  value: 0.19202733030177083
user: Frankee_88  value: 0.1630931726880383
user: holownia_live  value: 0.1540763992592527
user: xKasiaZx  value: 0.12656399984533948
user: piotrasik  value: 0.12594898590807985
user: Mark84801721  value: 0.11893522550392259
user: Yashimaa94  value: 0.11670077715278714
user: michalkobosko  value: 0.10651060109193444
```



Rys. 58 Użytkownicy według wektora własnego w grafie Szymona Hołowni

Na rys. 58 można zauważać, że podobnie jak w przypadku stopnia i pośrednictwa, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość wektora własnego. Tutaj również wyróżnia się konto szymon_holownia, co oznacza, że ma ono największy wpływ w sieci.



Rys. 59 Użytkownicy według miary Page Rank w grafie Szymona Hołowni

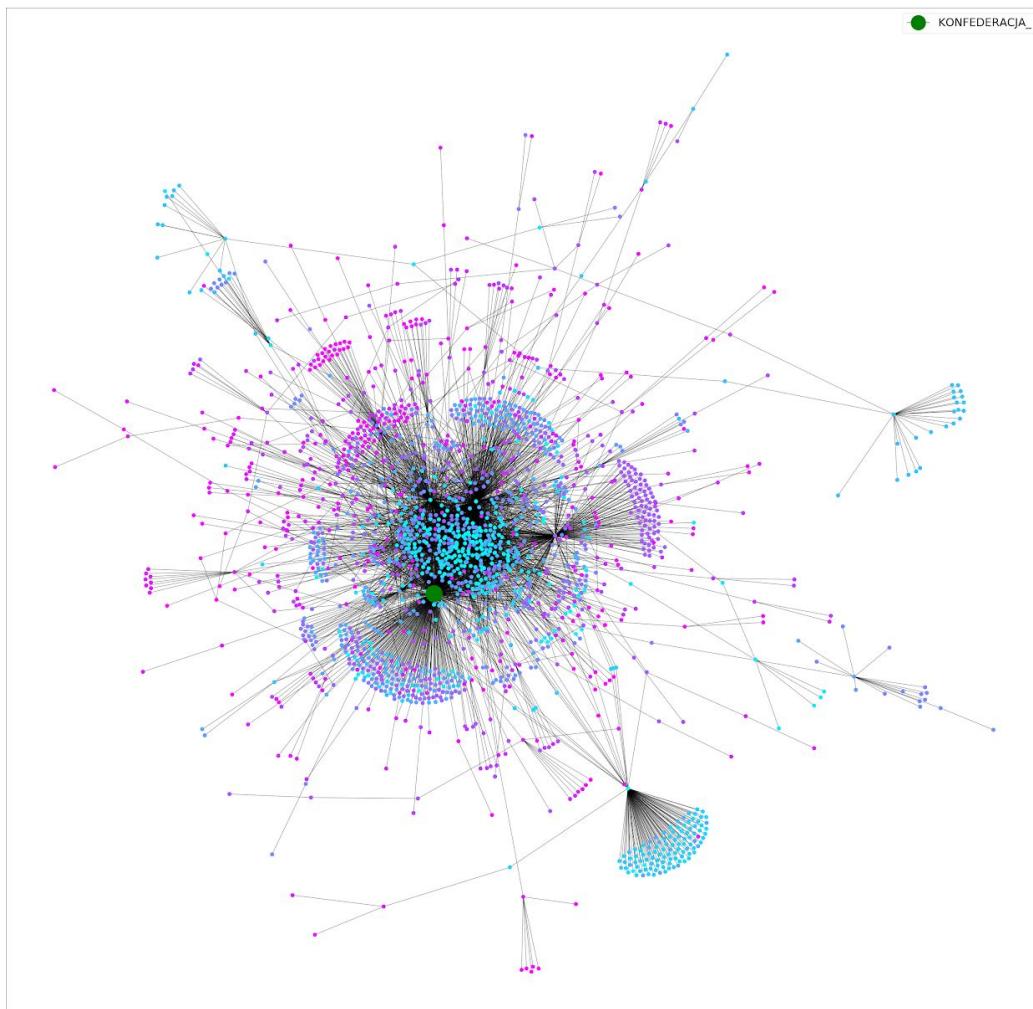
Na rys. 59 widać, że w przypadku miary PageRank u wszystkich użytkowników w sieci jest ona bardzo niska, najwyższa wartość wynosi jedynie ok. 0.1.

Na rys. 55-59 można zauważać, że wśród użytkowników o największych miarach centralności w grafie społeczności Szymona Hołowni znajdują się: szymon_holownia - konto kandydata, SondazeP - konto tweetujące o polityce i wyborach, michal_kobosko - pełnomocnik komitetu kandydata. Pozostałe konta to zwykli użytkownicy Twittera wspierający kandydata Szymona Hołowni na prezydenta.

Wartości poszczególnych miar pokazanych na rysunkach 55-59 zostały znormalizowane. Podobnie jak w grafie Andrzeja Dudy w przypadku miar przedstawionych na rys. 55 i 57-59 prawie wszyscy użytkownicy znajdują się w przedziale 0-0.05, jedynie kilku ma miarę wyższą niż 0.1. Jednym z tych kont jest szymon_holownia - oficjalne konto kandydata, co jest zrozumiałe, gdyż udostępnia ono liczne tweety z hashtagami kandydata i jest retweetowane częściej niż inni - w większości zwykłi użytkownicy Twittera. Natomiast w przypadku miary bliskości (rys. 56) wartości te są rozłożone znacznie bardziej równomiernie między wartościami 0 a 0.50 z największą liczbą użytkowników z wartością około 0.35. Jest związane z krótkimi średnimi odległościami między wierzchołkami w grafie.

6.3 Sieć Krzysztofa Bosaka

Sieć społeczna stworzona przez użytkowników Twittera, którzy w swoich wypowiedziach używali hashtagów dotyczących Krzysztofa Bosaka:



Rys. 60 Sieć ludzi tweetujących o Krzysztofie Bosaku

Podstawowe własności grafu:

```

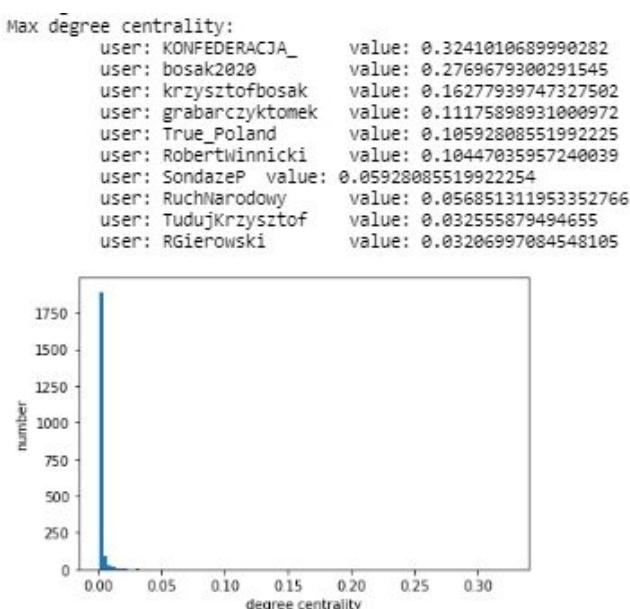
Number of nodes: 2208
Number of edges: 4450
Max degree: 667
Min degree: 1
Density: 0.0018263693255320685
Graph connected: False
Number of connected components: 71
Largest connected component:
Number of nodes: 2059
Number of edges: 4371
Max degree: 667
Min degree: 1
Density: 0.002063046824224729
Average clustering coefficient: 0.19263412737871238
Transitivity: 0.02217916009949596
Diameter: 12
Average distance between two nodes: 3.44

```

Rys. 61 Podstawowe własności grafu ludzi tweetujących o Krzysztofie Bosaku

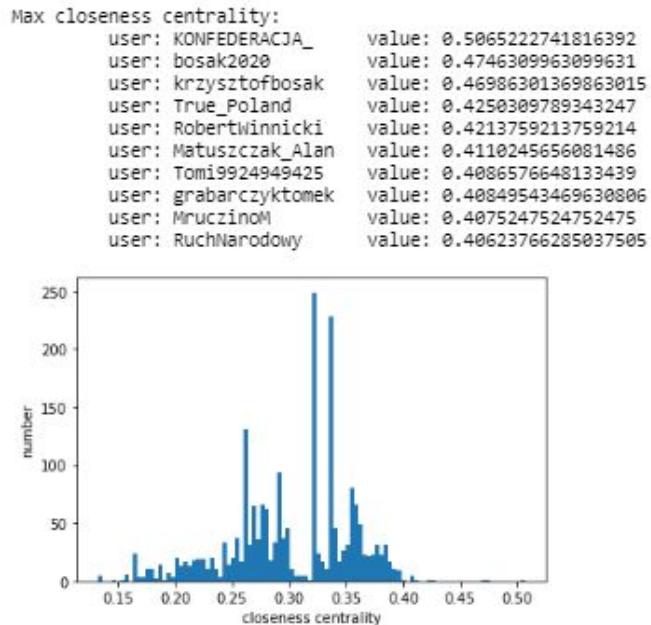
Na rys. 61 można zauważyc, że użytkownicy, których tweety pobraliśmy, nie stworzyli jednego spójnego grafu, jednakże jego największa wspólna składowa zawiera w sobie większość jego wierzchołków i krawędzi. Gęstość i współczynnik grupowania grafu są większe niż w przypadku Andrzeja Dudy i Szymona Hołowni. Średnia odległość między wierzchołkami jest niższa niż u Dudy, ale wyższa niż u Hołowni. Graf jest trochę większy niż Hołowni, ale użytkownicy go tworzący bardziej połączeni między sobą. W dalszej części analizowana jest największa wspólna składowa grafu.

Miary centralności w grafie:



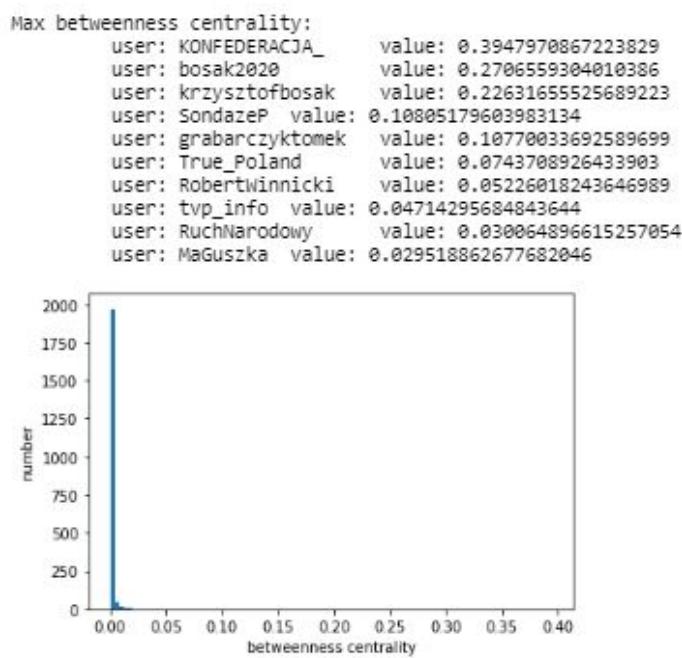
Rys. 62 Użytkownicy według stopnia w grafie Krzysztofa Bosaka

Na rys. 62 można zauważyc, że prawie wszyscy użytkownicy mają bardzo niski i zbliżony do siebie stopień w sieci. Najbardziej wyróżniają się konto Konfederacji (KONFEDERACJA_) oraz konto sztabu kandydata (bosak2020), co jest zrozumiałe. Są to najaktywniejsze konta w sieci.



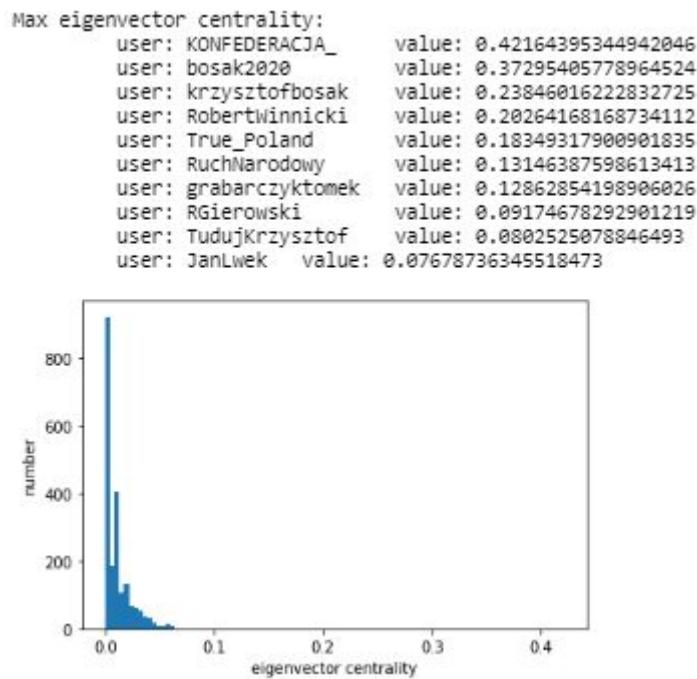
Rys. 63 Użytkownicy według bliskości w grafie Krzysztofa Bosaka

Na rys. 63 można zauważyc, że miara bliskości użytkowników w sieci jest już bardziej równomiernie rozłożona niż stopień i żaden z nich pod tym względem się szczególnie nie wyróżnia. Wynika z tego, że większość użytkowników ma dosyć krótkie średnie odległości od pozostałych, powodem jest zapewne fakt, że nie jest to duża sieć.



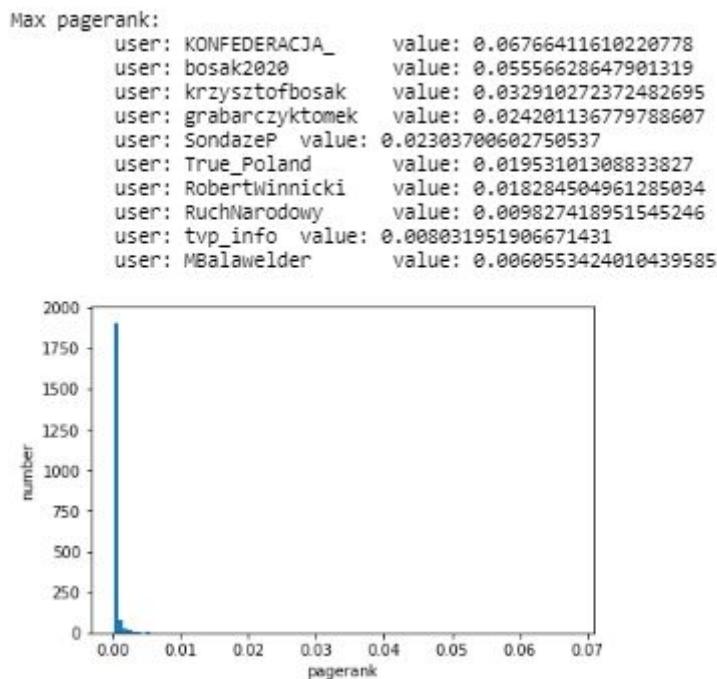
Rys. 64 Użytkownicy według pośrednictwa w grafie Krzysztofa Bosaka

Na rys. 64 można zauważyć, że podobnie jak w przypadku stopnia, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość pośrednictwa w sieci. Ponownie najbardziej wyróżniają się konta KONFEDERACJA_ i bosak2020, które mają największą kontrolę w sieci.



Rys. 65 Użytkownicy według wektora własnego w grafie Krzysztofa Bosaka

Na rys. 65 można zauważać, że podobnie jak w przypadku stopnia i pośrednictwa, prawie wszyscy użytkownicy mają bardzo niską wartość wektora własnego. Zakres tych wartości jest jednak trochę szerszy. Tutaj również wyróżniają się konta KONFEDERACJA_ i bosak2020, co oznacza, że mają one największy wpływ w sieci.



Rys. 66 Użytkownicy według miary Page Rank w grafie Krzysztofa Bosaka

Na rys. 66 widać, że w przypadku miary PageRank u wszystkich użytkowników w sieci jest ona bardzo niska, najwyższa wartość wynosi jedynie ok. 0.06.

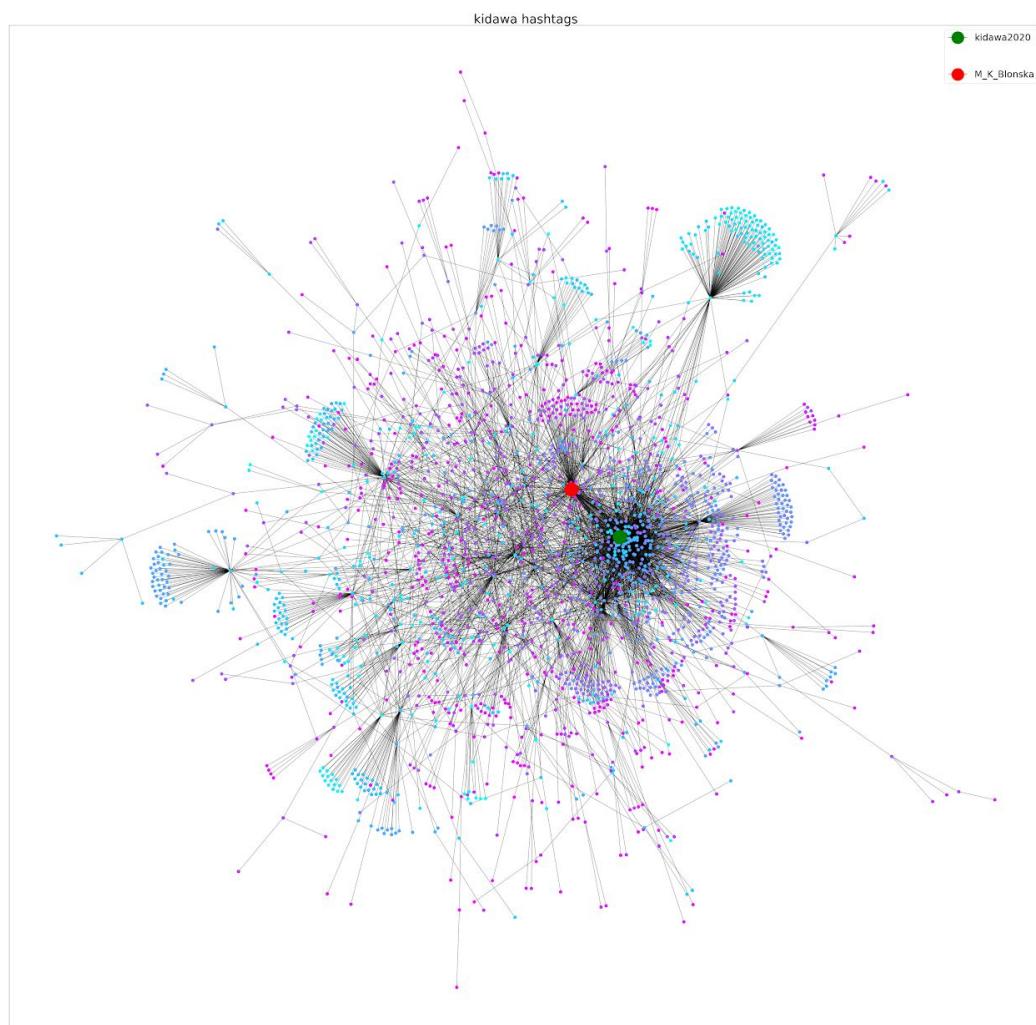
Analizując rys. 62-66 można zauważać, że wśród użytkowników o największych miarach centralności w grafie społeczności Krzysztofa Bosaka znajdują się: KONFEDERACJA_ - konto partii kandydata, bosak2020 - konto sztabu kandydata, krzysztofbosak - konto kandydata, grabarczyktomek - rzecznik prasowy Konfederacji, RobertWinnicki - prezes Ruchu Narodowego, RuchNarodowy - konto partii Ruch Narodowy, tvp_info - konto TVP INFO, RGierowski - członek Ruchu Narodowego, SondazeP - konto tweetujące o polityce i wyborach. Pozostałe konta to zwykli użytkownicy Twittera wspierający kandydatę Krzysztofa Bosaka na prezydenta. W tej społeczności nad potencjalnymi wyborcami przeważają konta polityków.

Wartości poszczególnych miar pokazanych na rysunkach 62-66 zostały znormalizowane. Podobnie jak w grafie Andrzeja Dudy i Szymona Hołowni w przypadku miar przedstawionych na rys. 62 i 64-66 prawie wszyscy użytkownicy mają wartości miar w przedziale 0-0.05, jedynie kilku ma miarę wyższą niż 0.1, a w

przypadku PageRank najwyższa wartość wynosi ok. 0.06 (rys. 66). Użytkownikami wyróżniającymi się spośród innych ze względu na wartości tych miar (stopień, pośrednictwo, wektor własny, pagerank) jest konto Konfederacji oraz osobiste i sztabowe konto Krzysztofa Bosaka. Są one bardzo aktywne, publikują najczęściej tweetów z hashtagami kandydata. Natomiast w przypadku miary bliskości (rys. 63) wartości te są rozłożone znacznie bardziej równomiernie między wartościami 0 a 0.50 z największą liczbą użytkowników z wartościami w przedziale 0.30-0.35. Jest związane z krótkimi średnimi odległościami między wierzchołkami w grafie.

6.4 Sieć Małgorzaty Kidawy-Błońskiej

Sieć społeczna stworzona przez użytkowników Twittera, którzy w swoich wypowiedziach używali hashtagów dotyczących Małgorzaty Kidawy-Błońskiej:



Rys. 67 Sieć ludzi tweetujących o Małgorzacie Kidawie-Błońskiej

Podstawowe właściwości grafu:

```

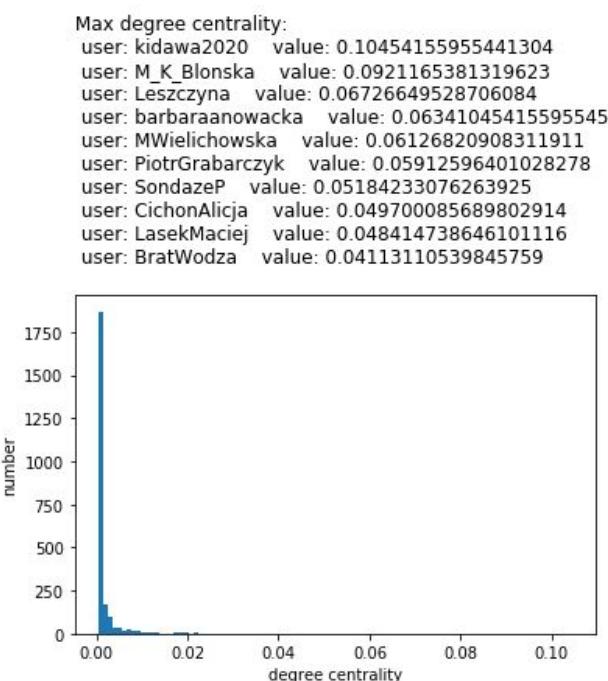
Number of nodes: 2585
Number of edges: 4967
Max degree: 244
Min degree: 1
Density: 0.0014872058973238078
Graph connected: False
Number of connected components: 111
Largest connected component:
Number of nodes: 2335
Number of edges: 4827
Max degree: 244
Min degree: 1
Density: 0.0017714119000566983
Average clustering coefficient: 0.07296198972008547
Transitivity: 0.07634848246579651
Diameter: 10
Average distance between two nodes: 4.23

```

Rys. 68 Podstawowe własności grafu ludzi tweetujących o Małgorzacie Kidawie-Błońskiej

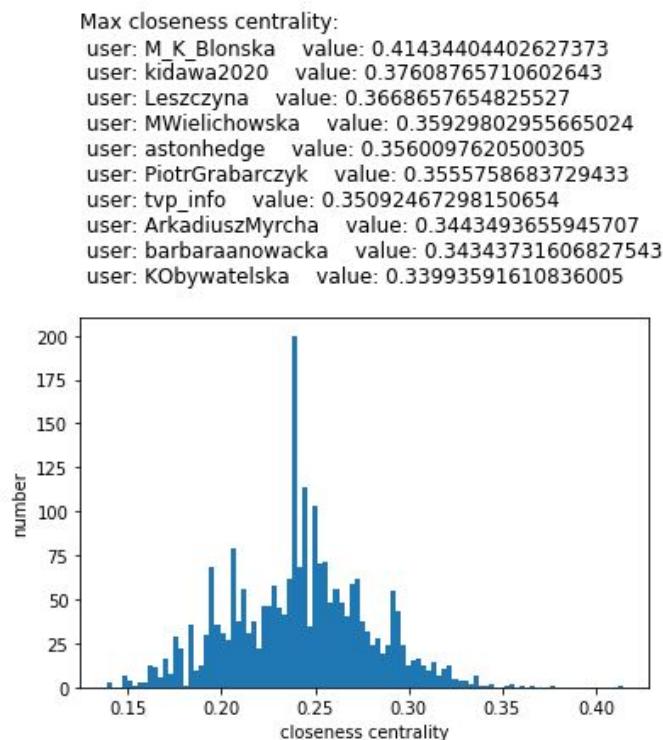
Na rys. 68 można zauważyc, że użytkownicy, których tweety pobraliśmy, nie stworzyli jednego spójnego grafu, jednakże jego największa wspólna składowa zawiera w sobie większość jego wierzchołków i krawędzi. Gęstość grafu jest większa niż w przypadku Andrzeja Dudy, ale mniejsza niż u Krzysztofa Bosaka. Średnia odległość między wierzchołkami jest wyższa niż u pozostałych kandydatów. Graf jest trochę większy niż Bosaka, ale użytkownicy go tworzący nie są tak dobrze połączeni. W dalszej części analizowana jest największa wspólna składowa grafu.

Miary centralności w grafie:



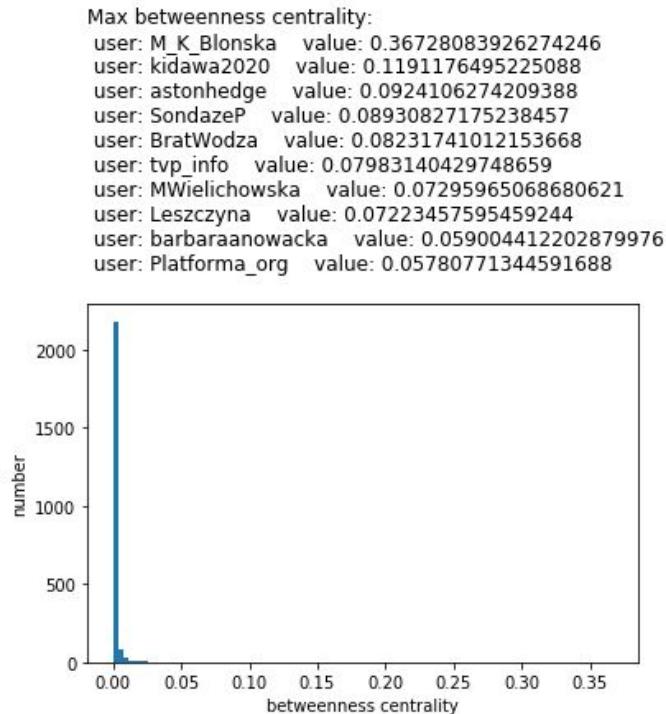
Rys. 69 Użytkownicy według stopnia w grafie Małgorzaty Kidawy-Błońskiej

Na rys. 69 można zauważyc, że prawie wszyscy użytkownicy mają bardzo niski i zbliżony do siebie stopień w sieci. Najwyższą wartość ma konto kampanii kidawa2020, jednak nie jest to wartość szczególnie wyróżniająca się spośród pozostałych.



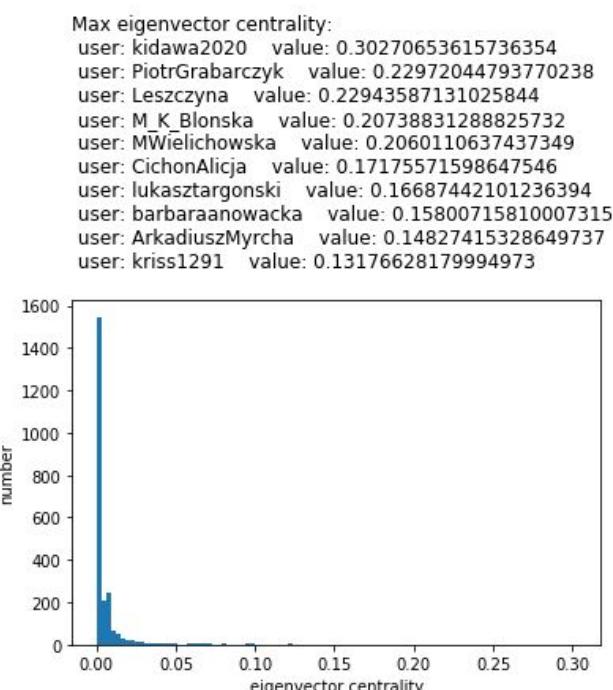
Rys. 70 Użytkownicy według bliskości w grafie Małgorzaty Kidawy-Błońskiej

Na rys. 70 można zauważyc, że miara bliskości użytkowników w sieci jest już bardziej równomiernie rozłożona niż stopień i żaden z nich pod tym względem się nie wyróżnia. Wynika z tego, że większość użytkowników ma dosyć krótkie średnie odległości od pozostałych, powodem jest zapewne fakt, że większość użytkowników znajduje się w centrum sieci. Dodatkowo, nie jest to duża sieć. Najwyższą wartość ma oficjalne konto kandydatki - M_K_Blonska, co oznacza, że jest ono najbardziej efektywne w sieci.



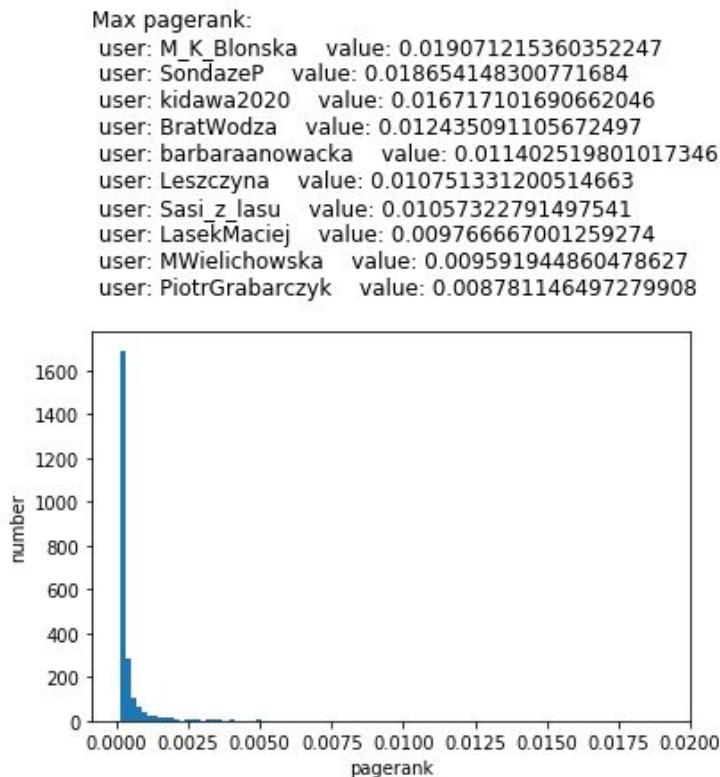
Rys. 71 Użytkownicy według pośrednictwa w grafie Małgorzaty Kidawy-Błońskiej

Na rys. 71 można zauważyć, że podobnie jak w przypadku stopnia, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość pośrednictwa w sieci. Wyróżnia się jedynie konto M_K_Blonska, które ma zdecydowanie największą kontrolę w sieci.



Rys. 72 Użytkownicy według wektora własnego w grafie Małgorzaty Kidawy-Błońskiej

Na rys. 72 można zauważać, że podobnie jak w przypadku stopnia i pośrednictwa, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość wektora własnego. Tutaj wyróżnia się konto kidawa2020, co oznacza, że ma ono największy wpływ w sieci.



Rys. 73 Użytkownicy według miary Page Rank w grafie Małgorzaty Kidawy-Błońskiej

Na rys. 73 widać, że w przypadku miary PageRank u wszystkich użytkowników w sieci jest ona bardzo niska, najwyższa wartość wynosi jedynie ok. 0.02.

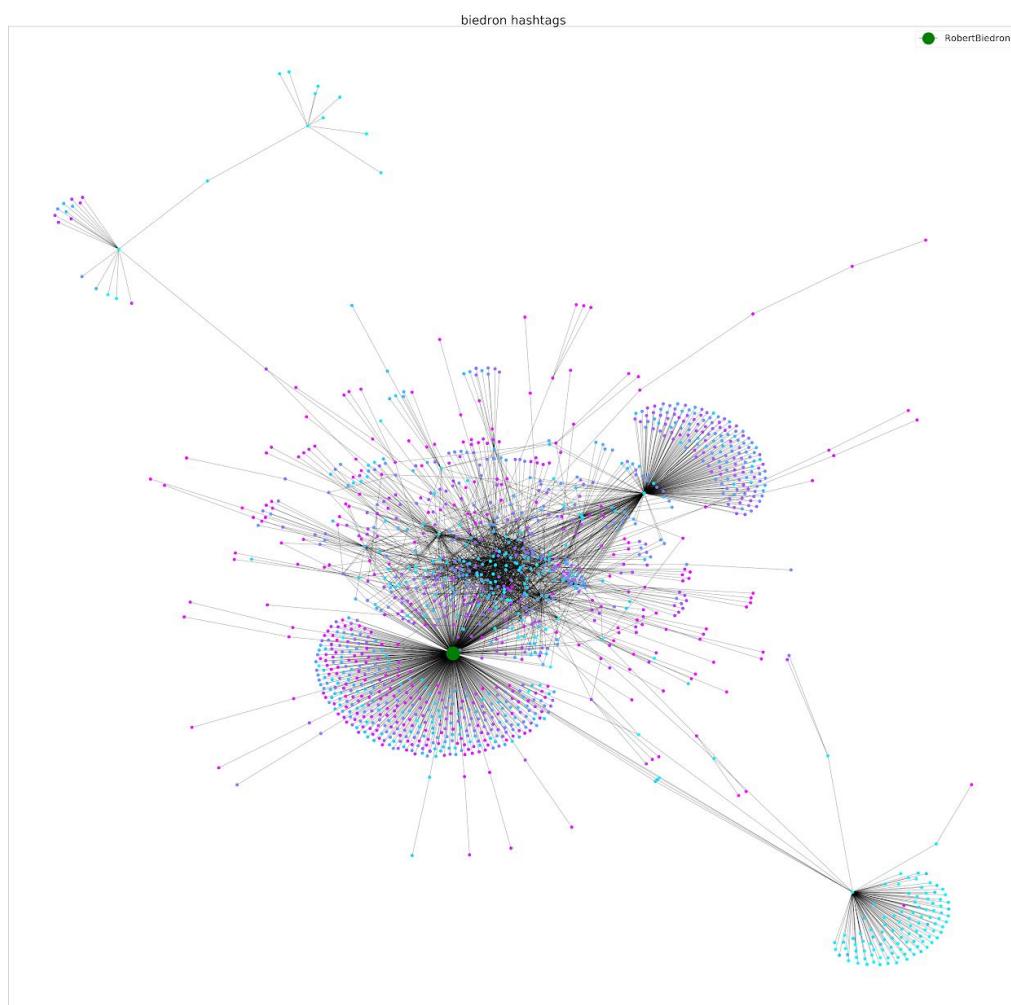
Analizując rys. 69-73 można zauważać, że wśród użytkowników o największych miarach centralności w grafie społeczności Krzysztofa Bosaka znajdują się: kidawa2020 - konto sztabu kandydatki, M_K_Blonska - konto kandydatki, Leszczyna - posłanka KO, barbaraanowacka - posłanka KO, MWielichowska - posłanka KO, tvp_info - konto TVP INFO, PiotrGrabarczyk - przewodniczący PO w powiecie olkuskim, LasekMaciej - poseł KO, ArkadiuszMyrcha - poseł KO, Platforma_org - konto Platformy Obywatelskiej, SondazeP - konto tweetujące o polityce i wyborach. Pozostałe konta to zwykli użytkownicy Twittera w większości wspierający kandydaturę Małgorzaty Kidawy-Błońskiej na prezydenta.

Wartości poszczególnych miar pokazanych na rysunkach 69-73 zostały znormalizowane. Podobnie jak w grafach poprzednich kandydatów w przypadku miar przedstawionych na rys. 69 i 71-73 (stopień, pośrednictwo, wektor własny,

pagerank) prawie wszyscy użytkownicy mają wartości miar w przedziale 0-0.05, jedynie kilku ma miarę wyższą niż 0.1, a w przypadku PageRank najwyższa wartość wynosi ok. 0.01 (rys. 73). Użytkownikami wyróżniającymi się spośród innych ze względu na wartości tych miar jest osobiste i sztabowe konto Małgorzaty Kidawy-Błońskiej. Natomiast w przypadku miary bliskości (rys. 70) wartości te są rozłożone znacznie bardziej równomiernie między wartościami 0 a 0.40 z największą liczbą użytkowników z wartościami wynoszącymi około 0.25. Jest to niższa wartość niż dla pozostałych kandydatów, gdyż średnie odległości między wierzchołkami są większe.

6.5 Robert Biedroń

Sieć społeczna stworzona przez użytkowników Twittera, którzy w swoich wypowiedziach używali hashtagów dotyczących Roberta Biedronia:



Rys. 74 Sieć ludzi tweetujących o Robertie Biedroniu

Podstawowe właściwości grafu:

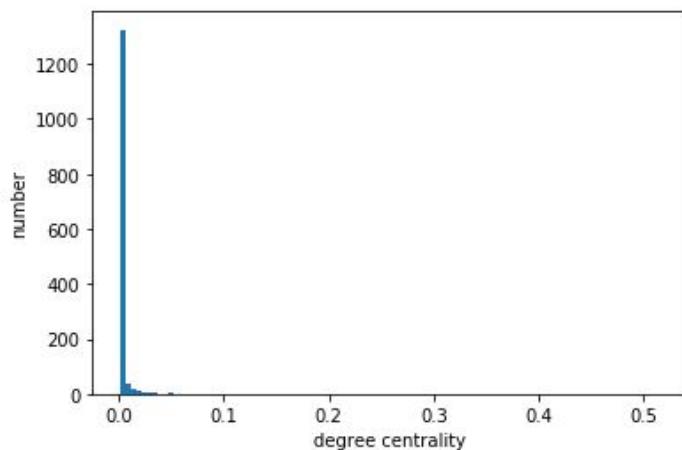
```
Number of nodes: 1568
Number of edges: 2736
Max degree: 728
Min degree: 1
Density: 0.0022270554680072414
Graph connected: False
Number of connected components: 64
Largest connected component:
Number of nodes: 1415
Number of edges: 2646
Max degree: 728
Min degree: 1
Density: 0.002644928803834447
Average clustering coefficient: 0.1579917015944656
Transitivity: 0.016738491024601166
Diameter: 11
Average distance between two nodes: 3.04
```

Rys. 75 Podstawowe własności grafu ludzi tweetujących o Robercie Biedroniu

Na rys. 75 można zauważyc, że użytkownicy, których tweety pobraliśmy, nie stworzyli jednego spójnego grafu, jednakże jego największa wspólna składowa zawiera w sobie większość jego wierzchołków i krawędzi. Rozmiar grafu jest mniejszy, a gęstość większa niż u reszty kandydatów, oprócz Władysława Kosiniaka-Kamysza. Średnia odległość między wierzchołkami jest najniższa spośród wszystkich kandydatów. Współczynnik grupowania jest wysoki w porównaniu do innych kandydatów, niższy tylko niż u Krzysztofa Bosaka. W dalszej części analizowana jest największa wspólna składowa grafu.

Miary centralności w grafie:

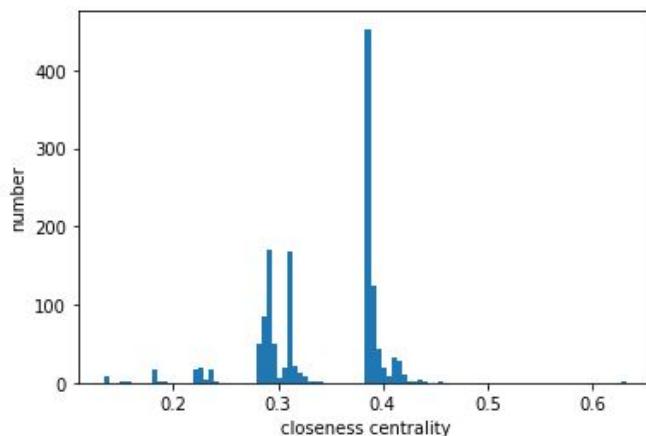
Max degree centrality:
 user: RobertBiedron value: 0.5148514851485149
 user: socjalisciLGBT value: 0.19943422913719944
 user: SondazeP value: 0.08203677510608204
 user: przedwiosnie_ value: 0.060820367751060825
 user: poselTTrela value: 0.051626591230551626
 user: wsawoniewicz value: 0.04809052333804809
 user: HannaGillPiatek value: 0.04526166902404526
 user: janczewski_p value: 0.04101838755304102
 user: little_heartu value: 0.03323903818953324
 user: JoankaSW value: 0.03253182461103253



Rys. 76 Użytkownicy według stopnia w grafie Roberta Biedronia

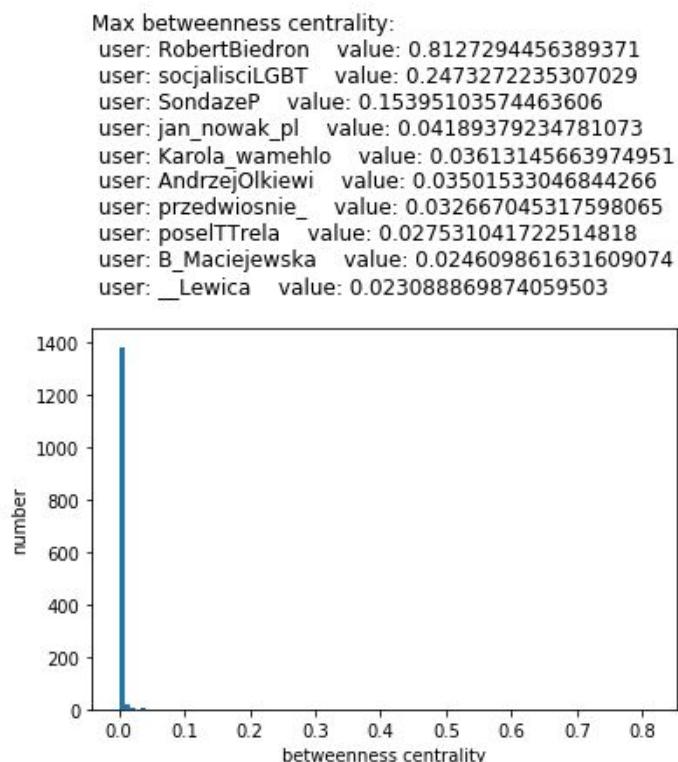
Na rys. 76 można zauważyć, że prawie wszyscy użytkownicy mają bardzo niski i zbliżony do siebie stopień w sieci. Wyróżnia się jedynie oficjalne konto kandydata - RobertBiedron, co jest zrozumiałe. Oznacza to, że jest to najaktywniejsze konto w sieci.

Max closeness centrality:
 user: RobertBiedron value: 0.6320965578900313
 user: socjalisciLGBT value: 0.45539452495974236
 user: HannaGillPiatek value: 0.43967661691542287
 user: CzerwonyKat value: 0.43655449212719977
 user: luuvmybaby value: 0.43561306223043744
 user: simon_wicz value: 0.43480934809348093
 user: juliat4r4 value: 0.4344086021505376
 user: pinksley value: 0.42939568782265414
 user: LewicowyHub value: 0.42848484848484847
 user: m0bi13 value: 0.4248798076923077



Rys. 77 Użytkownicy według bliskości w grafie Roberta Biedronia

Na rys. 77 można zauważyc, że miara bliskości użytkowników w sieci jest już bardziej rozłożona niż stopień. Można zauważyc dwie grupy użytkowników - z wartościami około 0.3 i 0.4. Wynika z tego, że prawie wszyscy użytkownicy mają dosyć krótkie średnie odległości od pozostałych, co jest zapewne spowodowane faktem, że sieć nie jest zbyt duża. Kontem o największej efektywności jest RobertBiedron.



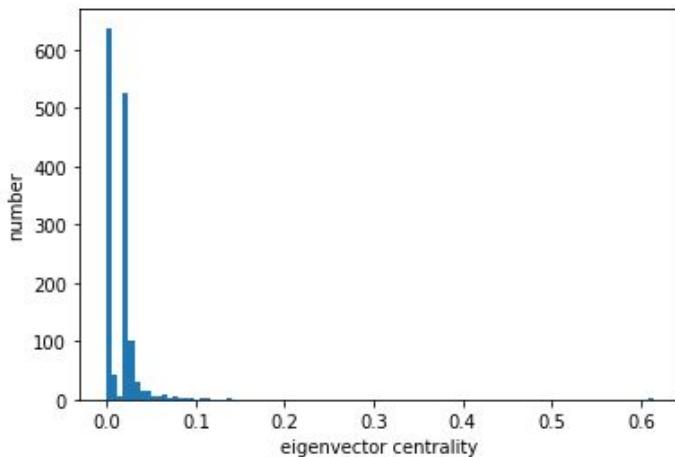
Rys. 78 Użytkownicy według pośrednictwa w grafie Roberta Biedronia

Na rys. 78 można zauważyć, że podobnie jak w przypadku stopnia, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość pośrednictwa w sieci. Ponownie zdecydowanie wyróżnia się konto RobertBiedron, przez które przechodzi większość połączeń w sieci.

```

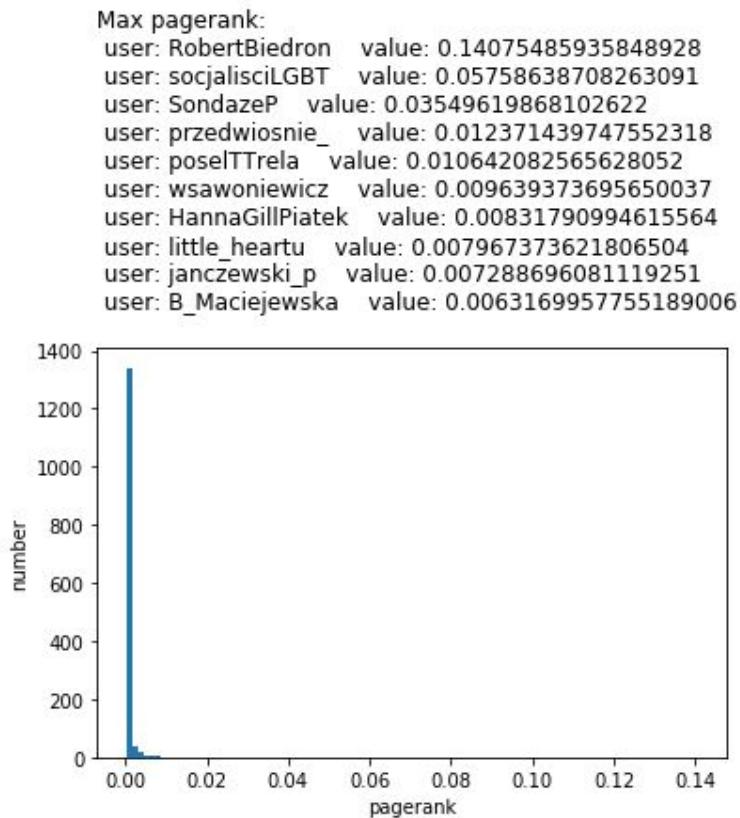
Max eigenvector centrality:
user: RobertBiedron  value: 0.6136006441341766
user: socjalisciLGBT  value: 0.13883115972086363
user: HannaGillPiatek  value: 0.11421875096765607
user: janczewski_p  value: 0.11028309989607847
user: przedwiosnie_  value: 0.1055207235028774
user: poselTTrela  value: 0.09470055558540767
user: simon_wicz  value: 0.08949866799492634
user: wsawoniewicz  value: 0.08868140807053562
user: pawlowska_pl  value: 0.08643989997483653
user: KrutulPawel  value: 0.08581684568862996

```



Rys. 79 Użytkownicy według wektora własnego w grafie Roberta Biedronia

Na rys. 79 można zauważyć, że podobnie jak w przypadku stopnia i pośrednictwa, prawie wszyscy użytkownicy mają bardzo niską wartość wektora własnego. Jednakże zakres tych wartości jest trochę szerszy. Tutaj również wyróżnia się konto RobertBiedron, co oznacza, że ma ono największy wpływ w sieci.



Rys. 80 Użytkownicy według miary Page Rank w grafie Roberta Biedronia

Na rys. 80 widać, że w przypadku miary PageRank u wszystkich użytkowników w sieci jest ona bardzo niska, najwyższa wartość wynosi jedynie ok. 0.14.

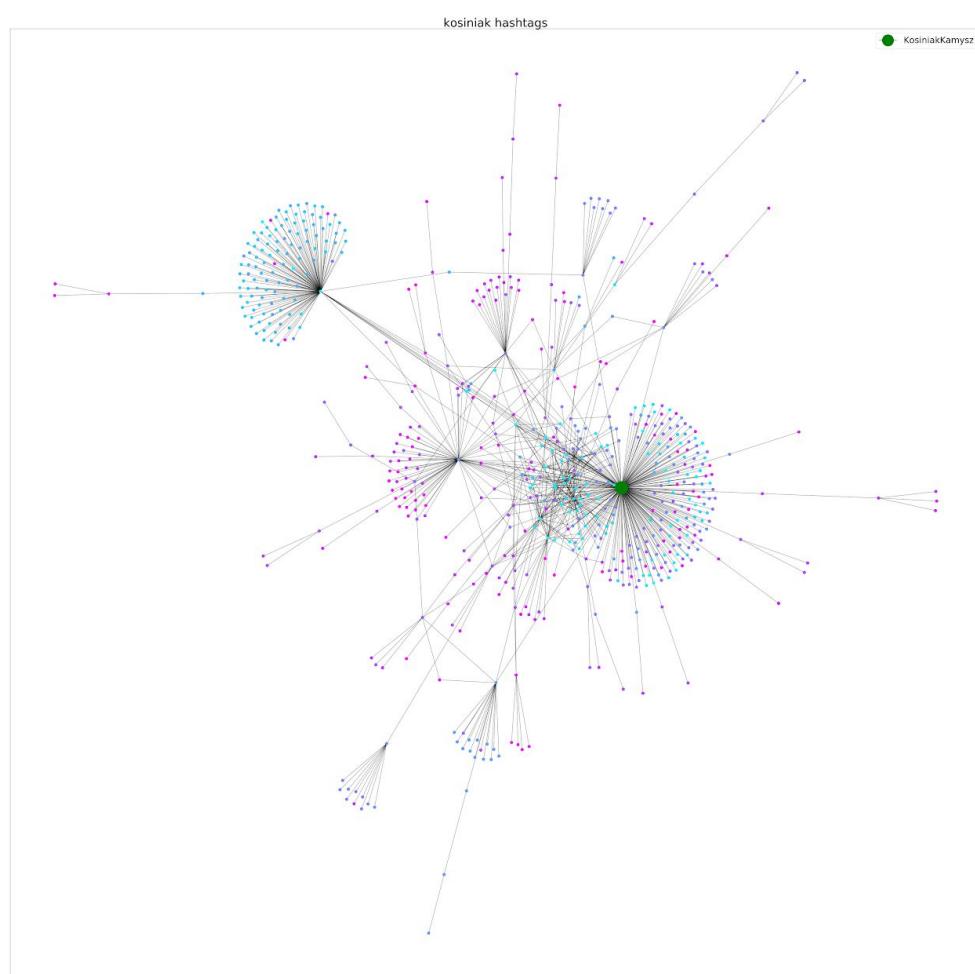
Analizując rys. 76-80 można zauważyc, że wśród użytkowników o największych miarach centralności w grafie społeczności Roberta Biedronia znajdują się: RobertBiedron - oficjalne konto kandydata, przedwiosnie_ - konto młodzieży partii Wiosna, posełTTrela- poseł Lewicy, HannaGilPiatek - posłanka Lewicy, pawlowska_pl - posłanka Lewicy, wsawoniewicz - działacz Lewicy, B_Maciejewska - posłanka KO, _Lewica - konto Lewicy, SondazeP - konto tweetujące o polityce i wyborach. Pozostałe konta to zwykli użytkownicy Twittera wspierający kandyturę Roberta Biedronia na prezydenta.

Wartości poszczególnych miar pokazanych na rysunkach 76-80 zostały znormalizowane. Podobnie jak w grafach poprzednich kandydatów w przypadku miar przedstawionych na rys. 76 i 78-80 (stopień, pośrednictwo, wektor własny, pagerank) prawie wszyscy użytkownicy mają wartości miar w przedziale 0-0.05, jedynie kilku ma miarę wyższą niż 0.1. Użytkownikiem wyróżniającym się spośród innych ze względu na wartości tych miar jest oficjalne konto Roberta Biedronia. Jest to zrozumiałe, gdyż kandydat jest aktywny w internecie i publikuje dużo tweetów. Drugi użytkownik w rankingach - konto socjalisciLGBT, wspierające Lewicę i jej

kandydata na prezydenta - jest znacznie mniej aktywne i wpływowe w powstałej sieci niż Biedroń. W przypadku miary bliskości (rys. 77) najwięcej użytkowników ma wartości około 0.30 i 0.40, bardzo niewiele znajduje się w przedziale 0-0.20. Są to wyższe wartości niż dla pozostałych kandydatów, gdyż średnie odległości między wierzchołkami są niższe.

6.6 Władysław Kosiniak-Kamysz

Sieć społeczna stworzona przez użytkowników Twittera, którzy w swoich wypowiedziach używali hashtagów dotyczących Władysława Kosiniaka Kamysza:



Rys. 81 Sieć ludzi tweetujących o Władysławie Kosiniaku-Kamyszu

Podstawowe własności grafu:

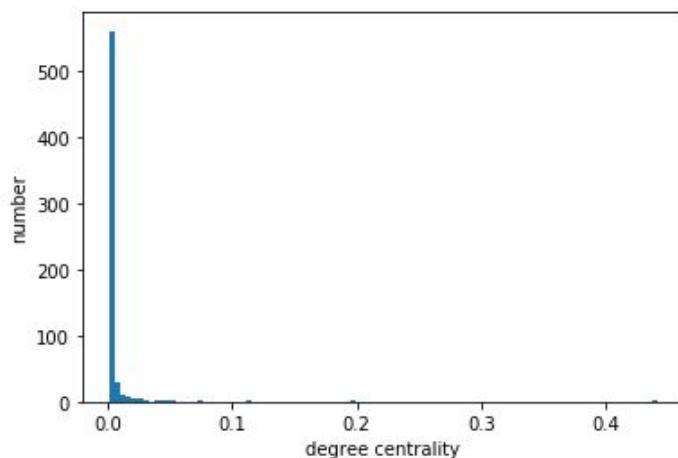
```
Number of nodes: 703
Number of edges: 944
Max degree: 276
Min degree: 1
Density: 0.003825688036214352
Graph connected: False
Number of connected components: 31
Largest connected component:
Number of nodes: 627
Number of edges: 898
Max degree: 276
Min degree: 1
Density: 0.004575772862303886
Average clustering coefficient: 0.10622351455239111
Transitivity: 0.01643431939249688
Diameter: 9
Average distance between two nodes: 3.20
```

Rys. 82 Podstawowe własności grafu ludzi tweetujących o Władysławie Kosiniaku-Kamyszu

Na rys. 82 można zauważyc, że użytkownicy, których tweety pobraliśmy, nie stworzyli jednego spójnego grafu, jednakże jego największa wspólna składowa zawiera w sobie większość jego wierzchołków i krawędzi. Graf społeczności Władysława Kosiniaka-Kamysza jest najmniejszy spośród wszystkich kandydatów. Wynika z tego, że zainteresowanie jego kandyaturą na Twitterze jest najmniejsze. Gęstość grafu jest bardzo wysoka w porównaniu do innych kandydatów, a niższą średnią odległość między wierzchołkami ma tylko Robert Biedroń. W dalszej części analizowana jest największa wspólna składowa grafu.

Miary centralności w grafie:

Max degree centrality:
 user: KosiniakKamysz value: 0.4408945686900958
 user: SondazeP value: 0.19808306709265175
 user: JanParadowski_ value: 0.1134185303514377
 user: magdasobkowiak value: 0.07507987220447285
 user: OjczynaMoja value: 0.051118210862619806
 user: Aneta_Marciszek value: 0.04792332268370607
 user: PSL_Malopolska value: 0.04313099041533546
 user: pkgrabowski value: 0.036741214057507986
 user: TomaszPanfic86 value: 0.03035143769968051
 user: MichalDziubak value: 0.027156549520766772



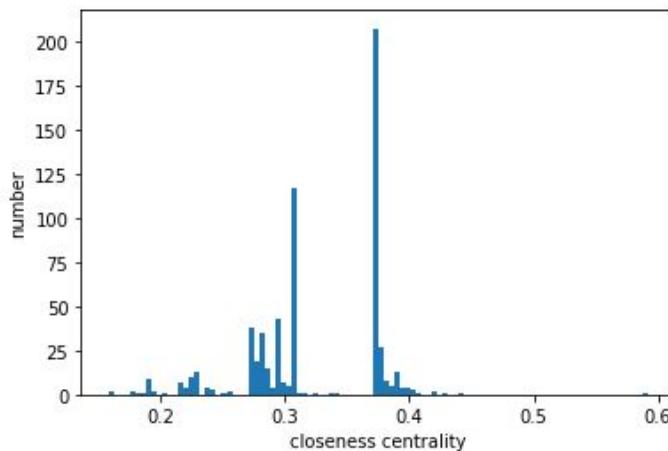
Rys. 83 Użytkownicy według stopnia w grafie Władysława Kosiniaka-Kamysza

Na rys. 83 można zauważyć, że prawie wszyscy użytkownicy mają bardzo niski i zbliżony do siebie stopień w sieci. Wyróżnia się jedynie oficjalne konto kandydata - KosiniakKamysz, co nie jest zaskakujące. Jest to najaktywniejsze konto w sieci.

```

Max closeness centrality:
user: KosiniakKamysz value: 0.5916824196597353
user: SondazeP value: 0.44397163120567373
user: AndrzejDuda value: 0.4275956284153005
user: kaliber308win value: 0.4204163868368032
user: JanParadowski_ value: 0.4181696726786907
user: pkgrabowski value: 0.40755208333333333
user: Krzyszt73414707 value: 0.4017971758664955
user: Wojciec41989761 value: 0.4017971758664955
user: mario_martens value: 0.4017971758664955
user: MikosMarcin value: 0.39746031746031746

```



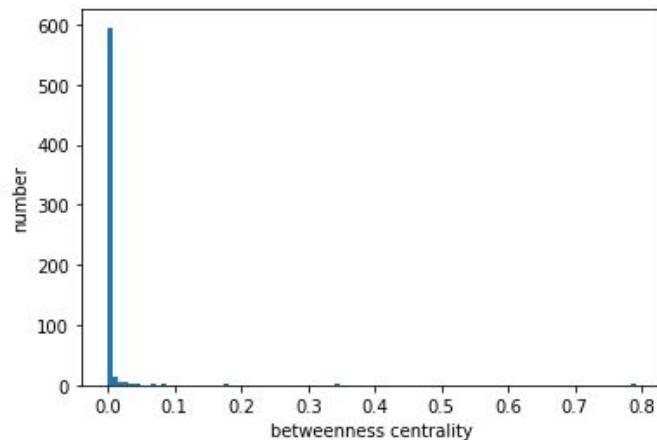
Rys. 84 Użytkownicy według bliskości w grafie Władysława Kosiniaka-Kamysza

Na rys. 84 można zauważyć, że miara bliskości użytkowników w sieci jest trochę bardziej równomiernie rozłożona niż stopień i żaden z nich pod tym względem się nie wyróżnia. Można zauważyć dwie większe grupy użytkowników - z wartościami około 0.3 i 0.4. Wynika z tego, że prawie wszyscy użytkownicy mają dosyć krótkie średnie odległości od pozostałych, co jest zapewne spowodowane faktem, że sieć jest dosyć mała. Kontem o największej efektywności jest KosiniakKamysz.

```

Max betweenness centrality:
user: KosiniakKamysz  value: 0.790508834829045
user: SondazeP  value: 0.3461221008397662
user: JanParadowski_  value: 0.1770149351859194
user: OjczynaMoja  value: 0.08271381120281167
user: TomaszPanfic86  value: 0.06578500293212085
user: kaliber308win  value: 0.04574546212266879
user: M_K_Blonska  value: 0.04195197322379435
user: Aneta_Marciszek  value: 0.03671347694034519
user: DariuszMatecki  value: 0.031718849840255595
user: LechNowacki  value: 0.029259834784856307

```



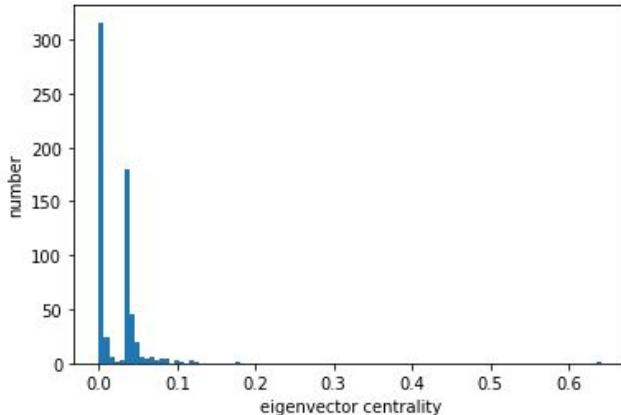
Rys. 85 Użytkownicy według pośrednictwa w grafie Władysława Kosiniaka-Kamysza

Na rys. 85 można zauważyć, że podobnie jak w przypadku stopnia, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość pośrednictwa w sieci. Ponownie zdecydowanie wyróżnia się konto KosiniakKamysz, przez które przechodzi większość połączeń w sieci.

```

Max eigenvector centrality:
user: KosiniakKamysz  value: 0.6416479302473788
user: magdasobkowiak  value: 0.17618313535404503
user: pkgrabowski  value: 0.12651283615693754
user: PSL_Malopolska  value: 0.1202472839782835
user: Aneta_Marciszek  value: 0.11673327423296298
user: MichałDziubak  value: 0.10430292914562982
user: JanParadowski_  value: 0.10113746117496629
user: SamOn07323044  value: 0.09823033959482118
user: SawickiEmil  value: 0.0876937470125599
user: Paslawska  value: 0.08548293501672545

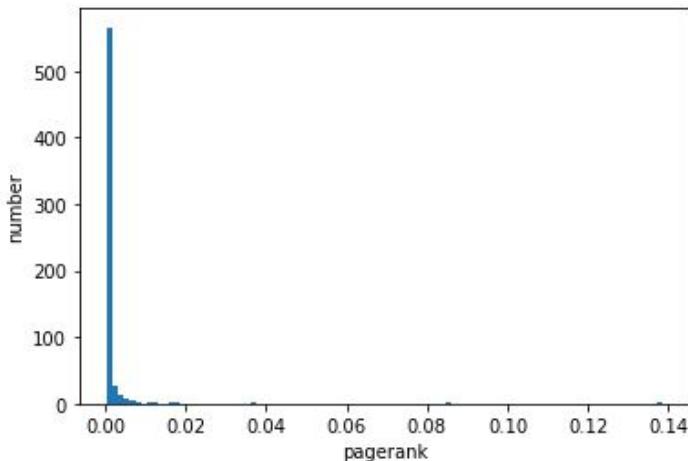
```



Rys. 86 Użytkownicy według wektora własnego w grafie Władysława Kosiniaka-Kamysza

Na rys. 86 można zauważyć, że podobnie jak w przypadku stopnia i pośrednictwa, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość wektora własnego. Tutaj również wyróżnia się konto KosiniakKamysz, co oznacza, że ma ono największy wpływ w sieci.

Max pagerank:
 user: KosiniakKamysz value: 0.13830577860781457
 user: SondazeP value: 0.08474050539838136
 user: JanParadowski value: 0.036975577507849196
 user: magdasobkowiak value: 0.017863657184872456
 user: OjczyznaMoja value: 0.016840884851250147
 user: Aneta_Marciszek value: 0.011805553720692846
 user: TomaszPanfic86 value: 0.011788492274440959
 user: PSL_Malopolska value: 0.010312215192450755
 user: pkgrabowski value: 0.008091147457843987
 user: DariuszMatecki value: 0.007957412588116922



Rys. 87 Użytkownicy według miary Page Rank w grafie Władysława Kosiniaka-Kamysza

Na rys. 59 widać, że w przypadku miary PageRank u wszystkich użytkowników w sieci jest ona bardzo niska, najwyższa wartość wynosi jedynie ok. 0.13.

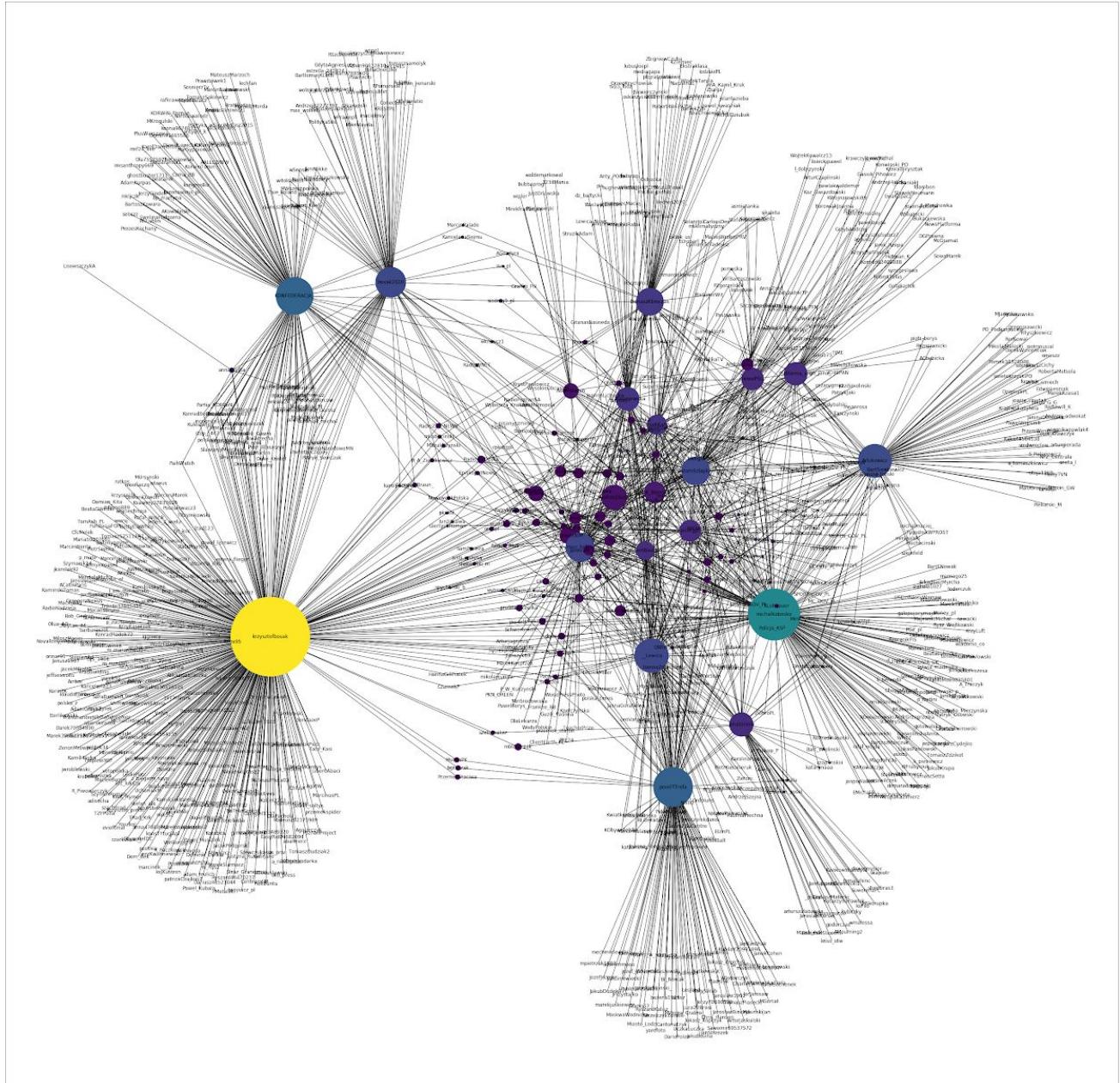
Analizując rysunki 83-87 można zauważyc, że wśród użytkowników o największych miarach centralności w grafie społeczności Władysława Kosiniaka-Kamysza znajdują się: KosiniakKamysz - oficjalne konto kandydata, magdasobkowiak - szefowa sztabu kandydata, AndrzejDuda - oficjalne konto prezydenta Andrzeja Dudy, M_K_Blonska - oficjalne konto Małgorzaty Kidawy-Błońskiej, pkgrabowski - działacz PSL, DariuszMatecki - radny PIS, Paslawska - posłanka PSL, SawickiEmil - członek sztabu, SondazeP - konto tweetujące o polityce i wyborach. Pozostałe konta to zwykli użytkownicy Twittera wspierający kandyturę Władysława Kosiniaka-Kamysza na prezydenta.

Wartości poszczególnych miar pokazanych na rysunkach 83-87 zostały znormalizowane. Podobnie jak w grafach poprzednich kandydatów w przypadku miar przedstawionych na rys. 83 i 85-87 (stopień, pośrednictwo, wektor własny, pagerank) prawie wszyscy użytkownicy mają wartości miar w przedziale 0-0.05, jedynie kilku ma miarę wyższą niż 0.1. Użytkownikiem wyróżniającym się spośród innych ze względu na wartości tych miar jest oficjalne konto Władysława Kosiniaka-Kamysza. Podobnie jak w sieci Roberta Biedronia w przypadku miary bliskości (rys. 84) najczęściej użytkowników ma wartości około 0.30 i 0.40, znacznie

mniej znajduje się w przedziale 0-0.20. Są to wyższe wartości niż dla pozostałych kandydatów, gdyż średnie odległości między wierzchołkami są niższe.

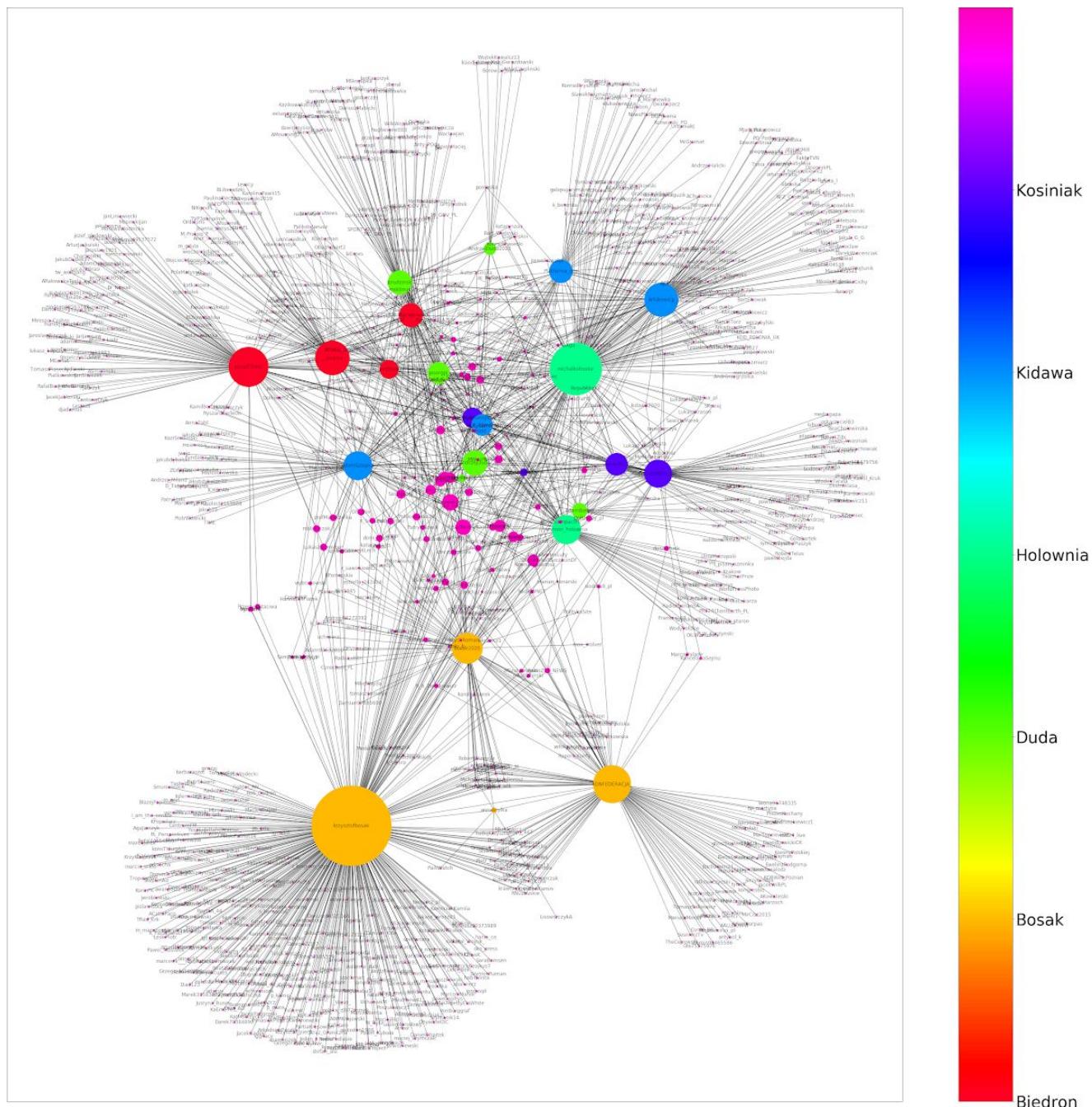
6.7 Sieć wszystkich kandydatów

Sieć społeczna stworzona przez konta wszystkich kandydatów oraz użytkowników, z którymi wchodzili oni w interakcje:



Rys. 88 Sieć społeczności wszystkich kandydatów

W grafie przedstawionym na rysunku 88 kolor wierzchołków zależy od ich stopnia (im jaśniejszy kolor tym większy stopień), a ich wielkość od miary pośrednictwa (im większy wierzchołek tym większe pośrednictwo). Najbardziej aktywnymi kontami są: krzysztofbosak, michalkobosko, konfederacja_, poseiT Treła, bosak2020 i Arlukowicz.



Rys. 89 Sieć społeczności wszystkich kandydatów z uwzględnioną przynależnością kont do zgrupowań

Na rysunku 89 przedstawiono tę samą sieć, ale z wierzchołkami przyporządkowanymi do poszczególnych kandydatów.

Podstawowe właściwości grafu:

```
Number of nodes: 1063
Number of edges: 1698
Max degree: 387
Min degree: 1
Density: 0.003008222119467874
Graph connected: True
Number of connected components: 1
Average clustering coefficient: 0.16666322185881993
Transitivity: 0.017948977965782497
Diameter: 5
Average distance between two nodes: 3.15
```

Rys. 90 Podstawowe właściwości grafu wszystkich kandydatów

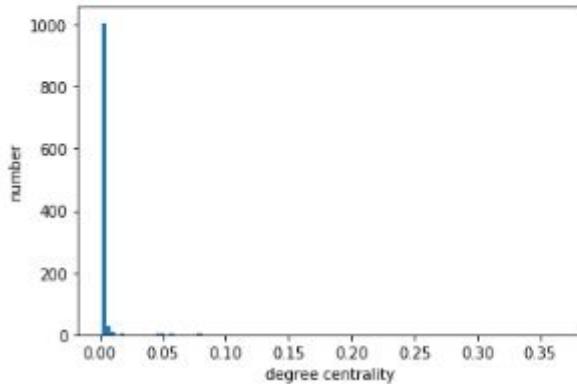
Jak widać na rysunku 90 graf stworzony przez konta wszystkich kandydatów i osób z nimi związanych oraz użytkowników, z którymi wchodzili w interakcje jest grafem spójnym. Oznacza to że między każdym ze sztabów istnieje jakieś połączenie - bezpośrednie lub pośrednie, np. poprzez retweetowanie tweetów tego samego konta. Gęstość i współczynnik grupowania są wyższe, a średnica i średnia odległość między wierzchołkami niższe, niż w przypadku społeczności poszczególnych kandydatów.

Miary centralności:

```

Max degree centrality:
user: krzysztofbosak      value: 0.3644067796610169
user: michalkobosko        value: 0.1784331450094162
user: KONFEDERACJA_         value: 0.1167608286252354
user: poselTTrela          value: 0.11205273069679848
user: Arlukowicz           value: 0.0847457627118644
user: bosak2020             value: 0.08097928436911488
user: adamSziapka          value: 0.0800376647834275
user: __Lewica              value: 0.0790900451977401
user: szymon_holownia       value: 0.07627118644067796
user: DariuszKlimczak       value: 0.06120527306967985

```



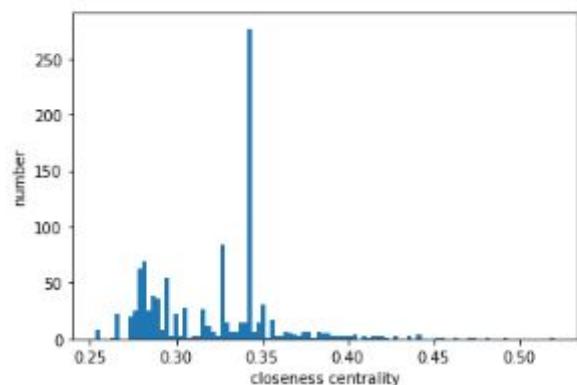
Rys. 91 Użytkownicy według stopnia w grafie wszystkich kandydatów

Na rys. 91 można zauważać, że prawie wszyscy użytkownicy mają bardzo niski i zbliżony do siebie stopień w sieci. Najbardziej wyróżnia się oficjalne konto Krzysztofa Bosaka - krzysztofbosak. Jest to najaktywniejsze konto w sieci.

```

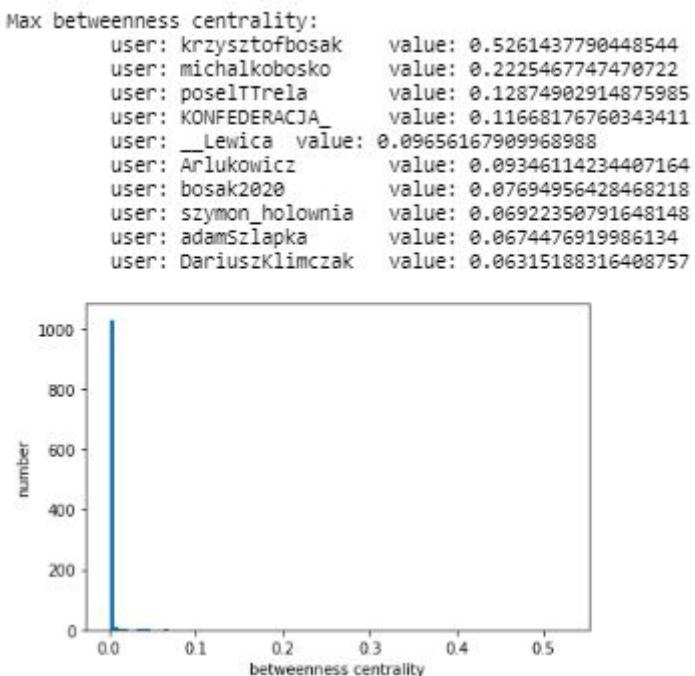
Max closeness centrality:
user: krzysztofbosak      value: 0.5200783545543585
user: AndrzejDuda          value: 0.4909847434119279
user: michalkobosko         value: 0.4822888283378747
user: M_K_Blonska            value: 0.4747429593205186
user: MorawieckiM           value: 0.46949602122015915
user: __Lewica              value: 0.4617391304347826
user: PolsatNewsPL           value: 0.45384615384615384
user: Jaroslaw_Gowin          value: 0.45076400679117146
user: SasinJacek              value: 0.44139650872817954
user: PremierRP                value: 0.44084682440846823

```



Rys. 92 Użytkownicy według bliskości w grafie wszystkich kandydatów

Na rys. 92 można zauważyc, że miara bliskości użytkowników w sieci jest już bardziej równomiernie rozłożona niż stopień i żaden z nich pod tym względem się nie wyróżnia. Wynika z tego, że większość użytkowników ma dosyć krótkie średnie odległości od pozostałych.



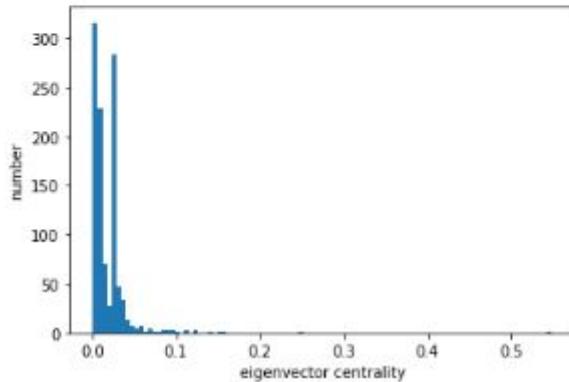
Rys. 93 Użytkownicy według pośrednictwa w grafie wszystkich kandydatów

Na rys. 93 można zauważyc, że podobnie jak w przypadku stopnia, prawie wszyscy użytkownicy mają bardzo niską i zbliżoną do siebie wartość pośrednictwa w sieci. Ponownie najbardziej wyróżnia się konto - krzysztofbosak, które ma największą kontrolę w sieci.

```

Max eigenvector centrality:
    user: krzysztofbosak      value: 0.5475701332698117
    user: michalkobosko       value: 0.2489808729642595
    user: szymon_holownia     value: 0.1591463256148735
    user: KONFEDERACJA_       value: 0.14929466268840313
    user: __Lewica           value: 0.14195603220412814
    user: poselTTrela         value: 0.12338694989016892
    user: KosiniakKamysz     value: 0.1212973284450516
    user: adamSzlapka        value: 0.11515292017540081
    user: M_K_Blonska         value: 0.11332133781781023
    user: bosak2020           value: 0.11087005268639528

```



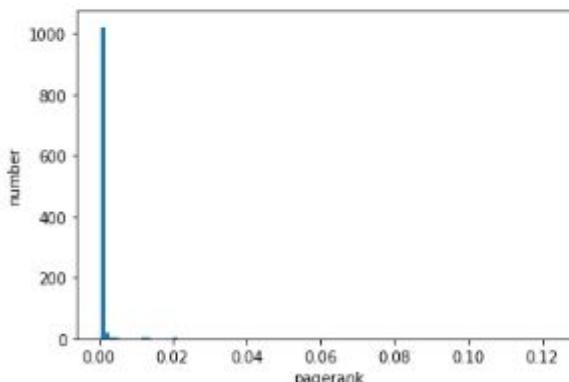
Rys 94. Użytkownicy według wektora własnego w grafie wszystkich kandydatów

Na rys. 94 można zauważyc, że podobnie jak w przypadku stopnia i pośrednictwa, prawie wszyscy użytkownicy mają niską wartość wektora własnego. Zakres tych wartości jest jednak trochę szerszy. Tutaj również wyróżnia się konto krzysztofbosak, co oznacza, że ma ono największy wpływ w sieci.

```

Max pagerank:
    user: krzysztofbosak      value: 0.1241542247403521
    user: michalkobosko       value: 0.047898717580923574
    user: KONFEDERACJA_       value: 0.03535584306702796
    user: poselTTrela         value: 0.03239635760553253
    user: Arlukowicz          value: 0.024142443408681855
    user: bosak2020            value: 0.023517064740914504
    user: adamSzlapka         value: 0.02091384525457343
    user: __Lewica             value: 0.020821477378606904
    user: szymon_holownia     value: 0.01924362003586645
    user: DariuszKlimczak     value: 0.016679793479059974

```



Rys. 95 Użytkownicy według miary Page Rank w grafie wszystkich kandydatów

Na rys. 95 widać, że w przypadku miary PageRank u wszystkich użytkowników w sieci jest ona bardzo niska, najwyższa wartość wynosi jedynie ok. 0.12 i ponownie należy do konta krzysztofbosak.

Na rysunkach 91-95 można zauważyc, że wśród użytkowników o największych miarach centralności w grafie społeczności wszystkich kandydatów znajdują się głównie kandydaci, ich sztabi i partie, ale także posłowie z różnych ugrupowań. Najbardziej aktywni i istotni w sieci są krzysztofbosak i michalkobosko, którzy pojawiają się na górze każdego rankingu. Z przedstawionych rankingów wynika, że niektórzy posłowie mają większy wpływ w sieci niż pozostali kandydaci.

Wartości poszczególnych miar pokazanych na rysunkach 91-95 zostały znormalizowane. Podobnie jak w przypadku grafów poszczególnych kandydatów w przypadku miar przedstawionych na rys. 91 i 93-95 (stopień, pośrednictwo, wektor własny, pagerank prawie wszyscy użytkownicy mają wartości miar w przedziale 0-0.05, niewielu ma miarę wyższą niż 0.1. Użytkownikiem wyróżniającym się spośród innych ze względu na wartości tych miar jest Krzysztofa Bosak. Natomiast w przypadku miary bliskości (rys. 92) wartości te są rozłożone znacznie bardziej równomiernie między wartościami 0 a 0.50 z największą liczbą użytkowników z wartościami około 0.35.

7. Przykłady użycia

7.1 Uruchomienie aplikacji

1. Klonowanie repozytorium
2. Stworzenie wirtualnego środowiska (nieobowiązkowe)
3. instalacja potrzebnych bibliotek

```
gniedziela@c7caeef1658e769b17e9c1e0e254cb54939a3e982:~/agh/semestr6/io/TwitterElection$ python3 -m venv env
gniedziela@c7caeef1658e769b17e9c1e0e254cb54939a3e982:~/agh/semestr6/io/TwitterElection$ source env/bin/activate
(env) gniedziela@c7caeef1658e769b17e9c1e0e254cb54939a3e982:~/agh/semestr6/io/TwitterElection$ pip install -U pip
Cache entry deserialization failed, entry ignored
Collecting pip
  Using cached https://files.pythonhosted.org/packages/43/84/23ed6a1796480a6f1a2d38f2802901d078266bda38388954d01d3f2e821d/pip-20.1.1-py3-none-any.whl
Installing collected packages: pip
  Found existing installation: pip 9.0.1
    Uninstalling pip-9.0.1:
      Successfully uninstalled pip-9.0.1
Successfully installed pip-20.1.1
(env) gniedziela@c7caeef1658e769b17e9c1e0e254cb54939a3e982:~/agh/semestr6/io/TwitterElection$ pip install -r requirements.txt
Collecting certifi==2020.4.5.1
  Using cached certifi-2020.4.5.1-py2.py3-none-any.whl (157 kB)
Collecting chardet==3.0.4
  Using cached chardet-3.0.4-py2.py3-none-any.whl (133 kB)
Collecting click==7.1.1
  Using cached click-7.1.1-py2.py3-none-any.whl (82 kB)
Collecting cyclere==0.10.0
  Using cached cyclere-0.10.0-py2.py3-none-any.whl (6.5 kB)
Collecting decorator==4.4.2
  Using cached decorator-4.4.2-py2.py3-none-any.whl (9.2 kB)
Collecting Flask==1.1.2
  Using cached Flask-1.1.2-py2.py3-none-any.whl (94 kB)
Collecting gunicorn==20.0.4
  Using cached gunicorn-20.0.4-py2.py3-none-any.whl (77 kB)
Collecting idna==2.9
  Using cached idna-2.9-py2.py3-none-any.whl (58 kB)
Collecting itsdangerous==1.1.0
  Using cached itsdangerous-1.1.0-py2.py3-none-any.whl (16 kB)
Collecting Jinja2==2.11.2
  Using cached Jinja2-2.11.2-py2.py3-none-any.whl (125 kB)
Collecting MarkupSafe==1.1.1
  Using cached MarkupSafe-1.1.1-cp35-cp35m-manylinux1_x86_64.whl (27 kB)
Collecting networkx==2.4
  Using cached networkx-2.4-py3-none-any.whl (1.6 MB)
Collecting numpy==1.18.3
```

4. uruchomienie serwera

Po kolej, wygląda to tak:

```
git clone https://github.com/mnabywan/TwitterElection
python3 -m venv env
pip install -r requirements.txt
cd server
gunicorn --bind 0.0.0.0:5000 wsgi:app
```

Aplikacja zostanie uruchomiona na podanym jako parametr `--bind` adresie.

7.2 Przykłady korzystania z aplikacji

- polubienia i retweety

Twitter Election

Home

Wykresy ▾

- Polubienia i reetweety
- Obserwujący
- Przyjaciele
- Retweety wg dat
- Utwierdzone wg dat

Wordcloudy słów ▾

Wordcloudy tagów ▾

Tagi ▾

Odpowiedzi ▾

Najpopularniejsze słowa ▾

Betweenness centrality ▾

Closeness centrality ▾

Degree centrality ▾

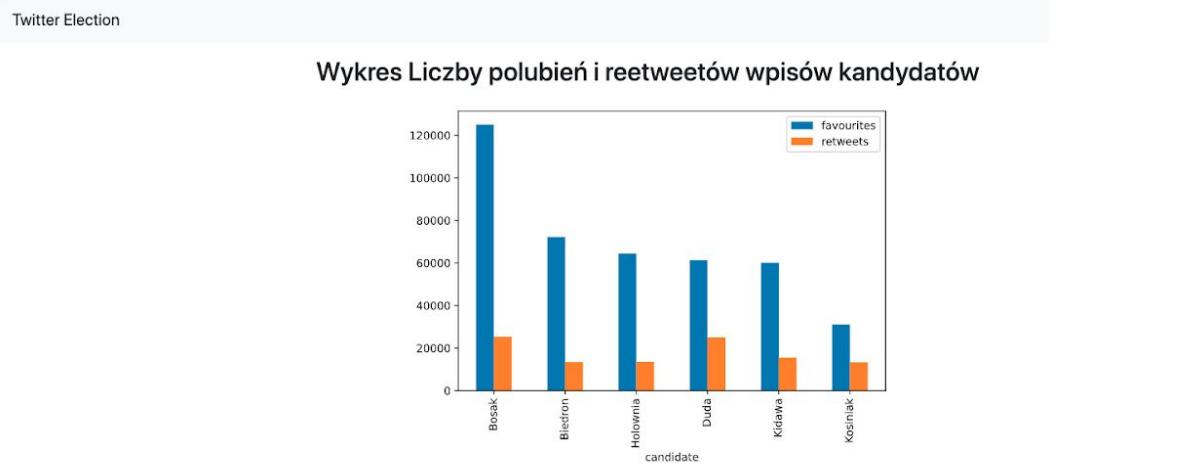
Eigenvector centrality ▾

Graph info ▾

Pagerank ▾

Hashtag graphs ▾

Rys. 96 Wykresy > Polubienia i retweety



Rys. 97 Wykres polubień i retweetów wpisów kandydatów

● Najpopularniejsze tagi

Twitter Election

Home

Wykresy ▾

Wordcloudy słów ▾

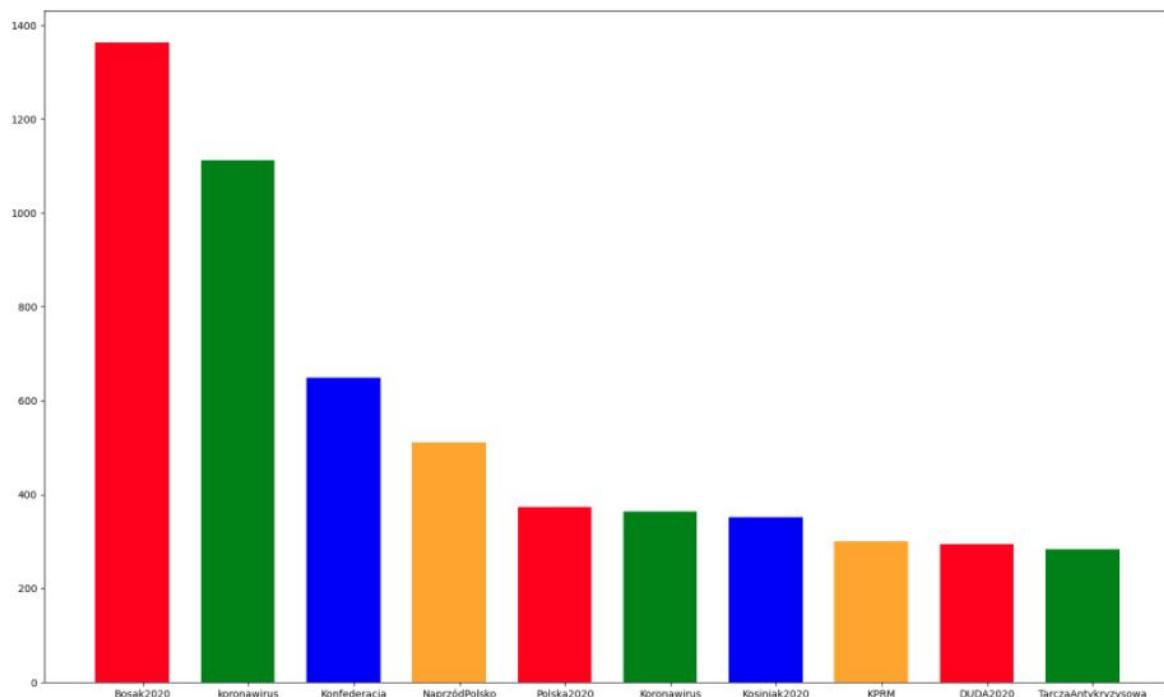
Wordcloudy tagów ▾

Tagi ▾

- Razem
- Biedron
- Bosak
- Duda
- Holownia
- Kidawa
- Kosiniak

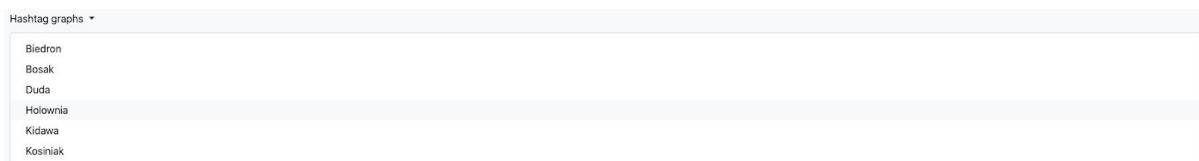
Rys. 98 Tagi > Razem

Najpopularniejsze tagi - wszyscy kandydaci



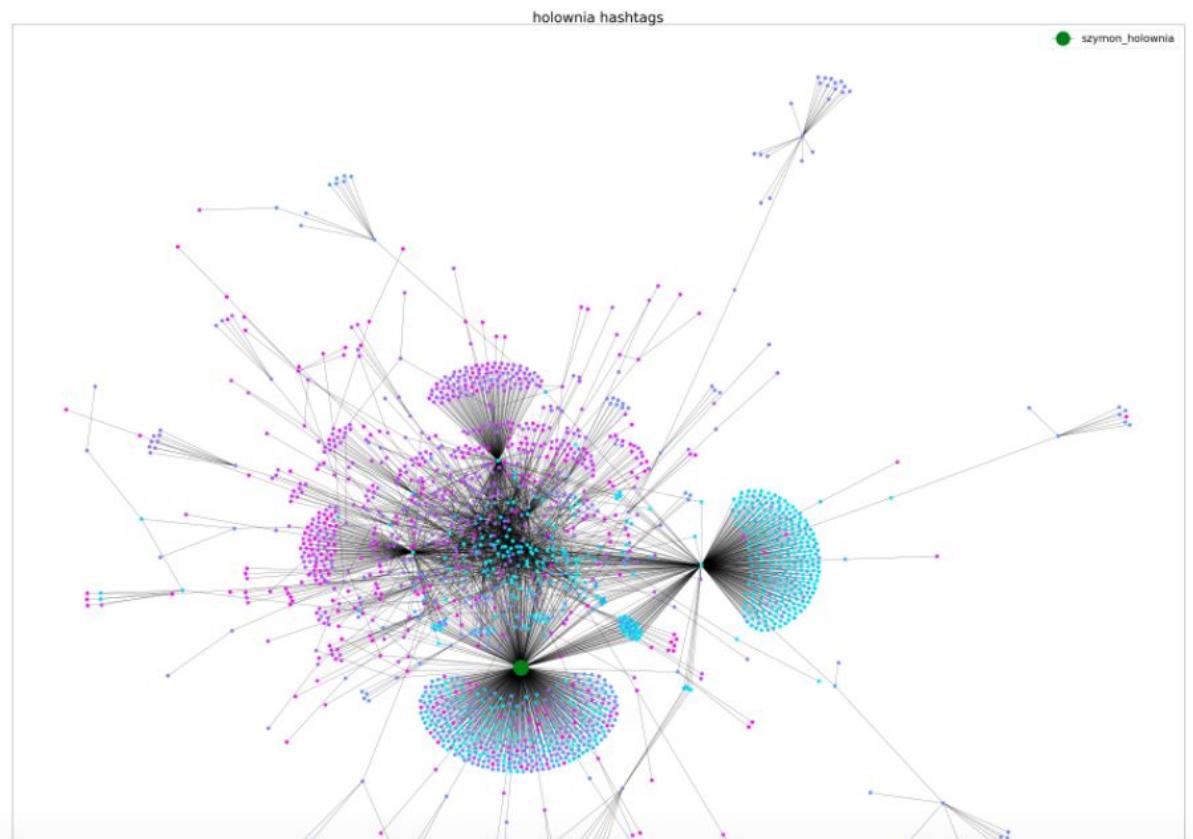
Rys. 99 Najpopularniejsze tagi dla każdego z kandydatów

- graf hashtagów



Rys. 100 Hashtag graphs > Hołownia

Hashtags - Holownia



Rys. 101 Graf hashtagów dla p. Hołowni

- retweety wg dat

Twitter Election

Home

Wykresy ▾

Polubienia i reetweety

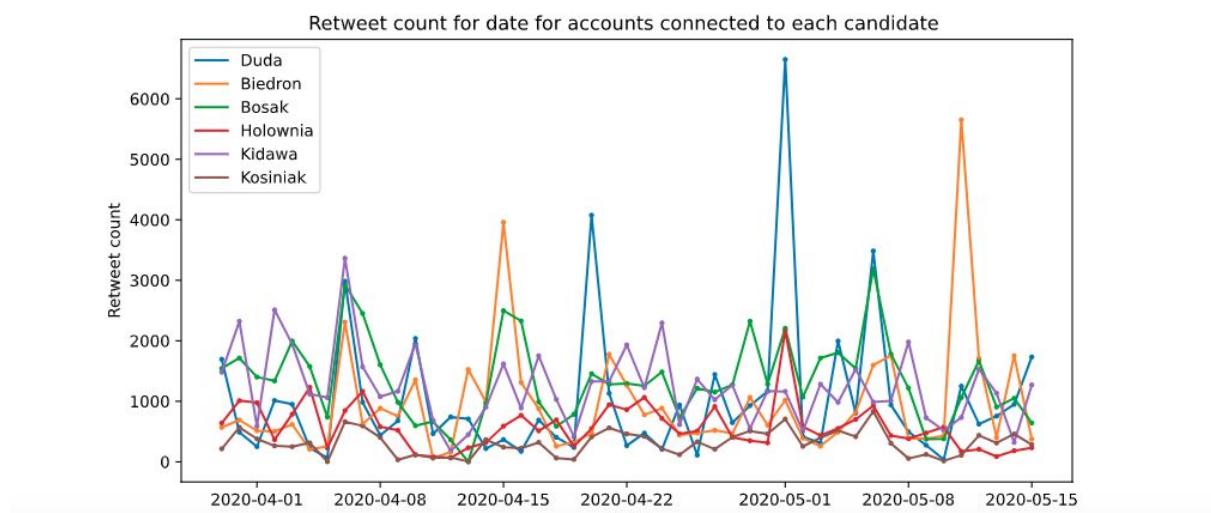
Obserwujący

Przyjaciele

Retweety wg dat

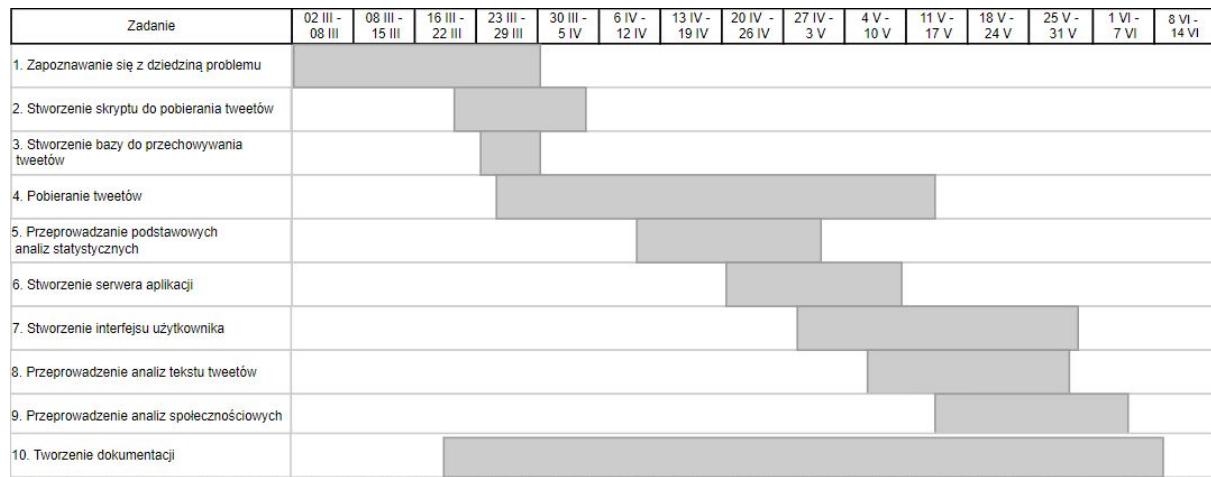
Rys. 102 Wykresy > Retweety wg dat

Liczba retweetów zsumowana wg dat dla wszystkich kont kandydatów



Rys. 103 Wykres retweetów według dat

8. Podział prac



Rys. 104 Diagram Gantta.

1. Mateusz Nabywaniec: zadania 1, 2, 3, 4, 5 10
2. Dominika Mlynarczyk: zadania 1, 5, 8, 9, 10
3. Grzegorz Niedziela: zadania 1, 6, 7, 10
4. Dominik Guz: zadania 1, 4, 5, 8, 10

8.1 Opis procesu wytwarzania

Realizację projektu zaczęliśmy od Pozyskania zbioru danych. W tym celu zgłosiliśmy do Twittera chęć pozyskania profesjonalnego klucza do API który miał większe możliwości. Ponadto w tym samym czasie został napisany skrypt, który zapisywał do bazy sqlite informację o tweetach: autora, sztab autora, id, ilość reweetów, likeów. Następnym krokiem było przeprowadzenie analiz statystycznych, ponadto zaczęliśmy szukać narzędzi, które umożliwią nam przeprowadzenie analiz społeczności i tekstowych. Zarówno SNA, jak i analiza tekstu potrzebowały jeszcze dodatkowych informacji w zbiorze danych, więc powstały odrębne skrypty, które dodatkowo pobierały przez API wymagane informacje. Po zdobyciu danych można było przeprowadzić analizy, jednocześnie rozpoczęła się praca nad serwerem i częścią webową projektu. Wyniki analiz i skrypty zostały umieszczone na serwerze. Cały projekt został zrealizowany w ustalonych przez nas ramach czasowych. Co ważne, nasz projekt wybrał szybkość nad elastyczność: wykresy są przygotowane wcześniej w postaci plików graficznych np JPG. Nie daje to możliwości ich dostosowania wedle własnych potrzeb, jednak sama aplikacja ładuje się bardzo sprawnie.

9. Podsumowanie

W ramach projektu od początku kwietnia do połowy maja tego roku zbierane były dane dotyczące wyborów prezydenckich w Polsce, pochodzące z portalu społecznościowego Twitter. Dane te zostały wykorzystane do przeprowadzenia licznych analiz statystycznych, analiz tekstu oraz analiz sieci społecznych. Powstała również aplikacja webowa umożliwiająca wyświetlanie wyników wykonanych analiz.

9.1 Wnioski końcowe z projektu

9.1.1 Wnioski z modułu statystycznego:

- Użytkownicy najczęściej reagowali na tweety pochodzące z kont kandydatów (bardziej niż np. konta partii, kampanii, rzeczników).
- Najaktywniejszymi sztabami pod względem liczby publikowanych tweetów były sztab Krzysztofa Bosaka oraz Roberta Biedronia.
- Większość kont związanych z partią popierającą kandydata tworzyła dużo tweetów, które jednak nie otrzymywały wielu reakcji
- Duża liczba tweetów nie przekładała się na liczbę polubień i retweetów w przeliczeniu na tweet. Ważniejsze są zasięgi danego konta.
- Konto prezydenta "AndzejDuda" publikowało zdecydowanie mniej tweetów niż konta konkurentów. Można podejrzewać, że pozostali kandydaci byli bardziej aktywni ponieważ łatwiej jest atakować, a ubiegający się o reelekcję może się tylko bronić wymieniając swoje osiągnięcia

9.1.2 Wnioski z modułu analizy tekstu:

- Język tweetów między kandydatami był bardzo zbliżony (podobieństwo w zakresie 45-80%). Można wywnioskować, że politycy operują schematycznym językiem pełnym wzorców.
- Pomimo iż panuje epidemia koronawirusa, nie jest ona wśród najpopularniejszych tematów tweetów.
- Najpopularniejsze słowa tweetów są powiązane z polityką np RZĄD. Można wywnioskować z tego, iż kandydaci w swoich wypowiedziach skupiają się przede wszystkim na komentowaniu wydarzeń politycznych.
- Jeśli chodzi o podobieństwo, najbardziej odróżniający był prezydent Andrzej Duda. Może to wynikać jego pozycji, gdzie inni kandydaci byli bezpartyjni lub reprezentowali opozycję
- ZDROWIE było jednym z najczęstszych słów używanych przez Władysława Kosiniaka-Kamysza (Reszta kandydatów skupiała się praktycznie tylko na

polityce). Być może jest to jeden z punktów, na których chce się skupić w swoim programem politycznym.

9.1.3 Wnioski z modułu analizy społeczności:

- Sieci społeczne powstałe dla większości kandydatów nie są szczególnie duże. Może to wynikać z faktu, że Twitter nie jest popularnym medium w Polsce. Ponadto, możliwe jest, że polityka nie znajduje się w centrum zainteresowań jego użytkowników.
- Największą sieć społeczną spośród wszystkich kandydatów ma Andrzej Duda. Oznacza to, że jego kandyatura wzbudza największe zainteresowanie wśród użytkowników Twittera. Wynika to zapewne z faktu, że jest on obecnym prezydentem i startuje z partii o bardzo dużym poparciu w Polsce. Przez to jest także głównym celem ataków opozycji i w jego sieci znajduje się wiele osób mu przeciwnych, podczas gdy w przypadku pozostałych kandydatów, są to głównie zwolennicy.
- Najmniejszą sieć społeczną ma Władysław Kosiniak-Kamysz. Można z tego wywnioskować, że ma on najmniejsze poparcie wśród użytkowników Twittera.
- W sieciach poszczególnych kandydatów największe znaczenie mają ich oficjalne konta, konta ich sztabów i partii. Oczywiście jest to zrozumiałe, gdyż konta te zajmują się promowaniem kandydatów, których reprezentują, więc często publikują tweety z hashtagami związanymi z kampanią. Ponadto mają one duże zasięgi i pozostały użytkownicy Twittera często odnoszą się do ich postów poprzez ich retweetowanie, cytowanie czy odpowiadanie na nie.
- Miary, za pomocą których badaliśmy sieci społeczne, rozkładają się podobnie w przypadku wszystkich kandydatów.
- Graf społeczności zbudowany na podstawie interakcji wszystkich kandydatów i ich sztabów jest grafem spójnym, z czego wynika, że między każdym ze sztabów istnieje jakieś połączenie - bezpośrednie lub pośrednie, powstałe, gdy sztabi retweetują czy odpowiadają na tweety należące do tego samego użytkownika.

9.2 Ewaluacja oceny ryzyka

Ryzyko wynikające z czynników zewnętrznych:

- możliwość zachorowania związana z sytuacją epidemiczną - udało nam się nie zachorować, przez co mogliśmy równo dzielić się pracą

- trudności w komunikacji związane ze słabą jakością internetu - mieliśmy pewne problemy, jednak udawało nam się dobrze skomunikować ze sobą i klientem aby omawiać postępy prac
- ograniczona ilość czasu związana z innymi zajęciami na uczelni, pracą itp. - daliśmy radę podzielić pracę w ten sposób, aby nie poświęcać zbyt dużo czasu na ten projekt. Udało nam się również w miarę mieścić się w deadlinach.

Ryzyko wynikające z czynników wewnętrznych:

- ograniczenia związane z API Twittera i ewentualna zbyt mała liczba tweetów - regularnie (ok. co tydzień) pobieraliśmy dane dotyczące wyborów, przez co nie mamy większych luk w bazie, co pozwoliło na dobre przeanalizowanie tworzenia tweetów np. w zależności od czasu.
- mała możliwość analizy tweetów z powodu zawieszenia kampanii/odwołania wyborów - okazało się, że istotnie wybory zostały przesunięte, ale decyzja zapadła późno, przez co kandydaci nadal prowadzili kampanię na Twitterze, co pozwoliło zebrac i analizować dane
- mała ilość danych do analizy np. najczęściej używanych słów gdyż głównym tematem jest epidemia koronawirusa - analiza tekstu wykazała, że mimo wyjątkowej sytuacji, kandydaci nie skupiali się na epidemii lecz na kampanii wyborczej, widać to dobrze w mapach słów związanych z kandydatami.
- dane i wyniki analiz mogą być przedstawione mało zrozumiałe - stworzyliśmy odpowiedni widok aplikacji, moduły tworzą prezentując dane w odpowiedni sposób, przez co wyniki analiz są dobrze widoczne.
- Dane mogą być przechowywane w sposób niewygodny i utrudniający analizę, zbyt wolne zapytanie o tweet przez id - zmieniliśmy koncepcję bazy danych. Oprócz bazy SQLite używamy plików JSON, które zawierają pełną treść tweeta, a także informacje o autorze. Dzięki temu możliwa jest analiza tych tweetów pod kątem tekstu i społeczności.

9.3 Wnioski z projektu dotyczące procesu wytwarzania produktu

Podczas wykonywania projektu pewne etapy trwały dłużej niż wynikałoby to z diagramu Gantta. Był to etap 3 (stworzenie bazy) z rysunku 96. W miarę rozwoju projektu musieliśmy zmienić koncepcję przechowywania danych, aby umożliwić pobieranie informacji o tekście tweetu oraz dane społecznościowe (któretweetował, odpowiadał na tweet). Rozważaliśmy także inne koncepcje (użycie MongoDb) opisane w punkcie 3.4. Z tego powodu ten etap trwał dłużej. Poszczególne analizy wymagały także doboru i poznania narzędzi, ale też można je było zrównoleglić (poszczególne rodzaje analiz były od siebie niezależne), więc prace nad nimi zaczęły się szybciej i ogólnie trwały przez większy okres czasu.

Projekt utwierdził nas w przekonaniu, że regularna praca i walidacja osiągnięć jest kluczowa w procesie wytwarzania oprogramowania. Ważne jest poświęcenie

czasu na początkową analizę problemu oraz wykonanie schematu oraz planu systemu. Czas ten nie powinniśmy traktować jako stracony, lecz jako inwestycja, ponieważ naprawianie błędów spowodowanych niepoprawną analizą kosztuje więcej czasu i stwarza więcej problemów niż dobra analiza.

Podczas wytwarzania oprogramowania w grupie ważna jest komunikacja między poszczególnymi osobami oraz inteligentne rozdzielenie zadań by optymalnie wykorzystać możliwości i czas każdego członka. Dzięki temu praca idzie sprawniej, a każdy z nas może rozwijać element, który go interesuje.

Częsta walidacja projektu przez klienta jest kluczową kwestią: najważniejszym celem jest dostarczenie jak najlepszego produktu, dlatego rozmowa o dotychczasowych dokonaniach, słuchanie uwag i dialog na temat rozwiązań pozwala dostarczyć oprogramowanie najbardziej zbliżone do potrzeb interesanta.

9.4 Ocena użyteczności i kierunek rozwoju projektu

W trakcie realizowania projektu z powodu epidemii koronawirusa wybory prezydenckie zostały przełożone na 21 czerwca. Został także zmieniony sposób przeprowadzania, umożliwiono bowiem głosowanie korespondencyjne. Jednym z pomysłów przed przystąpieniem do projektu było sprawdzenie w jaki sposób aktywność kandydatów na Twitterze przełożyła się na wynik wyborczy. Z opisanych wcześniej powodów stało się to niemożliwe. Zebraliśmy dane od marca do połowy maja więc nie bardzo można użyć ich do analizy wyników wyborów tym bardziej, że po ogłoszeniu nowej daty wyborów, z wyborów wycofała się Małgorzata Kidawa-Błońska (nowym kandydatem KO został Rafał Trzaskowski). Z powodu ograniczeń API Twittera (7 dni na pobieranie tweetów z danym hashtagiem, 30 dni na pobranie tweetów od danego użytkownika) nie ma możliwości kontynuowania zbiorów danych tak aby nie wystąpiła przerwa między zebranymi a nowymi danymi.

Zebrane dane pozwalają jednak na prowadzenie dalszych analiz zarówno na polu statystycznym, tekstowym i społeczności. Co do danych statystycznych, można próbować określić dane ludzi tweetujących o danym kandydacie kandydacie i stworzyć wykres tweetów z podziałem np. na deklarowany wiek. Analiza tekstu mogłaby pozwolić stwierdzić, czy tweetujący to naprawdę zainteresowani użytkownicy czy może trolle sprzyjający danemu kandydatowi, powielający dane treści.

Stworzony przez nas system możnaby wykorzystać do zbierania i analizy danych dotyczących innych wyborów. W tym celu należałyby zdefiniować odpowiednie tabele dotyczące kandydatów, czy też partii oraz zidentyfikować konta powiązane z nimi.

Projekt może być też rozwinięty pod kątem większego wpływu użytkownika, by miał on na przykład wpływ na parametry tworzonych wykresów, gdyż teraz są one generowane wcześniej i statyczne.

Innym kierunkiem rozwoju jest powiększenie źródła danych o dodatkowe witryny, na przykład facebook czy instagram. Dzięki temu analizy byłyby bardziej miarodajne.

9.5 Wnioski dotyczące używanych technologii

Cały nasz projekt był stworzony w języku programowania Python. Okazał się to być bardzo dobry wybór do stworzenia tego właśnie projektu. Od bibliotek do API Twittera, przez komunikację z bazami danych, komponentów do przeprowadzania i wizualizacji analiz, aż do aplikacji webowej, w której ich wyniki były pokazywane użytkownikowi - na każdym z tych etapów było wystarczająco możliwości oraz dokumentacji, co pozwoliło nam w miarę sprawnie wykonywać kolejne etapy projektu. Jeżeli mielibyśmy robić ten projekt od nowa, cały team zgodnie uważa ten język za najlepszy wybór.

Baza danych sqlite byłaby dobrym wyborem przy analizie mniejszej ilości danych. Podczas zmian struktury projektu i dodawania kolejnych analiz doszliśmy do wniosku, że szybsza nierelacyjna baza danych np. MongoDB byłaby lepszym wyborem.

Pandas jako podstawa analiz była świetnym wyborem. Jest to bardzo dopracowana biblioteka o dużych możliwościach, dokładnej dokumentacji i niskiej krzywej uczenia. Sprawdziła się ona zarówno podczas analizy tekstu, statystyce, jak i jako podstawa SNA. Radzi sobie ona z dużymi zbiorami danych, a jej wyniki można łatwo przedstawić w formie wykresu.

Jeżeli chodzi o analizę tekstową, ilość narzędzi radzących sobie z językiem polskim jest na razie mocno ograniczona. Morfeusz jako stemmer słów był bardzo skuteczny i spełniał nasze oczekiwania, jednak jego wyjście czasami trzeba było też odpowiednio przyciąć o końcówki nie należące do podstawowego słowa (rząd:v1 jako wyjście ze stemizacji). Trwają jednak prace nad innymi narzędziami, np. SpaCy ma zacząć obsługiwać język polski. Ponadto istnieją także płatne narzędzia do tego

W przypadku analiz sieci społecznych biblioteka NetworkX okazała się dobrym wyborem. Udostępnia wiele różnych funkcji pomocnych przy analizie grafów, dzięki czemu umożliwiła badanie powstałych sieci pod wieloma względami. Jedyną wadą była mała szybkość działania niektórych funkcji, szczególnie służących do tworzenia grafu i obliczania miar centralności. W przypadku większych zbiorów danych może być to problematyczne.

9.6 Co byśmy zmienili, gdybyśmy wykonali projekt od początku

Z racji na optymalność wielu narzędzi, tylko mała część zostałaby zmieniona.

Jeśli wymagane byłoby użycie bazy MongoDb to moglibyśmy ją wykorzystać, lecz z racji braku tych wymagań i wystarczająco dobrej bazy SQLite do potrzeb projektu nie wybraliśmy jej.

10. Raporty ze spotkań

Ten punkt został zamieszczony celem ukazania rozwoju całego projektu, zarówno aplikacji, jak i dokumentacji. Cotygodniowe raporty zawierają informacje co uzyskaliśmy w danym tygodniu, o czym dyskutowaliśmy na spotkaniu z klientem oraz co zostało ustalone.

10.1 Spotkanie z dnia 24.03.2020

Data i godzina spotkania: 24.03.2020, godzina 19:00

Obecni członkowie Zespołu Projektowego (Zespół 1):

Guz Dominik (Lider) - dominikguz1@gmail.com

Młynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Lista dyskutowanych tematów/ problemów, ew. rozwinięcie:

Kwestia użytej technologii i jej ograniczeń (zbieranie danych z twittera przez API), pomysły na analizę danych.

Ustalenia:

- Dane należy zacząć zbierać jak najszybciej
- Projekt będzie wykonany w Pythonie
- Wstępnie chcielibyśmy zrobić sondaż popularności kandydatów i przewidzieć ich wyniki (albo w tej sytuacji czy i kiedy się odbędą)
- Musimy zobaczyć jakie dane możemy pozyskiwać z korzystając z API twittera (np. wiek, lokalizacja użytkowników, którzy wypowiedzieli się na temat kandydata).

Deadline:

-ustalenie tego co chcemy uzyskać, a następnie prezentacja na po świętach

Co zostało do ustalenia:

Do ustalenia zostało jeszcze w jaki sposób chcemy zbierać dane i co dokładnie chcemy z nich wyłuskać oraz w jakie jeszcze sposoby możemy je przeanalizować.

10.2 Spotkanie z dnia 31.03.2020

Data i godzina spotkania: 31.03.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Guz Dominik (Lider) - dominikguz1@gmail.com

Mlynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. rozwinięcie:

- problem z otrzymaniem klucza do zespołowego konta Developer Twitter
- problem z wyborem interesujących danych (na podstawie autorów tweetów,
- obecna sytuacja (możliwa zmiana terminu wyborów), możliwość zmiany pozyskiwanych danych

Ustalenia:

- jeśli nie uzyskamy dostępu do zespołowego konta Developer Twitter to zacząć zbierać dane nie korzystając z klucza z konta zespołowego
- dane należy zacząć zbierać jak najszybciej (mamy dostęp do tweetów z ostatnich 30 dni)
- należy stworzyć schemat pozyskiwania danych z Twittera i wybrać bazę w której będziemy przechowywać informacje.
- warto sporządzić analizę danych które chcemy zbierać na podstawie m.in. autorów, hashtagów:

- stworzyć listę kandydatów, szefów ich sztabów ewentualnie osób z otoczenia zaangażowanych mocno w ich kampanię
- Stworzyć listę hashtagów dotyczących kandydatów, opierających się o hasła wyborcze (np. #Duda2020)
- Zastanowić się nad aktywnością najpopularniejszych dziennikarzy (często przychylnych danemu obozowi politycznemu), stworzyć ich listę

- powinniśmy zacząć od uzyskiwania podstawowych informacji i wykresów z twittów (ilości lajków, zasięgi hashtagów) a dopiero później zastanowić się nad szczegółowym przetwarzaniem danych np. analiza najczęstszych pojęć używanych przez kandydatów (np. Praca, imigracja, przyszłość, zdrowie)

- warto zacząć poznawać narzędzia umożliwiające przetwarzanie danych
- w związku z sytuacją epidemiczną, możliwą zmianą terminu wyborów możliwa będzie zmiana analizowanych danych tak aby było ich wystarczająco dużo.

Deadline:

- wysłanie raportu - w dniu spotkania
- stworzenie fragmentu dokumentacji zawierającej: opis problemu, cel pracy, wizję projektu, plan prac, używane narzędzia - do 6 kwietnia (poniedziałek) godz. 20:00

10.3 Spotkanie z dnia 07.04.2020

Data i godzina spotkania: 07.04.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Guz Dominik (Lider) - dominikguz1@gmail.com

Mlynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. Rozwinięcie:

- Recenzja obecnego fragmentu dokumentacji
- Prezentacja rodzaju i ilości obecnie zebranych danych
- Data i forma kolejnego spotkania

Ustalenia:

- dokumentacja zaakceptowana, zastrzeżenia odnośnie punktu 3 "Wizja rozwiązania"
- należy dokładnie wyszczególnić jakie analizy chcemy wykonywać na naszych danych, w jaki sposób chcemy tego dokonać i jakich wyników się spodziewamy
- warto dodać punkt 3.4 (schemat architektury)
- dodać punkt 3.5 (oczekiwane wyniki) - powinna znaleźć się tam lista analiz zawierająca te, których oczekuje klient (analizy statystyczne, analizy oparte o eksplorację danych i badanie sieci społecznych)
- należy rozwinąć punkt 4 (podział prac) - warto zamieścić diagram Gantta, wykonać tabelę z podziałem na tygodnie uwzględniając daty z możliwym przesunięciem związanym z ewentualnymi problemami
- należy uzasadnić wybór kandydatów (to że nie badamy kont mniej znanych kandydatów)

- Należy uzasadnić listę wybranych dziennikarzy (czy było to pod względem popularności, ilości followersów, czy np. subiektywnego wyboru), dodać do źródeł
- Warto wykonać wstępную analizę ryzyka, co może się nie udać w projekcie i jak temu chcemy przeciwdziałać
- Przygotować informacje o kontaktach kandydatów i innych kont członków sztabu (ilość followersów, ilość dotychczasowych twittów itd.)
- Należy rozwinać punkt 2 (cel i wymagania) - sprecyzować jakie tweety chcemy pobierać i w jakim celu, jak mają być one przechowywane (format), jakie analizy będziemy wykonywać
- Spotkanie za tydzień będzie normalnie w godzinach zajęć

Deadline:

- wysłanie raportu - w dniu spotkania
- Uzupełnienie dokumentacji o podane punkty - do 14 kwietnia
- Prezentacja postępów w projekcie - 21 kwietnia

10.4 Spotkanie z dnia 14.04.2020

Data i godzina spotkania: 14.04.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Guz Dominik (Lider) - dominikguz1@gmail.com

Mlynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. Rozwinięcie:

- Recenzja obecnego fragmentu dokumentacji
- Prezentacja proponowanych analiz
- Prezentacja rodzaju i ilości obecnie zebranych danych
- Wygląd prezentacji na następnym spotkaniu

Ustalenia:

- Zadowalająca ilość proponowanych analiz, jednak należy opisać je jeszcze bardziej szczegółowo, dodać konkretne scenariusze

- Uszczegółoić zebrane dane, dokonać wstępnej analizy - ile mamy tweetów z danego konta, ile followersów ma dane konto co będzie świadczyło o aktywności i popularności kandydatów
- Należy najpierw skoncentrować się na analizie popularności, analiza tekstu w dalszej części projektu
- W bazie należy zbierać nie tylko id ale także ilość polubień, retweetów, datę utworzenia, aby analizowanie danych było wydajne
- Należy rozwinąć punkt 2 (zagrożenia) - należy oszacować ryzyko i posortować od najwyższego
- Należy rozwinąć opis architektury - obrazek stanowi dobry wstęp, ale trzeba dokładniej opisać strukturę bazy, dodać schemat w UML-u,
- Przygotować prezentację na następne spotkanie - opis problemu, wizja, koncepcja, plan działań do 15 czerwca z 2 tygodniowym buforem

Deadline:

- wysłanie raportu - w dniu spotkania
- Uzupełnienie dokumentacji o podane punkty - do 21 kwietnia
- Prezentacja postępów w projekcie ok. 20 min - 21 kwietnia

10.5 Spotkanie z dnia 05.05.2020

Data i godzina spotkania: 05.05.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Guz Dominik (Lider) - dominikguz1@gmail.com

Mlynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. Rozwinięcie:

- Recenzja obecnego fragmentu dokumentacji
- Prezentacja proponowanych analiz
- Prezentacja harmonogramu
- Prezentacja przechowywanych danych

Ustalenia:

- Należy zastanowić się nad strukturą bazy i uzasadnić jej strukturę - czy nie będzie redundancji danych
- Należy opisać priorytety hashtagów, które odnoszą się do tweetów zawartych w tabeli "election_tweets" (tweety odnoszące się do ogólnych hashtagów wyborczych)
- Zadowalająca ilość proponowanych analiz, jednak należy opisać je jeszcze bardziej szczegółowo, dodać konkretne scenariusze
- Należy dokładniej napisać które dane będą używane w analizach, na podstawie jakiej listy słów znajdziemy informacje jakie o jakich tematach najczęściej tweetują kandydaci
- Umieścić informacje o rozważanych narzędziach w analizie tekstu i społeczności
- Należy opisać schemat architektury w UMLu

Deadline:

- wysłanie raportu - w dniu spotkania
- Uzupełnienie dokumentacji o podane punkty - do 11 maja

10.6 Spotkanie z dnia 19.05.2020

Data i godzina spotkania: 19.05.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Mlynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. Rozwinięcie:

- Recenzja obecnego fragmentu dokumentacji
- Propozycja zmiany bazy danych z SQLite na MongoDB
- Prezentacja nowej analizy statystycznej aktywności kandydatów
- Prezentacja proponowanych analiz tekstu
- Prezentacja analiz społecznościowych

Ustalenia:

- Należy opisać analizę statystyczną aktywności kandydatów od czasu, zaprezentować wszystkich kandydatów na wykresie, podpisać wykres i skomentować (co wynika z tego, kto jest najbardziej aktywny itd.)
- Warto porównać jak przekłada się ilość tweetów na ilość polubień oraz retweetów - wydaje się że jest to ściśle związane, ale warto sprawdzić. Można także wyliczyć np. średnią.
- Do analiz tekstu należy podać wzór względem którego będzie liczone podobieństwo, a także należy bardziej je rozwijać
- Przy analizach społeczności wypisać 10 najbardziej aktywnych użytkowników
- Przy opisie narzędzi dodać odnośniki w tekście do dokumentacji.
- Przy analizach społeczności dodać obliczanie gęstości grafu, PageRank, podsumować i zinterpretować wykresy, dodać wykresy z różnych okresów czasu, dodać histogram przedstawiający miary centralności użytkowników twittera, opisać użytkowników z wysokimi miarami centralności, dodać informacje o całym grafie przed wyodrębnieniem największej spójnej składowej

Deadline:

- wysłanie raportu - w dniu spotkania
- Uzupełnienie dokumentacji o podane punkty - do 25 maja
- Prezentacja końcowa - 2 czerwca - wyniki prac, rozwinięcie 1 prezentacji.

10.7 Spotkanie z dnia 26.05.2020

Data i godzina spotkania: 09.06.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Dominik Guz - dominikguz1@gmail.com

Mlynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. Rozwinięcie:

- Recenzja obecnej wersji dokumentacji
- Prezentacja wyników analiz i dotychczasowych osiągnięć

Ustalenia:

- W całym dokumencie wprowadzić poprawki edytorskie np. Justowanie, poprawną numerację rysunków
- Uzupełnić dokumentację o wnioski końcowe dotyczące każdego rodzaju analiz
- Uzupełnić wnioski o rozważania dotyczące używanych technologii - dla osób które miałyby kontynuować projekt
- Dodać do działu z diagramem Gantta opis procesu wytwarzania, ewaluację - co było dobre, co moglibyśmy poprawić, co zrealizowaliśmy - ustosunkować się do tego wszystkiego
- Odnieść się do oceny ryzyka - co się sprawdziło, co nie
- Dodać ocenę użyteczności, jak można by rozwijać projekt
- Pod każdym rysunkiem dodać wnioski odnoszące się do tego rysunku.

Deadline:

- wysłanie raportu - w dniu spotkania
- Uzupełnienie dokumentacji o podane punkty przed końcowymi zajęciami- do 13 czerwca

10.8 Spotkanie z dnia 09.06.2020

Data i godzina spotkania: 09.06.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Dominik Guz - dominikguz1@gmail.com

Młynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. Rozwinięcie:

- Recenzja obecnej wersji dokumentacji
- Prezentacja wyników analiz i dotychczasowych osiągnięć

Ustalenia:

- W całym dokumencie wprowadzić poprawki edytorskie np. Justowanie, poprawną numerację rysunków
- Uzupełnić dokumentację o wnioski końcowe dotyczące każdego rodzaju analiz
- Uzupełnić wnioski o rozważania dotyczące używanych technologii - dla osób które miałyby kontynuować projekt
- Dodać do działu z diagramem Gantta opis procesu wytwarzania, ewaluację - co było dobre, co moglibyśmy poprawić, co zrealizowaliśmy - ustosunkować się do tego wszystkiego
- Odnieść się do oceny ryzyka - co się sprawdziło, co nie
- Dodać ocenę użyteczności, jak można by rozwijać projekt
- Pod każdym rysunkiem dodać wnioski odnoszące się do tego rysunku.

Deadline:

- wysłanie raportu - w dniu spotkania
- Uzupełnienie dokumentacji o podane punkty przed końcowymi zajęciami- do 13 czerwca

10.9 Spotkanie z dnia 16.06.2020

Data i godzina spotkania: 16.06.2020, godzina 11:15

Obecni członkowie Zespołu Projektowego (Zespół 1):

Dominik Guz - dominikguz1@gmail.com

Młynarczyk Dominika - mlynarczyk.dd@gmail.com

Nabywaniec Mateusz - mateusz.nabywaniec@gmail.com

Niedziela Grzegorz - gregxsunday@gmail.com

Prowadzący:

Dr Małgorzata Żabińska-Rakoczy - prowadząca przedmiot

Dr Jarosław Koźlak

Lista dyskutowanych tematów/ problemów, ew. Rozwinięcie:

- Recenzja obecnej wersji dokumentacji
- Prezentacja aplikacji

Ustalenia:

- Do dokumentacji wstawić ustalenia ze wszystkich raportów w kolejności
- Umieścić repozytorium i dokumentację na UPEL zgodnie z tym co jest tam napisane
- Rozdzielić rozdział z wynikami na 3 rozdziały lub podrozdziały dotyczące każdego z modułów analizy
- Każdy rozdział powinien mieć wstęp - opis czego dotyczy
- Poprawić opis architektury na rys. 4 - ma być napisane że dotyczy danej części systemu z rys. 4 i opis ma być bardziej jednolity
- Sprawdzić, czy jest wyraźnie napisane, że w aplikacji nie da się dynamicznie tworzyć wykresów i czy jest wyjaśnione dobrze
- Dodać informację o kierunkach rozwoju np. O tworzeniu dynamicznych wykresów
- Napisać co byśmy zmienili, gdybyśmy jeszcze raz zaczynali ten projekt

Deadline:

- wysłanie raportu - w dniu spotkania
- Uzupełnienie dokumentacji o podane punkty przed końcowymi zajęciami- do 20 czerwca

11. Źródła

1. Anderson Martin - "Choosing a Python Library for Sentiment Analysis"
<https://www.iflexion.com/blog/sentiment-analysis-python>
2. Barber Lionel, Sevastopulo Demetri and Tett Gillian- "Donald Trump: Without Twitter, I would not be here — FT interview"
<https://www.ft.com/content/943e322a-178a-11e7-9c35-0dd2cb31823a> - wywiad Donalda Trumpa dla Financial Times
3. Czech Mateusz - "Co to jest analiza sentymentu oraz jak możesz ją wykorzystać?"
<https://brand24.pl/blog/co-to-jest-analiza-sentymentu-oraz-jak-mozesz-ja-wykorzystac/> - wstęp do analizy sentymentu
4. Dokumentacja twitter API - <https://developer.twitter.com/en/docs>
5. Dokumentacja NetworkX - <https://networkx.github.io/documentation/stable/>
6. "Dziennikarze na Twitterze (TOP20)"
<https://www.wirtualnemedia.pl/artykul/dziennikarze-na-twitterze-weglarczyk-kolton-i-sekielski-stracili-3-proc-obserwujacych-gmyz-i-janecki-zyskali-7-8-proc-top20> - ranking popularności polskich dziennikarzy na twitterze za rok 2018
7. Fromm Jennifer, Melzer Stefanie, Ross Björn and Stieglitz Stefan - "Trump Versus Clinton: Twitter Communication During the US Primaries."
https://www.researchgate.net/publication/324081259_Trump_Versus_Clinton_Twitter_Communication_During_the_US_Primaries - analiza wyborów prezydenckich USA na twitterze
8. Inzaugarat Euge - "Visualizing Twitter interactions with NetworkX"
<https://medium.com/future-vision/visualizing-twitter-interactions-with-networkx-a391da239af5> - jak wykorzystać networkx w SNA twittera
9. Maksymowicz Piotr - "Twitter - co to jest? Jak Twitter radzi sobie w Polsce?"
<https://www.whitepress.pl/baza-wiedzy/245/twitter-co-to-jest-jak-twitter-radzi-sobie-w-polsce> - popularność Twittera w Polsce
10. Roesslein Joshua - Tweepy Documentation
<http://docs.tweepy.org/en/latest/> - dokumentacja biblioteki tweepy
11. Sarata Joanna - "SNA, czyli sieć jako obiekt analizy"
<https://predictivesolutions.pl/sna-czyli-siec-jako-obiekt-analizy>
12. Strona główna projektu Morfeusz 2 służącego do lematyzacji słów -
<http://morfeusz.sjgp.pl>
13. Strona główna biblioteki NLTK - <https://www.nltk.org>
14. Strona główna projektu Wordcloud na githubie -
https://github.com/amueller/word_cloud

15. Social network analysis -

https://en.wikipedia.org/wiki/Social_network_analysis#Social_Media_Internet_Applications