Politechnika Warszawska

WYDZIAŁ ELEKTRYCZNY

Instytut Elektrotechniki Teoretycznej i Systemów Informacyjno-Pomiarowych
Zakład Elektrotechniki Teoretycznej i Informatyki Stosowanej

# Praca dyplomowa magisterska

na kierunku Informatyka
w specjalności Inżynieria oprogramowania

## Analiza wiadomości z serwisów społecznościowych na użytek rozpoznawania nastrojów.

Jakub Rzepliński
Numer albumu: 233608

promotor
dr inż. Marcin Kołodziej

Warszawa, 2018

**Streszczenie**

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.
**Słowa kluczowe:** jakieś, słowa, kluczowe, po polsku

**Abstract**

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.
**Keywords:** some, keywords, in, english

## Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że przedstawiona praca dyplomowa:

– została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami,

– nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego lub stopnia naukowego w wyższej uczelni

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

..... .....
data podpis autora (autorów) pracy

## Oświadczenie

Wyrażam zgodę / nie wyrażam zgody*1 na udostępnianie osobom zainteresowanym mojej pracy dyplomowej. Praca może być udostępniana w pomieszczeniach biblioteki wydziałowej. Zgoda na udostępnienie pracy dyplomowej nie oznacza wyrażenia zgody na jej kopiowanie w całości lub w części.
Brak zgody nie oznacza ograniczenia dostępu do pracy dyplomowej osób:

– reprezentujących władze Politechniki Warszawskiej,

– członków Komisji Akredytacyjnych,

– funkcjonariuszy służb państwowych i innych osób uprawnionych, na mocy odpowiednich przepisów prawnych obowiązujących na terenie Rzeczypospolitej Polskiej,

do swobodnego dostępu do materiałów chronionych międzynarodowymi przepisami o prawach autorskich. Brak zgody nie wyklucza także kontroli tekstu pracy dyplomowej w systemie antyplagiatowym.

..... .....
data podpis autora (autorów) pracy

*1 - niepotrzebne skreślić

# Spis treści

# Rozdział 1

# Wstęp

Witaj, to szablon pracy dyplomowej **częściowo** przystosowany do nowych wymagań edycyjnych PW[1]. Zapoznaj się z kodem źródłowym przed przystąpieniem do pracy. Na początek wymienię kilka ważnych uwag:

- ustawienia dokumentu znajdują się w pliku `config.tex`,
- niestety ze względu na użycie płatnych fontów w stronie tytułowej aktualna wersja używa plików .png[2],
- bibliografia używa stylu zbliżonego do stylu harwardzkiego(**ogata2010modern**), a cytowania są w większości zgodne z zaleceniami BG PW (**linh2002line**); można to zmienić w pliku konfiguracyjnym(**DUMMY:1**) na cytowania numeryczne,
- do kompilacji polecam TeXstudio 2.10.8 + MiKTeX w sekwencji pdflatex + biber + pdflatex + pdflatex[3] (czasem warto powtórzyć 2-3 razy, by zaktualizowała się bibliografia); do podglądu zaś Sumatra PDF,
- do zarządzania bibliografią polecam oprogramowanie Zotero i opcja Better BibTeX.

Pozdrawiam, Dominik Roszkowski.

## 1.1 Motywacja do pracy

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed

[1] Niektóre elementy z Zarządzenia Rektora nr 43/2016 są bardzo trudne do zrealizowania w LaTeX-u.

[2] Jeśli uda się dostać nagłówek i rodzaj pracy w krzywych, to zostaną one podmienione.

[3] By dodać w TeXstudio przejdź do Opcje > Build > User commands i dodaj: `txs:///pdflatex | txs:///biber | txs:///pdflatex | txs:///pdflatex`


**Rysunek 1.1:** Długi podpis grafiki opisujący na przykład, co znajduje się na niej po lewej a co po prawej stronie rozciągający się na więcej niż jedną linijkę

**Tablica 1.1:** My caption of this table

| L.p. | $\Re\{\underline{\mathfrak{X}}(m)\}$ | $-\Im\{\underline{\mathfrak{X}}(m)\}$ | $\mathfrak{X}(m)$ | $\frac{\mathfrak{X}(m)}{23}$ | $A_m$ | $\varphi(m)$ / ° | $\varphi_m$ / ° |
|---|---|---|---|---|---|---|---|
| 1 | 16.128 | 8.872 | 16.128 | 1.402 | 1.373 | −146.6 | −137.6 |
| 2 | 1.29 | 0.099 | 1.29 | 0.112 | 0.097 | −175.6 | −114.7 |
| 3 | 16.128 | 8.872 | 16.128 | 1.402 | 1.373 | −146.6 | −137.6 |
| 4 | 1.29 | 0.099 | 1.29 | 0.112 | 0.097 | −175.6 | −114.7 |
| 5 | 16.128 | 8.872 | 16.128 | 1.402 | 1.373 | −146.6 | −137.6 |
| 6 | 1.29 | 0.099 | 1.29 | 0.112 | 0.097 | −175.6 | −114.7 |
| 7 | 0.641 | −0.466 | 0.641 | 0.056 | 0.045 | 133.3 | −106.3 |

accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.


**(a)** A subfigure 1


**(b)** A subfigure 2

**Rysunek 1.2:** A figure with two subfigures that you can reference easily e.g. in Figure 1.2a and Figure 1.2b

### 1.1.1 Przykłady innych prac

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

## 1.2 Metodyka pracy

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipi-

scing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

## 1.3 Plan pracy

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

# Rozdział 2

# Przegląd technologii

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

## 2.1 Metodyka przeglądu

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

## 2.2 Wyniki przeglądu

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

## 2.3 Kolejna sekcja

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc

eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

# Rozdział 3

# Serwis społecznościowy Twitter

Serwis społecznościowy Twitter jest globalnym serwisem internetowym służącym głównie do zamieszczania wiadomości tzw. *tweet*, które użytkownicy tego serwisu mogą także czytać, komentować lub przekazywać dalej. Od kilku lat Twitter jest serwisem gdzie dochodzi do wymiany zdań na różny temat, dotyczących np. polityki, sportu, produktów, wydarzeń społecznych, a profile posiada wiele osób znanych publicznie oraz instytucji.

## 3.1 Historia

Serwis ten, nazywany SMS internetu, został założony w 2006 r. w Stanach Zjednoczonych przez Jacka Dorseya, Ev Williamsa, Noah Glassa oraz Biza Stone'a i od początku powstania sukcesywnie zwiększał swoją popularność poprzez wzrost liczby użytkowników odwiedzających jego witrynę oraz wysyłających wiadomości. W 2012 r. osiągnął ponad 100 milionów użytkowników, którzy zamieszczali łącznie ponad 340 milionów wiadomości dziennie oraz obsługiwał średnio około 1.6 miliarda wyszukujących zapytań dziennie. W 2013 r. Twitter stał się jedną z najczęściej odwiedzanych stron w całym internecie. W tym samym roku inżynierowie Twittera podali informację, że serwis ten obsługuje ok. 143 tys. wiadomości na sekundę. Na początku 2016 r. serwis ten posiadał ponad 319 milionów użytkowników aktywnych podczas każdego miesiąca. Od listopada 2013 r. akcje Twittera są obecne na nowojorskiej giełdzie.

## 3.2 Tweet

Tweet, czyli krótka wiadomość tekstowa, była początkowo ograniczona do 140 znaków, ale limit ten został podwojony w 2017 r. dla wszystkich języków oprócz chińskiego, japońskiego i koreańskiego. Użytkownicy mają możliwość wyróżniania wybranych przez siebie tematów przez dodanie do nich znaku '#', co czyni takie wyrażenie tagiem. Inną możliwością oferowaną przez Twittera jest odpowiadanie innym użytkownikom lub zamieszczenie referencji do nich przez dodanie znaku '@' poprzedzającego nazwę profilu innej osoby.

## 3.3 Architektura

Serwis społecznościowy Twitter opierał się początkowo o typową architekturę trójwarstwową składająca się z warstwy prezentacji, logiki biznesowej oraz warstwy danych. Do napisania tej aplikacji został użyty framework Ruby on Rails wykorzystujący język Ruby, a warstwa bazodanowa opierała się o technologię MySQL. Jednak wraz ze wzrostem ilości przetwarzanych danych inżynierowie Twittera podjęli decyzję w 2011 r. o zmianie technologii na język Scala, który działa na maszynie wirtualnej Javy oraz zrezygnowano z dotychczasowej architektury na

rzecz budowy rozproszonych serwisów komunikujących się między sobą. Wraz z przeprowadzonymi zmianami zanotowano ponad 10-krotne polepszenie obsługi tweetów.

## 3.4 API

Twitter jest platformą otwartą i udostępnia programowalny interfejs API w dwóch postaciach: Search API oraz Streaming API.

### 3.4.1 Zakres działania

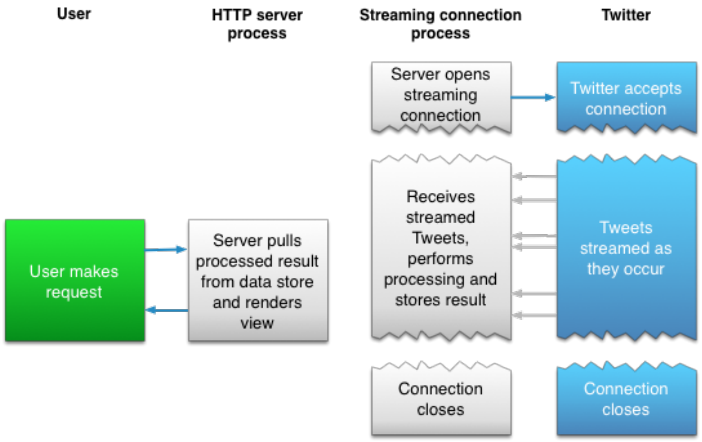
Programiści korzystający z Search API są w stanie uzyskać dostęp tylko do danych historycznych, które zostały już wcześniej zamieszczone na łamach serwisu Twitter. Natomiast w przypadku Streaming API dostajemy możliwość śledzenia strumienia danych, które są do naszego dostępu nawet już kilka sekund po zamieszczeniu w serwisie Twitter. Po podłączeniu do takiego strumienia możemy cały czas obserwować nowe wiadomości. Przyjęło się stosować nazywnictwo, że analiza Search API to analiza *back in time*, a Streaming API to śledzenie *real time*.

### 3.4.2 Rejestracja

Obie formy API wymagają wcześniejszej rejestracji na stronie *https://developer.twitter.com/en/apply-for-access* przeznaczonej dla deweloperów zainteresowanych wykorzystywaniem Twitter API. Po przejściu pomyślnej rejestracji dostajemy dane, które po nawiązaniu połączenia z serwisem Twitter umożliwiają mu jednoznacznie określić, że możemy uzyskać dostęp do API.

### 3.4.3 Sposób działania

Search API powstało z wykorzystaniem standardu REST - *Representational State Transfer*. Oba rodzaje API wykorzystują protokół HTTP: do poprawnego działania Streaming API potrzebne jest ciągłe połączenie HTTP, a w przypadku drugiego z nich każda operacja jest wykonywana przy nawiązaniu oddzielnego połaczenia.




**Rysunek 3.1:** Schemat działania dwóch rodzajów programistycznego interfejsu API udostępnianego przez serwis społecznościowy Twitter: Search API i Streaming API.

Search API posiada ściśle określone paramtery, które mogą być przesłane w żądaniu. W tabeli 3.1 zaprezentowano ich wykaz.

**Tablica 3.1:** Parametry żądania Twitter Search API

| Parametr | Wymagany/Opcjonalny | Opis | Przykład |
|---|---|---|---|
| q | wymagany | zapytanie wyszukujące o maksymalnej długości 500 znaków | nasa |
| geocode | opcjonalny | zwraca wiadomości użytkowników oddalonych o podany promień od podanej szerokości i długości geograficznej, promień może być podany w milach lub kilometrach | 37.781157 -122.398720 1mi |
| lang | opcjonalny | ogranicza wiadomości do wybranego języka spośród dostępnych kodów ISO 639-1 | pl |
| locale | opcjonalny | specyfikuje język wysyłanego zapytania, obecnie tylko *j*a jest skuteczny | ja |
| result_type | opcjonalny | określa typ zwracanych wiadomości, obecnie dostępne są trzy wartości tego parametru: *recent* (zwracane są najnowsze wiadomości), *popular* (zwracane są najbardziej popularne wiadomości) *mixed* (wartość domyślna, zwracane wyniki obejmują najnowsze i najbardziej popularne wiadomości) | mixed |
| count | opcjonalny | specyfikuje ilość zwracanych wiadomości; maksymalna wartość to 100, a domyślna to 15 | 100 |

| | | | |
|---|---|---|---|
| until | opcjonalny | ten parametr odpowiada za zwracanie wiadomości, których data utworzenia jest starsza o maksymalnie tydzień niż podana; obowiązuje format YYYY-MM-DD | 2015-07-19 |
| since_id | opcjonalny | dzięki temu parametrowi zwracane są wiadomości o ID większym niż podane czyli nowsze wiadomości niż określono | 12345 |
| max_id | opcjonalny | dzięki temu parametrowi zwracane są wiadomości o ID mnijeszym lub równym niż podane czyli starsze lub takie same wiadomości niż określono | 54321 |

Streaming API nie posiada takich ograniczeń. W języku programowania Java dostępny jest pakiet *twitter4j* zawierający interfejsy *User* oraz *Status*, na które mapowane są przychodzące ze strumienia informacje. W tabeli 3.2 zamieszczam ich dokumentację.

### 3.4.4 Ograniczenia

Korzystając z Search API mamy możliwość wysłania 720 zapytań na godzinę, a maksymalna ilość wiadomości jaka może być zwrócona na jedno zapytanie to 100. Jeśli wykorzystalibyśmy ten limit w maksymalny sposób to daje nam to 72 000 wiadomości na godzinę. W przypadku Streaming API głównym ograniczeniem jest dostęp do ok. 1 % danych ze strumienia, a maksymalna ilość wiadomości w czasie jednej minuty to 3 000. W przypadku tego API w ciągu godziny możemy uzyskać 180 000 wiadomości na godzinę. Są to ograniczenia, które obowiązują dla rozwiązań typu *open-source*.

## 3.5 Dostępne narzędzia analityczne

Udostępnienie API przez serwis społecznościowy Twitter oraz rosnące znaczenie danych generowanych przez użytkowników tego serwisu spowodowało, że wiele firm oraz instytucji zaczęło przywiązywać dużą wagę do analizy opinii wyrażanych na swój temat lub na tematy pokrewne, zainteresowania pewnymi tematami oraz kształtujących się trendów. Dlatego powstały aplikacje internetowe służące do wyświetlania takich informacji i przeprowadzające wstępną analizę zebranych danych. W dalszej części pracy znajduje się omówienie najważniejszych z nich.

### 3.5.1 Twitter Analytics

Pierwszym z narzędzi, którym warto poświęcić uwagę jest *Twitter Analytics*. Aplikacja ta posiada trzy zakładki. Na pierwszej z nich wyświetla statystyki dotyczące profilu z serwisu Twitter, którym logujemy się do niej: naszą najbardziej popularną wiadomość, najpopularniejszą wzmiankę o naszym profilu oraz wykresy trendów m. in. liczby osób śledzących nasz profil i odwiedzin. Na kolejnej zakładce mamy możliwość tworzenia wiadomości, które zostaną wygenerowane na naszym profilu oraz dodania do nich plików graficznych, materiałów audio lub wideo. Na ostatniej stronie użytkownik ma szansę poznania informacji takich jak np. lokalizacja geograficzna i wiek osób śledzących nasz profil czyli tzw. *followers*.

### 3.5.2 Hootsuite

Kolejną aplikacją jest *Hootsuite*. Jest to narzędzie, które umożliwia prowadzić kampanie w kilku mediach społecznościowych na raz np. *Facebook*, *Instagram*, *Linkedin*. Posiada wiele rozbudowanych funkcji. Co ciekawe Hootsuite pozwala na analizę nastrojów społecznych dla wybranych

**Tablica 3.2:** Dokumentacja interfejsu Status

| Typ zwracany | Nazwa metody | Opis |
|---|---|---|
| long[] | getContributors() | |
| java.util.Date | getCreatedAt() | zwraca datę utworzenia wiadomości |
| long | getCurrentUserRetweetId() | zwraca id użytkownika, którego wiadomość została podana dalej |
| int | getDisplayTextRangeEnd() | |
| int | getDisplayTextRangeStart() | |
| int | getFavoriteCount() | zwraca informację ile razy została polubiona wiadomość |
| GeoLocation | getGeoLocation() | zwraca lokalizację użytkownika zamieszczającego tą wiadomość |
| long | getId() | zwraca id wiadomości |
| java.lang.String | getInReplyToScreenName() | zwraca nazwę użytkownika, do którego kierowana jest odpowiedź |
| long | getInReplyToStatusId() | zwraca id wiadomości, do którego kierowana jest odpowiedź |
| java.lang.String | getLang() | zwraca język zamieszczonej wiadomości |
| Place | getPlace() | zwraca obiekt Place przypisany do tej wiadomości |
| Status | getQuotedStatus() | zwraca obiekt Status cytowanej wiadomości |
| long | getQuotedStatusId() | zwraca id cytowanej wiadomości |
| URLEntity | getQuotedStatusPermalink() | zwraca obiekt URLEntity reprezentujący bezpośredni odnośnik do cytowanej wiadomości |
| int | getRetweetCount() | zwraca ile razy wiadomość została podana dalej |
| Status | getRetweetedStatus() | zwraca oryginalny status, który jest podany dalej w tej wiadomości |
| Scopes | getScopes() | zwraca obiekt typu Scopes posiadający informację o id miejsc, do których odnosi się ta wiadomość |
| java.lang.String | getSource() | zwraca źródło wiadomości |
| java.lang.String | getText() | zwraca tekst wiadomości |
| User | getUser() | zwraca obiekt typu User powiązany z tą wiadomością |
| java.lang.String[] | getWithheldInCountries() | zwraca tablicę nazw krajów, w których wiadomość została wstrzymana |

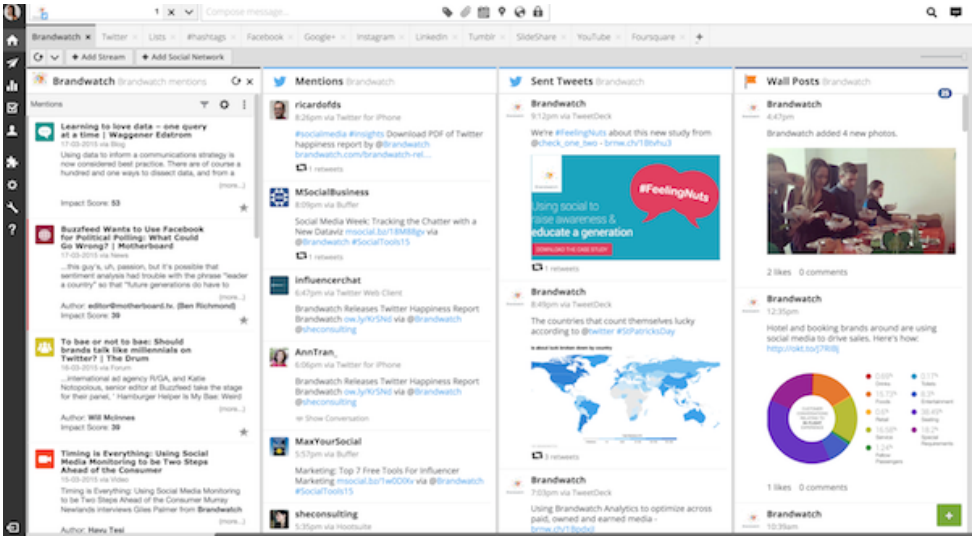
| | | |
|---|---|---|
| boolean | isFavorited() | zwraca informację czy wiadomość została polubiona |
| boolean | isPossiblySensitive() | zwraca informację czy wiadomość zawiera link do chronionych informacji |
| boolean | isRetweet() | zwraca informację czy tweet jest podaną dalej wiadomością |
| boolean | isRetweeted() | informuje czy wiadomość jest podana dalej |
| boolean | isRetweetedByMe() | informuje czy wiadomość jest podana dalej przez tego użytkownika |
| booleand | isTruncated() | informuje czy wiadomość jest skrócona (zakończona znakiem "...") |

**Tablica 3.3:** Dokumentacja interfejsu User z pakietu twitter4j. Zamieszczone zostały tylko najważniejsze z metod.

| Typ zwracany | Nazwa metody | Opis |
|---|---|---|
| java.util.Date | getCreatedAt() | zwraca datę utworzenia profilu użytkownika |
| java.lang.String | getDescription() | zwraca opis konta użytkownika |
| java.lang.String | getEmail() | zwraca adres e-mail powiązany z tym kontem |
| int | getFavouritesCount() | zwraca liczbę wiadomości, którą polubił ten użytkownik |
| int | getFollowersCount() | podaje ilość użytkowników śledzących profil |
| int | getFriendsCount() | podaje ilość śledzonych profili |
| long | getId() | zwraca id użytkownika |
| java.lang.String | getLang() | zwraca język preferowany przez użytkownika |
| java.lang.String | getLocation() | zwraca lokalizację użytkownika |
| java.lang.String | getName() | podaje nazwę użytkownika |
| java.lang.String | getScreenName() | zwraca nazwę konta |
| Status | getStatus() | zwraca obiekt typu Status reprezentujący wiadomość wysłaną przez użytkownika |
| int | getStatusesCount() | podaje ilość wiadomości wysłanych przez użytkownika |
| java.lang.String | getTimeZone() | podaje strefę czasową użytkownika |
| java.lang.String | getURL() | zwraca URL do profilu |
| boolean | isVerified() | podaje informację czy profil jest zweryfikowany |


**Rysunek 3.2:** Narzędzie Twitter Analytics.

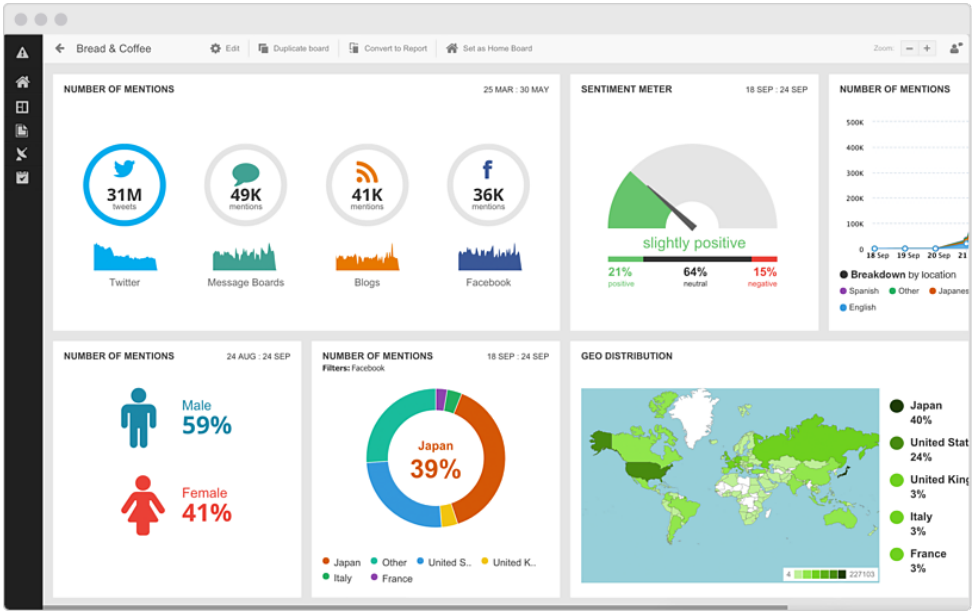
słów kluczowych. Wyświetlane informacje zawierają ogólny zarys użytkowników zamieszczających wiadomości na dany temat np. lokalizację geograficzną, podział ze względu na płeć i język oraz wykres trendu. Aplikacja ta wyświetla wszystkie wiadomości, w których użytkownicy używają wybranego słowa kluczowego. Jak podają twórcy tego narzędzia ma to główne zastosowanie jako pomoc w kampaniach marketingowych w dotarciu do użytkowników krytykujących produkt i posiadających największą ilością osób śledzących.




**Rysunek 3.3:** Narzędzie Hootsuite - główny pulpit.

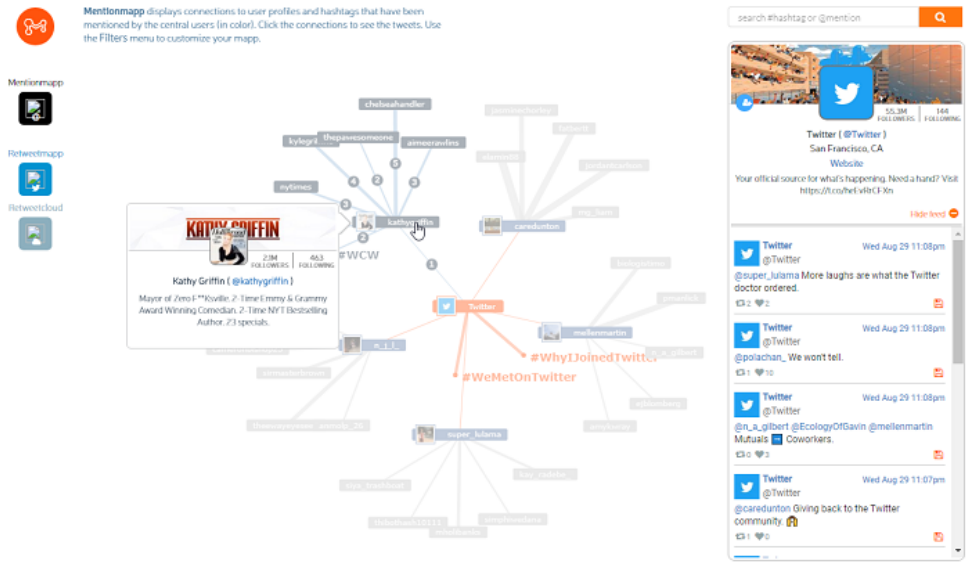
### 3.5.3 Mentionmap

Trzecim narzędziem wartym omówienia jest *Mentionmap*. Jest to aplikacja, która wyróżnia się spośród innych tym, że rysuje wykres powiązań pomiędzy użytkownikami wchodzącymi w interakcje z danym profilem, a także z ich profilami. Posiada także podstronę umożliwiającą badanie nastrojów społecznych osób zamieszczających wiadomości z konkretnym słowem klu-




**Rysunek 3.4:** Narzędzie Hootsuite - badanie nastrojów społecznych.

czowym. Nastroje te są przedstawione w postaci chmury nazw kont użytkowników, gdzie kolor nazwy zależy od nastoroju prezentowanego przez użytkownika pod kątem słowa kluczowego.




**Rysunek 3.5:** Narzędzie Mentionmap - wykres powiązań.

### 3.5.4 Podsumowanie

Podsumowując warto zauważyć, że nie ma obecnie na rynku aplikacji, która umożliwiałaby śledzenie występowania dowolnego słowa w wiadomościach zamieszczanych w czasie rzeczywistym w serwisie Twitter, rysowałaby wykres zależności pomiędzy użytkownikami tego serwisu, analizowałaby nastroje społeczne użytkowników z wyświetleniem informacji o nastroju wyrażanym w poszczególnych wiadomościach oraz pozwalałaby na analizowanie danych historycznych i zamieszczanych w czasie rzeczywistym. Taka sytuacja pozwala na utworzenie nowej aplikacji, która udostępniałaby te funkcjonalności.


**Rysunek 3.6:** Narzędzie Mentionmap - badanie nastrojów społecznych.

# Rozdział 4

# Analiza danych Big Data

Termin *Big Data* odnosi się do dużych, zmiennych i różnorodnych zbiorów danych, których przetwarzanie i analiza jest pracochłonne, ale może prowadzić do ciekawych wniosków oraz pozyskania nowej wiedzy. Zbieranie oraz przechowywanie dużej ilości danych do analizy było praktykowane od bardzo dawna, jednak dokładniejsza koncepcja Big Data została poznana w 2001 roku kiedy to analityk Doug Laney zaprezezentował znaną dzisiaj definicję *3V*: *volume*, *velocity*, *variety* czyli: ilość, szybkość, złożoność, a później dodano jeszcze czwarty atrybut *veracity* czyli: wiarygodność.

## 4.1 Pojęcie Big Data

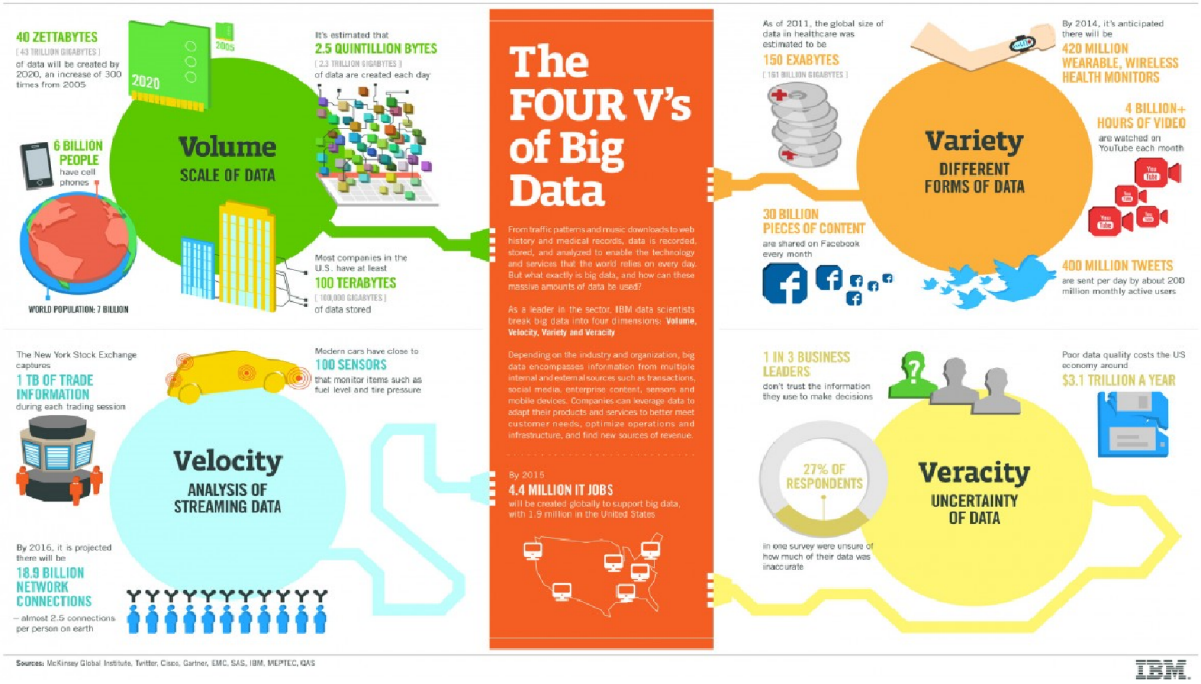
Wraz ze wzrostem zainteresowania Big Data podjęto próby dokładniejszego opisania tego terminu. Obecnie definiując to pojęcie trzeba odnieść się do nowych rozwiązań technologicznych dotyczących wielkich wolumenów danych o innym charakterze ilościowym oraz jakościowym niż dotychczas.
Jedna z pierwszych definicji Big Data została zaprezentowana przez M. Cox i D. Ellsworth w 1997 r. jako duża ilość danych, którą należy zwiększać, aby wydobyć wartości informacyjne. Inna i najbardziej popularna, została przestawiona w 2001 r. przez pracującego dla firmy analityczno-doradczej wspomnianego analityka D. Laney, opiera się o trzy atrybuty: ilość, szybkość i złożoność. W 2012 r. ta sama firma dodała do swojej definicji kolejne dwa atrybuty: zmienność i złożoność. Autorzy innej publikacji *Big Data: Issues, Challenges, Tools and Good Practices* z 2013 r. definiują pojęcie Big Data jako wymagające stosowania nowych technologii i architektur z powodu potrzeby ekstrakcji wartości płynącej z tych danych.
Podsumowując, określenie Big Data to pojęcie odnoszące się do zbiorów danych, które jednocześnie charakteryzują się dużą objętością, różnorodnością, strumieniowym napływem w czasie rzeczywistym, zmiennością, złożonością oraz wymagają stosowania innowacyjnych technologii i narzędzi, aby możliwe było wydobycie z nich wartościowych informacji.

## 4.2 Charakterystyka

Termin Big Data charakteryzują atrybuty: objętość, szybkość, różnorodność, zmienność, złożoność i wartość. W dalszej cześci tej pracy dyplomowej przedstawiono omówienie każdego z tych atrybutów.

### 4.2.1 Objętość

Atrybut ten odnosi się do dużej ilości danych, które wymagają nowych technologii. Rozmiar danych zależy od dziedziny i może wynosić od terabajtów lub petabajtów w zagadnieniach ta-




**Rysunek 4.1:** Definicja *Big Data* w ujęciu 4V.

kich jak np. analiza zderzeń cząstek elementarnych w fizyce do megabajtów lub gigabajtów np. w telekomunikacji przy analizie połączeń wykonywanych przez abonentów. Najnowsze badania prognozują, że ilość danych wzrośnie do 2020 r. o 40% zeta bajtów, co będzie skutkować 50-krotnym wzrostem od początku 2010 r..