Akademia Górniczo-Hutnicza

im. Stanisława Staszica w Krakowie

Wydział Informatyki, Elektroniki i Telekomunikacji

Katedra Informatyki



***Projekt dyplomowy***

*Analiza wpływu twittów na kurs akcji*

**Marek Grzyb**

**Opiekun Projektu: dr inż. Robert Marcjan**

**Kraków 2019**

Spis
treści

[1. Wstęp](#_Toc264_WPSOffice_Level1) [1](#_Toc264_WPSOffice_Level1)

[2. Notowania spółek](#_Toc16068_WPSOffice_Level1) [2](#_Toc16068_WPSOffice_Level1)

[3. Twitter](#_Toc98_WPSOffice_Level1) [3](#_Toc98_WPSOffice_Level1)

[3.1. Konto deweloperskie](#_Toc16068_WPSOffice_Level2) [3](#_Toc16068_WPSOffice_Level2)

[3.2. Autoryzacja](#_Toc98_WPSOffice_Level2) [4](#_Toc98_WPSOffice_Level2)

[3.3. Pobieranie twittów](#_Toc8055_WPSOffice_Level2) [4](#_Toc8055_WPSOffice_Level2)

[3.4. Struktura wiadomości](#_Toc8730_WPSOffice_Level2) [5](#_Toc8730_WPSOffice_Level2)

[4. Przetwarzanie języka naturalnego](#_Toc8055_WPSOffice_Level1) [6](#_Toc8055_WPSOffice_Level1)

[5. Analiza](#_Toc8730_WPSOffice_Level1) [7](#_Toc8730_WPSOffice_Level1)

[5.1. Studium przepadku: PKN ORLEN](#_Toc6010_WPSOffice_Level2) [7](#_Toc6010_WPSOffice_Level2)

[5.2. Studium przepadku: CD PROJECT](#_Toc10416_WPSOffice_Level2) [8](#_Toc10416_WPSOffice_Level2)

[5.3. Podsumowanie](#_Toc9596_WPSOffice_Level2) [8](#_Toc9596_WPSOffice_Level2)

[6. Referencje](#_Toc6010_WPSOffice_Level1) [9](#_Toc6010_WPSOffice_Level1)

# Wstęp

Celem pracy jest zbadanie jaki wpływ na decyzje inwestorów mają informacje zamieszczane w serwisach społecznościowych. W tym opracowaniu przedstawię związki między informacjami umieszczanymi na Twitterze a kursami akcji spółek notowanych na Giełdzie Papierów Wartościowych.

Podczas analizy wykorzystam informacje o kursie akcji pobrane za pośrednictwem API Bankier.pl oraz treść Twittów.

W pacy przeprowadzę analizę twittów, sprawdzę skorelowanie ich sentymentu z kursem akcji. Będę poszukiwał twittów które wyprzedzają zmianę kursu oraz osoby mające znaczny wpływ na kurs akcji.

# Notowania spółek

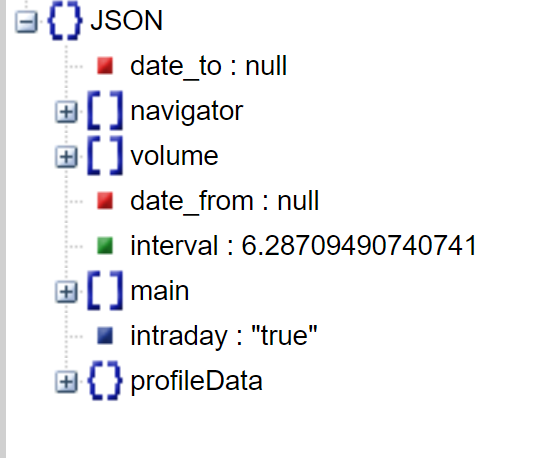
W przeprowadzonym eksperymencie wykorzystam informacje o kursach akcji spółek notowanych na Giełdzie Papierów Wartościowych uwzględnionych w indeksie WIG20.

Notowania spółek pobierane są ze strony bankier.pl. Do pobierania danych wykorzystywane jest zapytanie HTTP GET:

[https://www.bankier.pl/new-charts/get-data?symbol=](https://www.bankier.pl/new-charts/get-data?symbol={symbol}&intraday=true&type=area)**[{symbol}](https://www.bankier.pl/new-charts/get-data?symbol={symbol}&intraday=true&type=area)**[&intraday=true&type=area](https://www.bankier.pl/new-charts/get-data?symbol={symbol}&intraday=true&type=area)

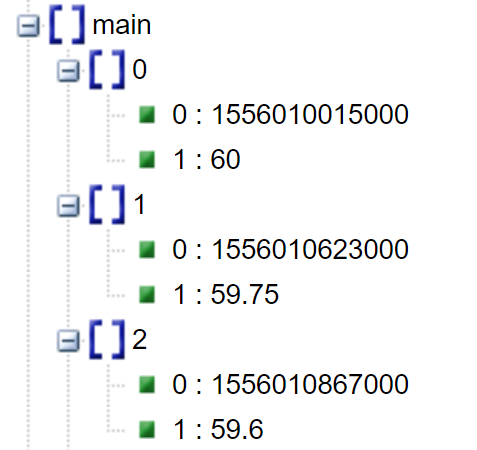
gdzie {symbol} oznacza symbol spółki giełdowej której notowania chcemy pozyskać.

Pobrane dane mają format JSON o strukturze zobrazowanej na Rys. 1.



Rys. 1. Struktura wiadomości JSON zwracanej prze API bankier.pl

Atrybut “main” zawiera kolekcje kursy spółki w postaci:



Rys. 2. Struktura JSON kursów akcji zwracana przez API bankier.pl

gdzie pierwsza wartość oznacza czas wyrażony w formacje POSIX (liczba sekund od roku 1970) dodatkowo pomnożona przez 1000, drugi to cena.

Pobrane dane wyrażają kurs spółki z interwałem 1 min w dni robocze od 9:00 do 17:00 (godziny pracy GPW).

# Twitter

Twitter udostępnia API dzięki któremu możemy w prosty sposób pobierać informacje o treściach zamieszczanych na portalu.

Wyszukiwanie dostępne jest w trzech wersjach:

* Standard

Wyszukiwanie jedynie na próbce twittów z ostatnich 7 dni.

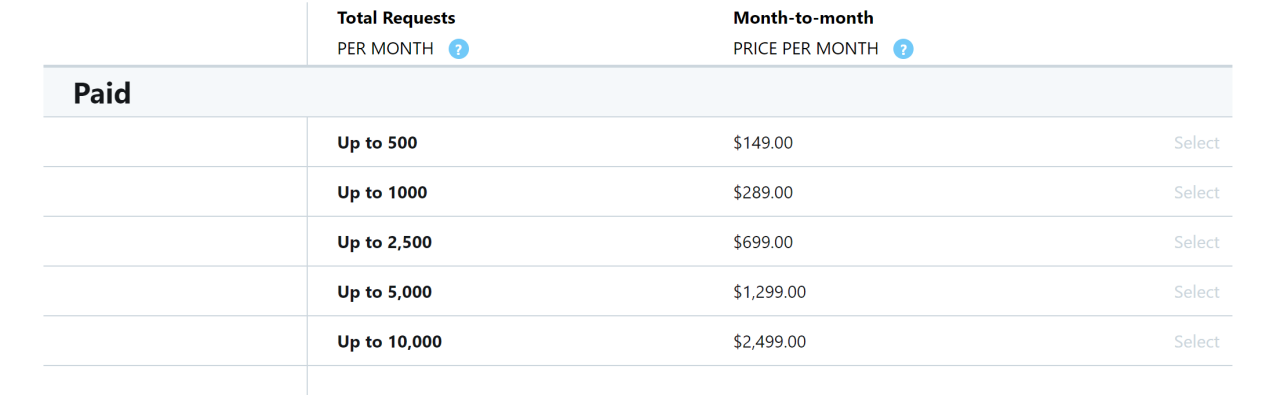
* Premium

Wyszukiwanie na pełnym zbiorze twittów. Darmowe konto ma jednak ograniczenia, 250 zapytam miesięcznie albo 1MB danych miesięcznie. Postaram się zmieścić w tym limicie dla celów eksperymentu.

* Enterprsie

Dostępne tylko po podpisana stosownej umowy.

Cennik dodatkowych wyszukiwani w opcji Premium został zamieszczony na Rys 3.



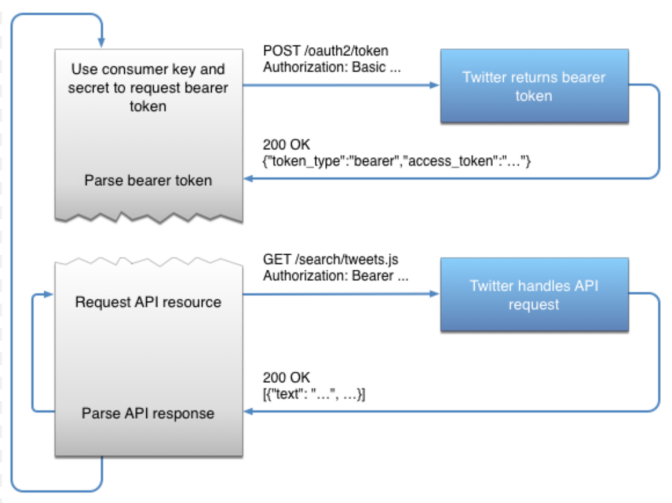
Rys. 3. Cennik usługi Premium Twittera

## Konto deweloperskie

Aby móc korzystać z API twitttera potrzebne jest konto deweloperskie, aby je założyć należy szczegółowo wyjaśnić w jakim cepu dane pochodzące z twistera będą wykorzystane.

## Autoryzacja

Dostęp do API wymaga autoryzacji protokołem OAuth 2.0, polega on na przesłaniu nazwy aplikacji oraz sekretnego klucza, w odpowiedzi otrzymujemy token który pozwala wykonywać udostępniane metody API.



Rys. 4. Diagram obrazujący sposób wymiany wiadomości podczas autoryzacji OAuth 2.0

## Pobieranie twittów

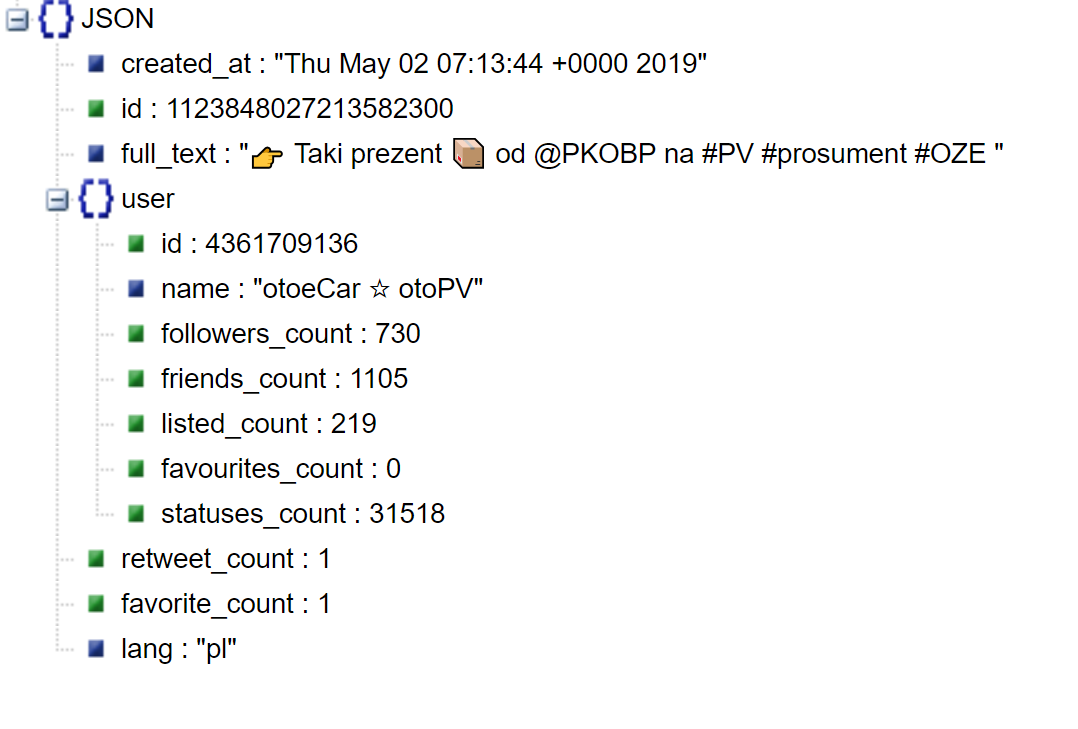
Wykorzystane API dostępne jest pod adresem: [https://api.twitter.com/1.1//tweets/search/30day/DEV.json](https://api.twitter.com/1.1/search/tweets.json)

Mechanizm komunikacji z tym API został zaimplementowany w wielu bibliotekach, jedna z nich to ‘tweepy’ (strona projektu <http://www.tweepy.org> ), która zostanie wykorzystana w tym opracowaniu.

## Struktura wiadomości

Pobrane dane mają format JSON. Z pośród kilkunastu dostępnych atrybutów do dalszej analizy wybierzemy tylko kilka najważniejszych:

* id - identyfikator
* created\_at – data utworzenia
* full\_text - pełny tekst wiadomości
* user.id, user.name - identyfikator i nazwa użytkownika
* user.followers\_count, user.friends\_count, user.listed\_count, user.favourites\_count – parametry wskazujące jak popularny i aktywny jest to użytkownik
* retweet\_count, favorite\_count – parametry wskazujące jak popularny jest to twitt



Rys. 5. Struktura wiadomości JSON opisująca Twitt

# Przetwarzanie języka naturalnego

Pierwszym etapie procesowania tekstu Twittów to jego podział na wyrazy (tokenizacja). Najprostszym sposobem podziału jest rozdzielenie tekstu po spacjach, strategia ta nie jest doskonała gdyż nie uwzględnia wyrazów rozdzielonych myślnikiem np.: czarno-biały. W tym opracowania tokenizacja została wykona przy użyciu biblioteki NLTK (strona projektu <http://www.nltk.org>).

Drugim etapem jest sprowadzenie rozpoznanych wyrazów do formy podstawowej np.: ropę -> ropa, w tym cepu wykorzystamy słownik fleksyjny pobrany ze strony <https://sjp.pl/slownik/odmiany/>, zawiera on 224 tysiące form bazowanych wyrazów. Dla każdej z form bazowych słownik zawiera formy odmienione np.: ropa, rop, ropach, ropami, ropą, ropę, ropie, ropo, ropom, ropy. Wycinek słownika fleksyjnego pokazuje Rys.6.

Z uwagi na fakt, że twitty mogą nie zawiera akcentów, przy dopasowaniu uwzględnimy mapowanie:

a - ą, c - ć , e - ę ,l - ł, n - ń, o - ó, s - ś, z - ż, z - ź

Tak przygotowane dane zostaną wykorzystane do dalszej analizy. W kolejnych krokach sprawdzimy powiazanie twittu ze spółką oraz ich wydźwięk (sentyment).



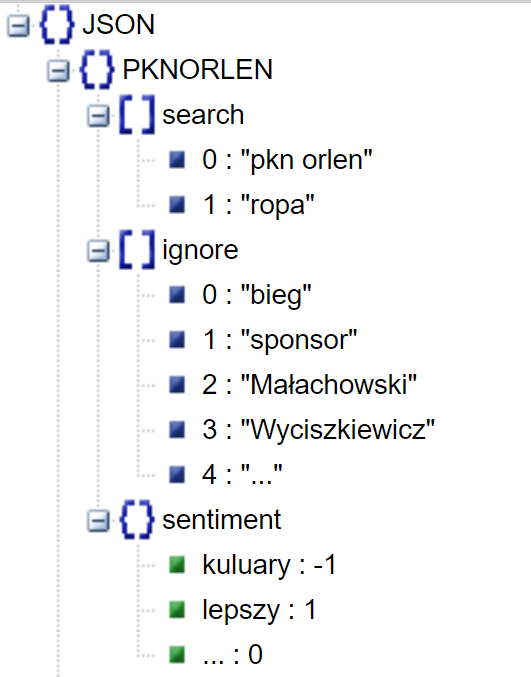
Rys. 6. Wycinek słownika fleksyjnego

# Analiza

W tym opracowaniu przedstawie szczegółową analizę dla kilku spółek notowanych na GPW. Wybrałem spółki z index-u WIG 20, są to jedne z bardziej znanych w swojej kategorii:

* PKN ORLEN - największa polska firma petrochemiczna
* CD PROJECT - najbardziej rozpoznawalny twórca gier komputerowych w Polsce, twórca kultowej gry Wiedźmin,
* GRUPA AZOTY - największy koncert chemiczny w Polsce
* LPP - polskie przedsiębiorstwo odzieżowe zajmujące się projektowaniem, produkcją i dystrybucją odzieży, mające w swoim portfolu mareki: Reserved, House, Cropp, Mohito i Sinsay
* mBank - czwarty co do wielkości bank w Polsce

Dla każdej przeanalizowanej spółki została przygotowana konfiguracja przedstawiona na Rys. 7.

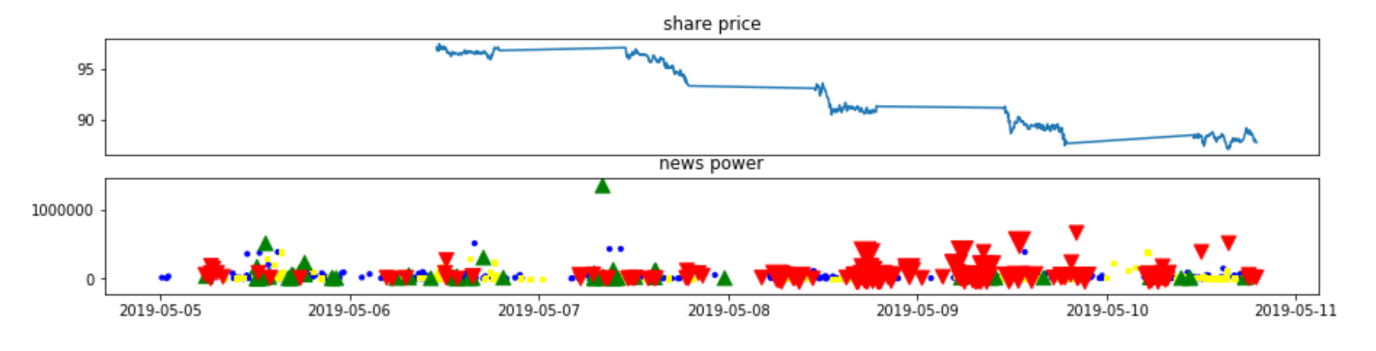


Rys. 7. Struktura konfiguracji dla jednej ze spółek

Gdzie

* search - słowa kluczowe wykorzystane przy zapytaniach o twitty
* ignore– słowa które powodują ignorowanie danego twittu, dla PKN ORLEN informacje o sportowcach sponsorowanych prze firmę
* sentiment - słowa kluczowe oraz wskażnik sentymentu z jakim one sie wiążą

Wyniki analizy przedstawie są na wykresie, przykładowy wykres na Rys. 8.



Rys. 8. Wyniki analizy

Górna część wykresu przedstawia cenę akcji w czasie dla analizowanej spółki, cena zmienia się tylko w godzinach 9-17, w dni robocze, jest to związane z godzinami pracy Giełdy Papierów Wartościowych. Wykres oparty jest o dane godzinowe.

Dolny wykres przedstawia twitty w czasie. Os X jest wspólna z wykresem kursu akcji. Os Y przedstawia moc oddziaływania twitta oraz osoby która go opublikowała. Kształty i korowy twittów oznaczają:

* niebieskie koło – twitt niesklasyfikowany
* żółty kwadrat - twitt zignorowany
* czerwony trójkąt - informacja negatywna
* zielony trójkąt - informacja pozytywna

## Studium przepadku: PKN ORLEN

Dla PKN ORLEN na twitterze znajdowało się sporo informacji o sportowcach sponsorowanych przez formę oraz informacji o zanieczyszczeniu rosyjskiej ropy.

## Studium przepadku: CD PROJECT

Dla CD PROJEKT twitter informował o otwarciu sklepu z gadżetami oraz o posterach w produkcji nowej gry. Brak jest negatywnych informacji, a mimo to kurs spółki nie wzrastał w sposób ciągły.

## Studium przepadku: GRUPA AZOTY

## Studium przepadku: LPP

## Studium przepadku: mBank

## Podsumowanie

Na podstawie przeanalizowanych przykładów nie możemy stwierdzić że twitter odgrywa kluczową role dla kursu cen akcji. Przytoczone informacje były również publikowane na wielu innych portalach. Dla żadnej z przeanalizowanej spółki nie odnalazłem kluczowej osoby, informacji która była by ‘breaking news’ opublikowana tylko na twitterze i powodująca gwałtowną reakcje giełdy.

# Referencje

1. <https://www.bankier.pl> - strona z informacjami o kursach akcji i nie tylko
2. <http://www.tweepy.org> - strona biblioteki implementującej komunikacje z twitterem
3. <https://github.com/PyConPL/Book/tree/master/2013/przetwarzanie_jezyka_naturalnego_w_praktyce> - praktyczne informajce o analizie tekstu w jezyku polskim
4. <http://www.nltk.org> - biblioteka do procesowania języka naturalnego
5. <https://github.com/agh-glk/pydic> - narządzie do sprowadzania wyrazów do formy podstawowej
6. <http://jsonviewer.stack.hu> - narzędzie do wizualizacji json
7. <https://sjp.pl/slownik/odmiany> - słownik fonetyczny dla języka polskiego