

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

Wstęp.....	1
Notowania spółek.....	2
Twitter.....	3
Konto deweloperskie.....	3
Autoryzacja.....	4
Pobieranie twittów.....	4
Struktura wiadomości.....	5
Przetwarzanie języka naturalnego.....	6
Analiza.....	7
Studium przypadku: PKN ORLEN.....	7
Studium przypadku: CD PROJECT.....	8
Podsumowanie.....	8
Referencje.....	9

## 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.

## Notowania spółek

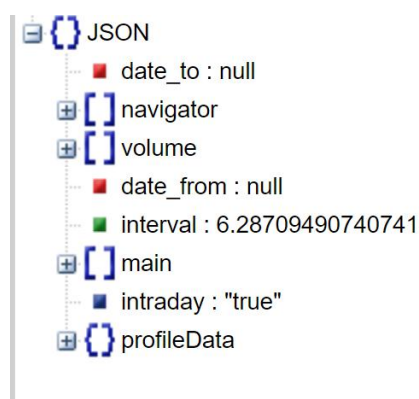
W przeprowadzonym eksperymencie wykorzystam informacje o kursach akcji spółek notowanych na Giełdzie Papierów Wartościowych uwzględnionych w indeksie WIG20.

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

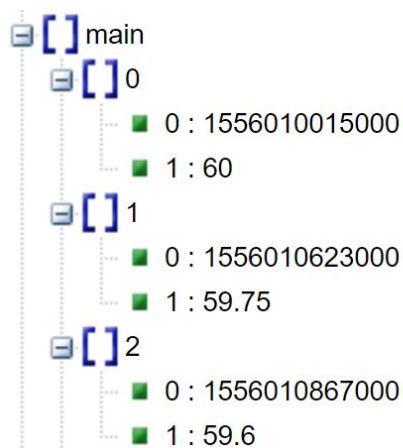
<https://www.bankier.pl/new-charts/get-data?symbol={symbol}&intraday=true&type=area>

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

Pobrane dane mają format JSON o strukturze:



Atrybut "main" zawiera kolekcję kursy spółki w postaci:



gdzie pierwsza wartość oznacza czas wyrażony w formacie 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 zapytań miesięcznie albo 1MB danych miesięcznie. Postaram się zmieścić w tym limicie dla celów eksperymentu.

- Enterprise

Dostępne tylko po podpisaniu stosownej umowy.

Cennik dodatkowych wyszukiwań w opcji Premium:

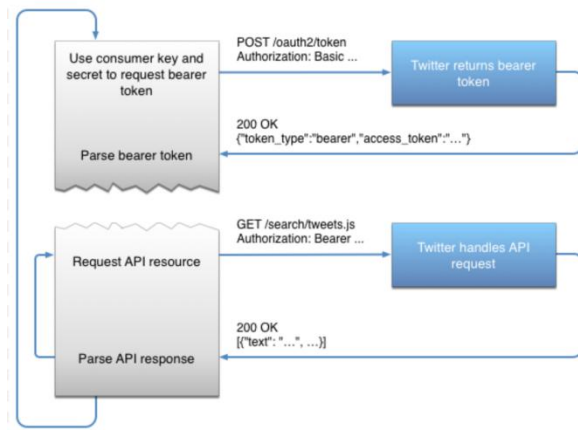
	Total Requests PER MONTH ?	Month-to-month PRICE PER MONTH ?	
<b>Paid</b>			
	Up to 500	\$149.00	Select
	Up to 1000	\$289.00	Select
	Up to 2,500	\$699.00	Select
	Up to 5,000	\$1,299.00	Select
	Up to 10,000	\$2,499.00	Select

### Konto deweloperskie

Aby móc korzystać z API twittera potrzebne jest konto deweloperskie, aby je założyć należy szczegółowo wyjaśnić w jakim celu 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.



## Pobieranie twittów

Wykorzystane API dostępne jest pod adresem:

<https://api.twitter.com/1.1/tweets/search/30day/DEV.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



## Przetwarzanie języka naturalnego

Pierwszym etapem 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 opracowaniu tokenizacja została wykonana przy użyciu biblioteki NLTK (strona projektu <http://www.nltk.org>).

Drugim etapem jest sprowadzenie rozpoznanych wyrazów do formy podstawowej np.: ropę -> ropa, w tym celu 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.

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

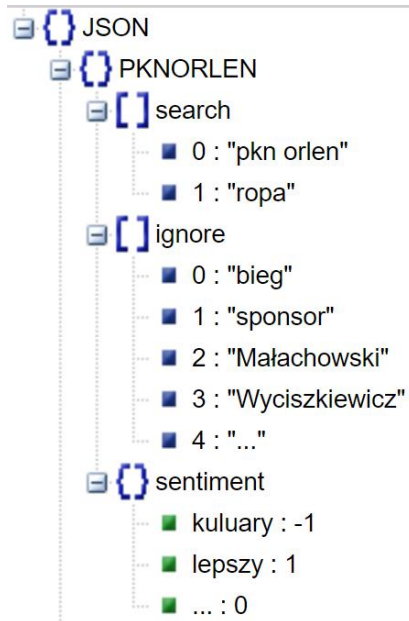
a - ą, c - ć, e - ę, l - ł, n - ń, o - ó, s - ś, z - ź, z - ż

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



## Analiza

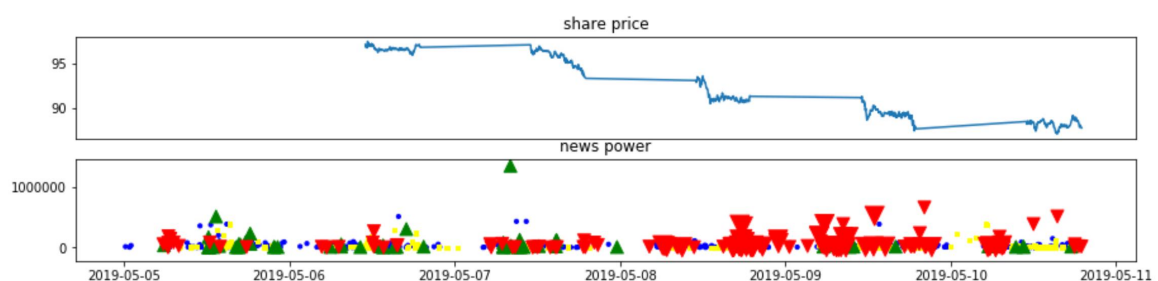
Szczegółową analizę wyniku przeprowadziłem dla kilkunastu firm, w tym opracowaniu przedstawię dwie przykładowe: PKN ORLEN i CD PROJECT. Dla każdej spółki została przygotowana konfiguracja:



Gdzie

- search - słowa kluczowe wykorzystane przy zapytaniach o twitty
- ignore— słowa które powodują ignorowanie danego twittu, dla PKN ORLEN informacje o sportowcach sponsorowanych przez firmę
- sentiment - słowa kluczowe oraz wskaźnik sentymentu z jakim one się wiążą

## Studium przypadku: PKN ORLEN



Pierwszy wykres przedstawia ceny akcji, GPW działa w godzinach 9-17, więc poza tymi godzinami cena się nie zmienia.

Dolny wykres przedstawia twitty, os Y to ich moc oddziaływania, kształty i kolory oznaczają:

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

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

### Studium przypadku: 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.

### Podsumowanie

Na podstawie przeanalizowanych przykładów nie możemy stwierdzić że twitter odgrywa kluczową rolę 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 reakcję giełdy.

## Referencje

1. <https://www.bankier.pl> - strona z informacjami o kursach akcji i nie tylko
2. <http://www.tweepy.org> - strona biblioteki implementującej komunikację z twitterem
3. [https://github.com/PyConPL/Book/tree/master/2013/przetwarzanie\\_jezyka\\_naturalnego\\_w\\_praktyce](https://github.com/PyConPL/Book/tree/master/2013/przetwarzanie_jezyka_naturalnego_w_praktyce) - praktyczne informacje o analizie tekstu w języku polskim
4. <http://www.nltk.org> - biblioteka do procesowania języka naturalnego
5. <https://github.com/agh-gl/kydic> - narzędzie 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