Akademia Górniczo-Hutnicza

im. Stanisława Staszica w Krakowie

Wydział Informatyki, Elektroniki i Telekomunikacji

Katedra Informatyki

Projekt dyplomowy

Analiza wpływu twittów na kurs akcji spółek notowanych na GPW

Marek Grzyb

Opiekun Projektu: dr inż. Robert Marcjan

Kraków 2019

Spis treści

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

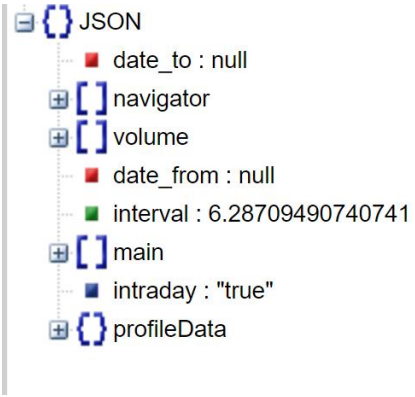
W przeprowadzonym eksperymencie wykorzystam informacje o kursach akcji spółek notowanych na Giełdzie Papierów Wartościowych uwzględnionych w indeksie WIG20.

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

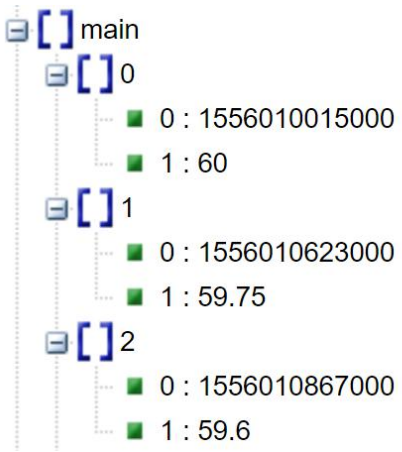
https://www.bankier.pl/new-charts/get-data?symbol=**{symbol}**&intraday=true&type=area

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

Pobrane dane mają format JSON o strukturze:



Atrybut “main” zawiera kolekcje kursy spółki w postaci:



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:

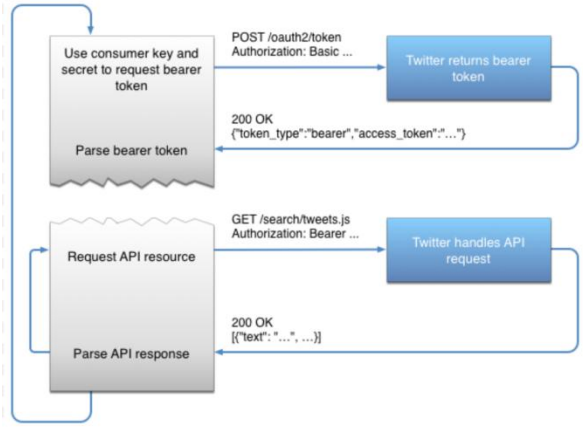
	Total Requests PER MONTH	Month-to-month PRICE PER MONTH	
Paid			
	Up to 500	$149.00	Select
	Up to 1000	$289.00	Select
	Up to 2,500	$699.00	Select
	Up to 5,000	$1,299.00	Select
	Up to 10,000	$2,499.00	Select

Konto deweloperskie

Aby móc korzystać z API 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.



Pobieranie twittów

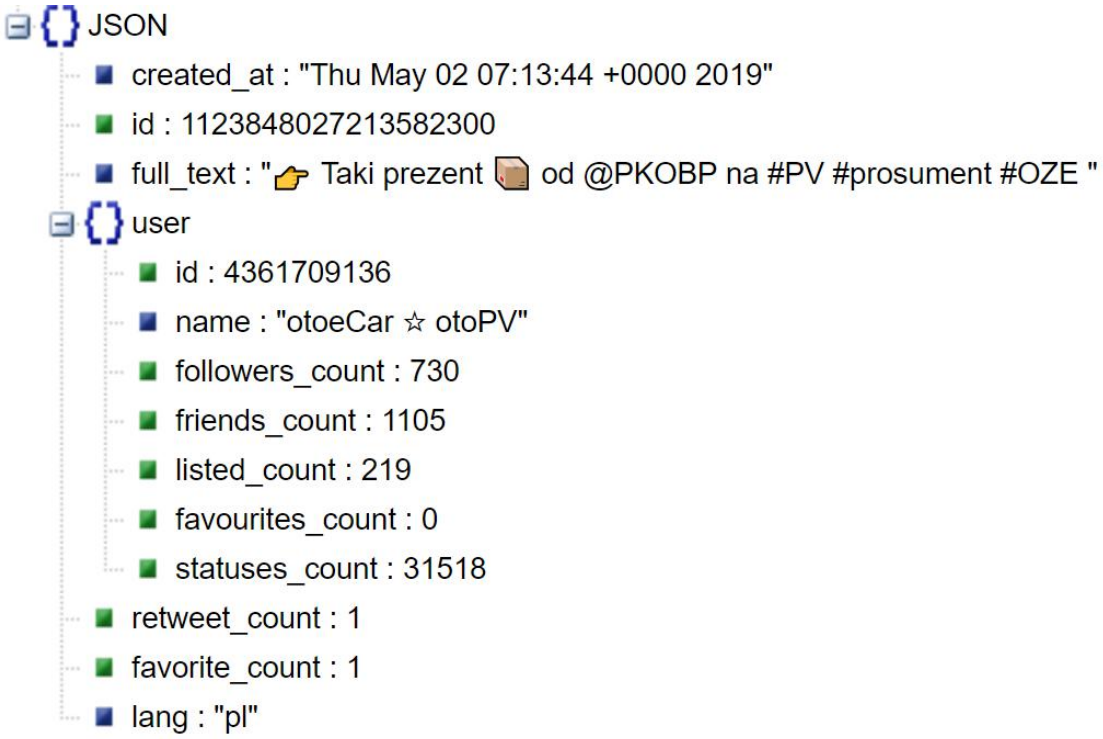
Wykorzystane API dostępne jest pod adresem:
https://api.twitter.com/1.1//tweets/search/30day/DEV.json

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 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.

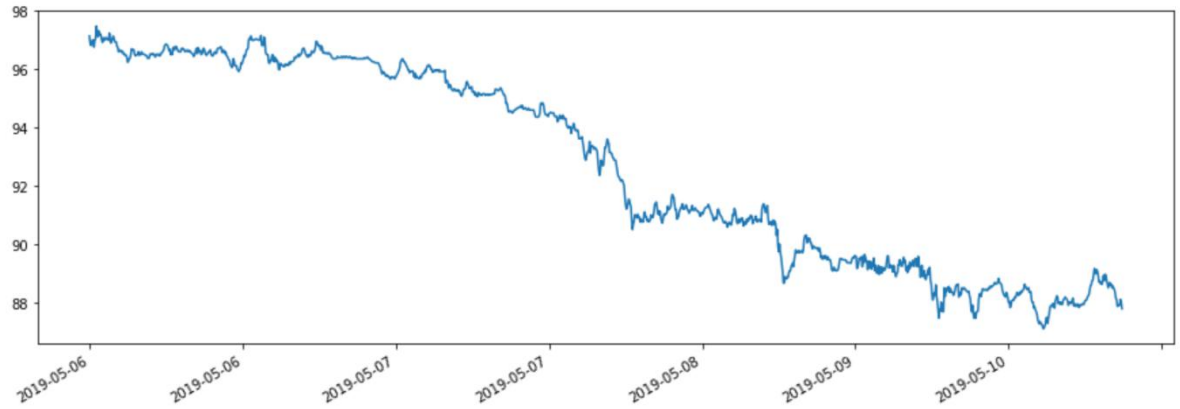
Z uwagi na fakt, że twitty mogą nie zawiera akcentów, przy dopasowaniu uwzględnimy mapowanie:

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).

Analiza

Wykonałem analizę dla spółce PKN Orlen, dane 2019-05-06 do 2019-05-12. Wykres akcji przedstawia się następująco:



Konfiguracje użyta do analizy twittów:

```
KeyWords= { 'PKNORLEN' : {
                'search': ['pkn orlen','ropa'] ,
                'ignore':['Małachowski','Kszczot','Lisek','Kubica','Dąbrowskiego','RobertKubicaKlub','William
                'sentiment':{'spada':1,'rośnie':-1 }
                }
        }
```

- search - słowa kluczowe stosowane do wyszukiwania twittów, w przpadku PKN Orlen jest to nazwa spółki i słowo ropa
- ignore - słowa kluczowe które wykluczaja twitt z dalszej analizy, np.: zawodnicy Orlen Team (lekkoatletyka), Kubica (F1)
- sentiment - słowa kluczowe oraz ich wpływ na współczynik sentymentu dla danego twitta

Wyniki sentymentu twittów a kurs akcji:

Przygotowany program

Referencje

1. https://www.bankier.pl
2. http://www.tweepy.org
3. https://github.com/PyConPL/Book/blob/master/2013/przetwarzanie_jezyka_naturalnego_w_praktyce/text.md
4. http://www.nltk.org
5. https://github.com/agh-glk/pydic
6. http://jsonviewer.stack.hu
7. https://sjp.pl/slownik/odmiany