Eksploracja danych

Grzegorz Najderek

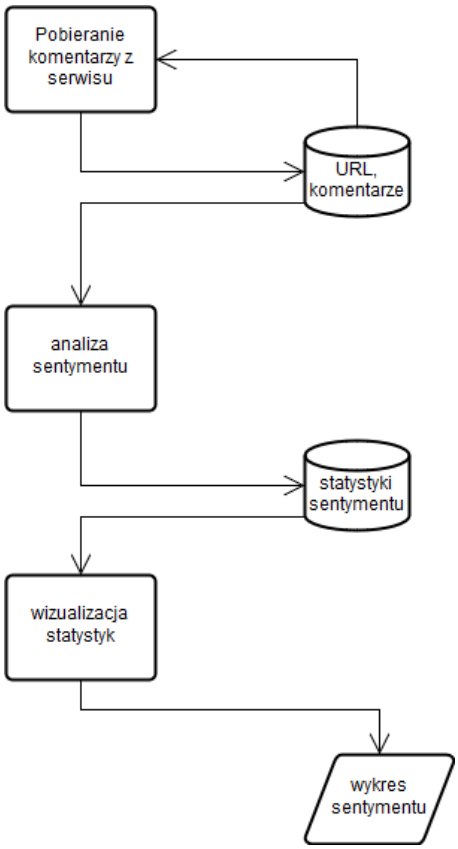
Analiza wyborów prezydenckich USA na przykładzie wybranego portalu (np. HuffingtonPost, Politico)

1. Cel projektu

Celem projektu była próba zbadania nastrojów panujących wśród wyborców na podstawie ich wypowiedzi w komentarzach w serwisach politycznych.

2. Architektura

Projekt zrealizowany został w języku Python. Funkcjonalnie został podzielony na 3 niezależne części zgodnie ze schematem:



Pierwsza baza danych zawierała 3 zbiory danych potrzebnych parserowi: listę **odwiedzonych** adresów, na bieżąco uzupełnianą listę adresów **do odwiedzenia**, oraz prostą strukturę przechowującą **datę oraz treść** komentarza, która była wykorzystywana na dalszym etapie programu.

Druga baza zawierała prostą strukturę przechowującą **datę** oraz wyliczoną **wartość sentymentu** (liczba) przypadającą na dany dzień. Dwa zbiory danych, po jednym na kandydata.

Trzecia przechowywana struktura to gotowy, wygenerowany wykres zmian sentymentu.

3. Realizacja

Tworzenie programu przebiegało etapowo, rozpoczynając od części pobierającej i parsującej dane z serwisu internetowego. Część ta przysporzyła najwięcej trudności.

Początkowo obiektem zainteresowania był portal HuffingtonPost.com, jednak doświadczenie pokazało, że sprawia bardzo wiele trudności samo wczytanie strony z serwisu. Ponadto Portal ten nie posiadał żadnego sposobu na wyszukiwanie artykułów po czasie zamieszczenia, ani po tagach, co bardzo utrudniało przeszukiwanie treści.

Drugi wybór, serwis Politico.com pozwalał na proste wyszukiwanie artykułów z dziedziny aktualnych wyborów, można było także wyszukiwać je w ten sam sposób wstecz. Niestety, obydwa serwisy korzystają z systemu komentarzy kontrolowanych przez portal Facebook.com. Technologia ta w dużej mierze opiera się na skryptach ładujących na bieżąco komentarze i niezwykle trudno jest zrobić z niej użytek bez uruchamiania strony w prawdziwej przeglądarce. Ponadto, nawet przeglądarki mają kłopoty, żeby ją dobrze obsługiwać (np. Mozilla Firefox).

Dlatego właśnie do wczytywania stron i zbierania z nich komentarzy został użyty framework Selenium ze sterownikiem do przeglądarki Internet Explorer. To połączenie pozwoliło wczytywać strony i część komentarzy, dalej jednak nie obsługiwało ładowania większej liczby komentarzy, co ograniczyło liczbę zebranych komentarzy z jednego artykułu do 10 i prawdopodobnie miało duży wpływ na końcowe wyniki. Ostatecznie zebrano ponad 75000 komentarzy.

Wszystkie wyniki, ze względu na prostą strukturę danych były przechowywane w plikach csv.

Druga część projektu przetwarzała zebrane komentarze i analizowała je pod względem sentymentu.

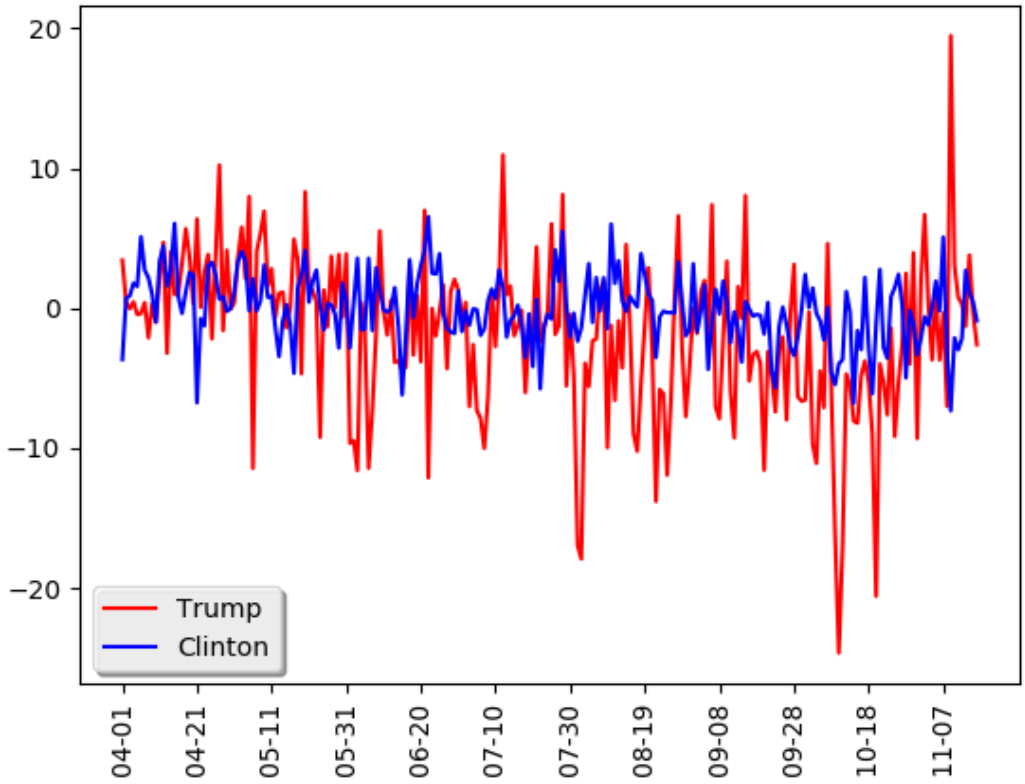
Jeśli w danym zdaniu, albo je poprzedzających, padło jedno ze słów kluczowych dotyczących któregoś z kandydatów, sentyment tych zdań zapisywany był na poczet tegoż kandydata.

Samo rozpoznawanie sentymentu zdania próbowano zrealizować na różne sposoby, jednak ostatecznie najlepszym i najprostszym rozwiązaniem okazał się gotowy do wykorzystania SentimentIntensityAnalyser z leksykonu Vader. Bazuje on na słowniku, w którym słowa mają przypisaną wagę - pozytywny lub negatywny sentyment. W trakcie korzystania z tego rozwiązania zaobserwowano subiektywnie dobrą skuteczność tego podejścia.

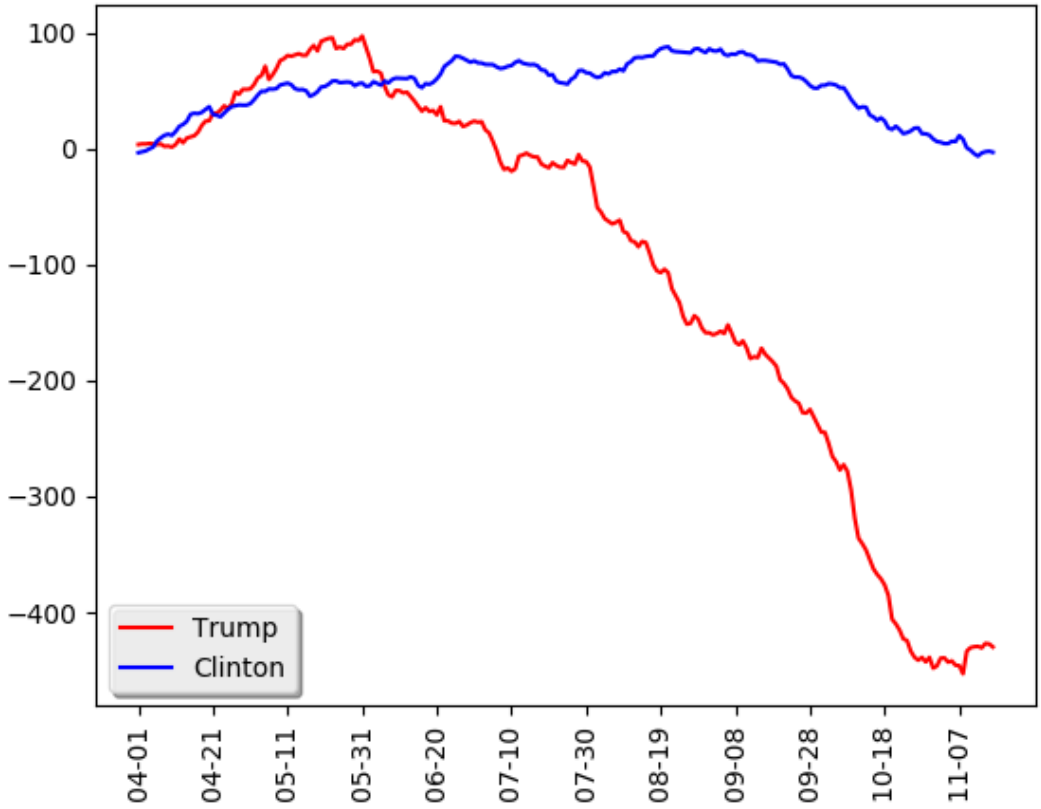
Sentyment obliczony dla wszystkich komentarzy z danego dnia był sumowany i zapisywany, naturalnie dla każdego z kandydatów osobno.

Ostatni element programu tworzył wykres popularności kandydatów bazując na komentarzach z danego dnia. Powstał zarówno wykres samych wartości sentymentu (pozytywne albo negatywne komentarze oraz ich ilość mają wpływ na dodatni lub ujemny wskaźnik) jak i wykres różnicowy, który jako start dla danego dnia przyjmuje wartość z dnia poprzedniego.

Wykres 1: Wartości obliczonego sentymentu od daty:

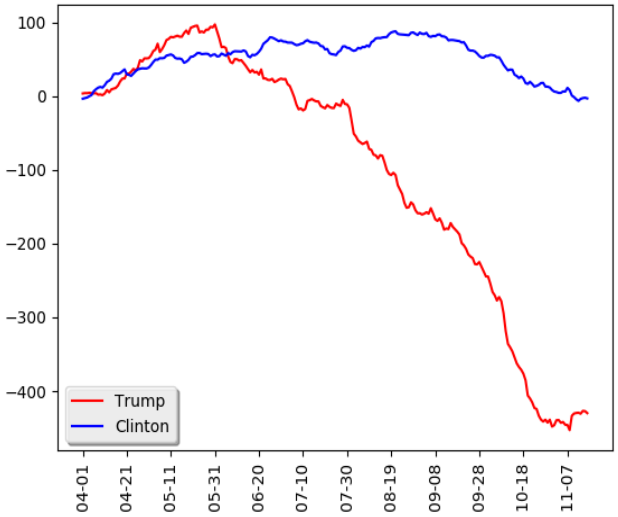


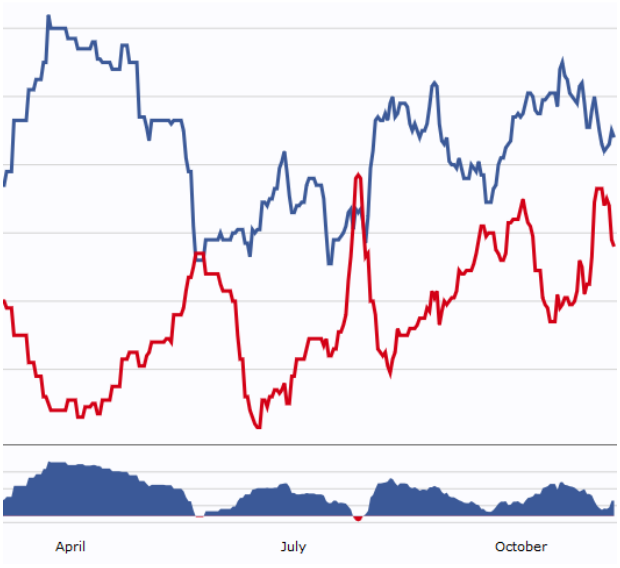
Wykres 2: Różnicowy wykres sentymentu od daty:



4. Wyniki

Wynikiem, który można interpretować i porównywać, jest wykres drugi. Trzeba przyznać, że różni się on znacznie od wykresów, które można znaleźć na portalach badających poparcie dla kandydatów.





Po lewej: otrzymane wyniki, po prawej: statystyki z realclearpolitics.com (kolory znaczą to samo)

Widzimy wyraźnie, że w otrzymanych wynikach stosunek do Donalda Trumpa mocno maleje i praktycznie nie rośnie, a wyniki Hillary Clinton zmieniają się w niewielkim stopniu.

Powodów do tego może być kilka.
Po pierwsze, najprawdopodobniej jest to spowodowane niedostateczną liczbą zebranych komentarzy.
Po drugie, może to być błąd metody, może sumowanie wyników wszystkich komentarzy o danym polityku z jednego dnia jest techniką zbyt prostą, aby otrzymać miarodajne wyniki.
Po trzecie, możliwe też, że wyniki są dobre, ale na tym właśnie portalu panowały silnie pogarszające się nastroje względem Donalda Trumpa oraz delikatnie pozytywny nastrój względem Clinton, pogarszający się nieco pod koniec trwania kampanii.

Nie jest łatwo ocenić, który z tych czynników miał największy wpływ na wyniki, pewne jest natomiast, że Trump był znacznie częściej podmiotem silnie oceniających komentarzy, co ilustrują obydwa wygenerowane wykresy.

Aby zbadać poprawność zebranych wyników, można by dodatkowo:
-ponownie zebrać komentarze, w większej ilości
-wykorzystać inną technikę liczenia sentymentu wypowiedzi
-wykorzystać bardziej złożoną technikę sumowania sentymentu
-wykorzystać dane rozwiązanie na innym serwisie i porównać charakter otrzymanych wykresów