|  |
| --- |
|  |
| Text Mining – Twitter sentiment analysis |
|  |

|  |
| --- |
| Tomasz Garski (260784)  12/16/2018 |

# Opis danych:

Przedmiotem analizy przeprowadzanej w ramach tego projektu jest baza wpisów na twitterze. Zawiera ona łącznie 31961 wpisów. Każdemu z nich przyporządkowano 3 kolumny – id, label oraz zawartość komentarza.

Projekt wykonany został w języku Python 3.6 przy użyciu bibliotek sklearn oraz TextBlob przeznaczonych do analizy tekstu i obrazów.

Program można podzielić na 3 główne składowe:

* Obliczanie wartości statystycznych
* Preprocessing
* Przetwarzenie tekstu

# Dane statystyczne:

Stworzona do zbierania danych statystycznych klasa Counter pozwala na wyciągnięcie wszelkich interesujących danych o tekście. Są to dane takie jak średnia ilość słów w poście, co mogłoby zostać użyte do badań społeczności internetowej, ale również informacje przydatne w dalszym przetwarzaniu tekstu (najczęściej występujące słowa, najrzadziej występujące słowa).

# Preprocessing:

## Usuwanie najczęsciej i najrzadziej występujących słów:

Usunięcie najczęściej występujących słów znacznie upraszcza analizę tekstu, jako że słowa tego typu zazwyczaj są spójnikami lub powszechnymi zwrotami niewpływającymi na znaczenie/wydźwięk tekstu. Należy tu oczywiście szczególną uwagę zwrócić na słowo ‘nie’ oraz inne słowa mogące mieć istotne znaczenie, które często występują w tekście.

Usuwanie najrzadziej występujących słów pozwala uniknąć wiązania ich z konkretnym wydźwiękiem wypowiedzi (które to powiązanie może być z dużym prawdopodobieństwem błędne).

## Poprawa składni:

Poprawianie literówek i błędów pozwala z większą precyzją analizować tekst, poprawnie łącząc posty zawierające te same słowa, które mogłyby być uznane w trakcie przetwarzania za różne z powodu niewielkich błędów.

## Tokenizacja:

Rozbicie ciągu słów na pojedyncze elementy zawierające poszczególne słowa. Przydatne w niektórych przypadkach przetwarzania tekstu.

## Lemmatyzacja:

Bardziej efektywna i zalecana wersja tzw. ‘stemmingu’, który polega na usuwaniu ze słów sufiksów decydujących o tym, do jakich części mowy należy dane słowo. Upraszczając, w języku angielskim usuwane są końcówki wyrazów takie jak –ly, -ing, -s, pozostawiając słowo w jego podstawowej formie.

Lemmatyzacja, zamiast usunąć sufiks, co może być w niektórych przypadkach niewystarczające, konwertuje słowo do formy bezokolicznika, co sprawia, że w każdym poście będzie ono postrzegane tak samo.

# Przetwarzanie tekstu:

W ramach projektu wykorzystane zostały trzy metody przetwarzania i analizy tekstu. Są to kolejno: Bag of words, Words Embeddings, Sentimental Analysis.

## Sentimental Analysis:

Analiza tego typu pozwala na określenie, czy dana wypowiedź jest negatywna czy pozytywna. Każdy post oceniany jest w skali <-1;1>, gdzie -1 jest oznacza wypowiedź negatywną, zaś 1 oznacza wypowiedź w pełni pozytywną. 0 oznacza wartość neutralną.

## Bag of words:

Podejście Bag of words polega na założeniu, że jeśli wypowiedzi zawierają podobne zestawy słów, powinny one mieć podobne znaczenie.