Laboratorium

Przetwarzanie tekstu i analiza opinii

Zadanie 1

Zapoznaj się ze stronami NLTK: https://www.nltk.org/ oraz https://www.nltk.org/book/ . Na potrzeby tego zadania przejrzyj też poniższe samouczki.

- https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk
- https://realpython.com/python-nltk-sentiment-analysis/
- https://www.geeksforgeeks.org/tokenize-text-using-nltk-python/
- https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
- https://www.geeksforgeeks.org/python-lemmatization-with-nltk/

Punkty do wykonania:

a) Wybierz dowolny i niezbyt krótki artykuły z dowolnego portalu angielskojęzycznego (BBC, NBC, Nature lub inne) . Temat artykułu dowolny – może być polityczny, społeczny, naukowy. Skopiuj go (ręcznie) i zapisz w pliku txt.

b) Dokonaj tokenizacji dokumentu. Podaj liczbę słów po tym etapie.

c) Usuń stop-words z artykułu używając standardowej listy dla słów angielskich. Podaj liczbę słów po tym etapie.

d) Sprawdź czy w naszym zestawie słów (bag of words) są jeszcze jakieś pominięte niepotrzebne słowa. Wówczas dodaj do listy stopwords dodatkowe słowa ręcznie (np. za pomocą komendy append lub extend). Podaj liczbę słów po tym etapie.

e) Dokonaj lematyzacji dokumentu. Jaki lematyzer został wybrany? Alternatywnie: możesz dokonać stemmingu. Podaj liczbę słów po tym etapie.

f) Podaj przetworzony dokument w formie wektora zliczającego słowa. Następnie wyświetl na wykresie słupkowym 10 najczęściej występujących słów (oś X: słowa, oś Y: liczba wystąpień słowa w tekście).

g) Stwórz chmurę tagów (word cloud) dla Twojego dokumentu. Pomocne linki:
https://www.datacamp.com/community/tutorials/wordcloud-python
https://amueller.github.io/word_cloud/
https://pypi.org/project/wordcloud/

Zadanie 2

Wykorzystaj paczkę NLTK Vader (https://www.nltk.org/_modules/nltk/sentiment/vader.html, https://www.nltk.org/howto/sentiment.html) do sprawdzenia jak radzi sobie z analizą opinii/sentymentu. Następnie porównaj to z

a) Wejdź na stronę z hotelami (np. https://www.booking.com/ , https://www.tripadvisor.com/) i znajdź jedną pozytywną opinię o jakimś hotelu i jedną zdecydowanie negatywną. Wybierz opinie w języku angielskim składające się z przynajmniej kilku zdań.

b) Używając narzędzia Vader sprawdź w jakim stopniu obie opinie są pozytywne (pos), negatywne (neg), neutralne (neu) i jaki jest wynik zagregowany wszystkich opinii (compound), który waha się od -1 (negatywny) do 1 (pozytywny).

c) Teraz wykorzystaj paczkę Text2Emotion, żeby sprawdzić jak obie opinie są tagowane wg pięciu emocji.

d) Czy wyniki dla obu narzędzi są zgodne z oczekiwaniami?

e) Spróbuj dodać parę pikantnych słów do obu tych recenzji, tak aby oceny sentymentu były silniejsze i powtórz eksperyment dla obu narzędzi.

Zadanie 3

Analizowaliśmy pojedyncze teksty, ale żeby robić poważne badania trzeba mieć większe bazy dany tekstu lub narzędzia do pozyskiwania tekstów.

Szczególnie ciekawe do badań są tweety (twitter.com), które są zwięzłymi tekstami otagowanymi dodatkowo hasztagami. Można ich pobrać setki, mieć mnóstwo różnych opinii i punktów widzenia na przeróżne tematy.

Celem tego zadania jest automatyczne pobranie przynajmniej 100 tweetów na wybrany temat. Można to zrobić za pomocą różnych narzędzi:

- Tweepy (popularne, wykorzystujące Twitter Api, wymaga posiadania konta developera, niestety można ściągać tweety tylko z ostatnich dni)
- Snscrape (mniej popularna, nie wymaga posiadania konta developera, więc to „szara strefa" pozyskiwania danych, nie ma ograniczeń czasowych)

a) Wybierz temat tweetów (może być jakiś konkretny hasztag) np. #ukraine, #christmas, #blackfriday lub inny

b) Wykorzystaj Tweepy lub Snscrape do pozyskania około 100 tweetów na dany temat. Wyświetl te tweety.

c) Zbadaj czy można pobierać tweety z różnych okresów czasu, różnych lokalizacji. Spróbuj zebrać dodatkowo 50 tweetów z wybranego okresu czasu z okolic Gdańska.

Zadanie 4

Wykorzystaj wiedzę, z zad 1, 2, 3 aby zrobić analizę opinii dla tweetów na wybrany przez Ciebie temat.

Pobieramy bazę danych tweetów na wybrany temat. Temat powinien być na tyle ciekawy (i kontrowersyjny), że wzbudza różne emocje u ludzi.
Takie propozycje były w roku 2021/22 (można pomyśleć nad bardziej aktualnymi):

a) **Szczepienia na COVID-19**. Temat można potraktować całościowo lub spróbować wyodrębnić na dwa podtematy: przeciwnicy i zwolennicy szczepienia. Jakie hasztagi i słowa klucze są specyficzne dla obu tych grup?
b) **Wybory prezydenckie w USA (Trump vs Biden).** Porównanie dla obu kandydatów. Jak zmieniało się nastawienie ludzi wobec nich na przestrzeni czasu. Czy emocje w tweetach są skorelowane z sondażami wyborczymi?
c) **Wybory prezydenckie w Polsce** (Duda vs Trzaskowski vs Inni). Pytania jak wyżej (uwaga: język polski?).
d) **Porównanie partii politycznych.** (Zjednoczona Prawica vs Koalicja Obywatelska vs Polska 2050 itd.) Jak się zmieniają emocje na twisterze dla poszczególnych partii i czy jest to skorelowane z sondażami wyborczymi.
e) **Protesty kobiet, wyrok TK, aborcja**. Temat można potraktować całościowo lub spróbować wyodrębnić na dwa podtematy: przeciwnicy i zwolennicy restrykcyjnych przepisów aborcyjnych. Uwaga: polskie tweety?
f) **Rasizm** (np. #blacklivesmatter, #blm, #alllivesmatter). Można się zastanowić jakie grupy społeczne posługują się hasztagiem #blacklivesmatter, #blm a jakie hasztagiem #alllivesmatter? Czy obie grupy można jakoś scharakteryzować?
g) **Seksizm / molestowanie kobiet** (np. #metoo). Jest tu duża osób solidaryzujących się z ofiarami przemocy. Czy jest też grupa ignorująca lub sprzeciwiająca się tej akcji? Jakie hasztagi i emocje ją charakteryzują?
h) **Opinie o znanych osobach**, które mają zwolenników i przeciwników, lub opinia o nich zmieniła się w czasie (politycy, hierarchowie kościoła, gwiazdy Hollywood – można wybrać jedną osobę, lub kilka do porownania).
i) **Opinie na temat serialu** – jak się zmieniały. Warto wybrać serial, który trzymał nierówny poziom (np. miał słaby finał). Np. „How I met your mother?", „Game of Thrones", "Orange is the new black", "Riverdale" lub inne. Można wybrać też serial, które wzbudziły kontrowersje (np. czy seriale powinny mieć aktorów o różnym tle etnicznym / kolorze skóry – patrz: Wiedźmin, Bridgertonowie).
j) **Opinie na temat gier komputerowych**. Może być tak, że premiera gry była sukcesem lub klapą (patrz Cyberpunk: #CyberPunk2077 i #CDProjektRED i nieudana premiera + wyciek danych + konsola vs pc). Czy opinie na twitterze są

skorelowane ze zmniejszeniem się wartości akcji firmy? Inne gry: Animal Crossing (Nintendo), Sims.

k) **Klęski żywiołowe, tragiczne wypadki**. Jak zmieniały się emocje (szok, smutek, współczucie, chęć pomocy) na przestrzeni czasu. Szczególnie interesujące byłyby tutaj klęski żywiołowe w Polsce.
l) **Zmiany klimatyczne** (czy przez człowieka).
m) **Płaskoziemcy**. Premiera filmu na Netflixie – czy wpłynęła na opinie.
n) (dla 2023) **Ukraina i Rosja**.
o) (dla 2023) Inflacja.

W eksperymentach należy uwzględnić poniższe aspekty.

- Jakie dane zawiera wybrana **baza danych** (okres, lokalizacja, język). W jaki sposób dane zostały pozyskane? Szczególnie ważny jest dobór okresu czasu, w którym było najwięcej tweetów.
- Jak tweety zostały **przetworzone**? Czy wyciągano hasztagi, słowa kluczowe? Jak rodzaj lematyzacji zastosowano? Itp.
- Przedstawienie **często występujących istotnych słów** w formie wykresu słupkowego i chmury tagów. Można też podzielić grupy tweetów w zależności od podtematu (np. dla dwóch kandydatów na prezydenta można zrobić osobne chmury tagów).
- Przyjęcie sensownej taktyki **zgłębiania emocji**. Jak podział wybrać: pozytywne vs negatywne? Czy uda się wyodrębnić więcej emocji np. wesołość/smutek/szok/złość itp.? Jak skutecznie oceniać tweety pod względem emocji (narzędzia, techniki, listy słów)?
- Zrobienie osobnych statystyk (np. chmury tagów) dla tweetów o **różnych emocjach**. Czym się różnią?
- Analiza zmian emocji w czasie np. na przestrzeni tygodni lub miesięcy (np. 100 tweetów pobieramy codziennie przez 3 miesiące = 3000 tweetów). Następnie rozrysowanie na **wykresie liniowym** natężenia emocji w tweetach (np. oś X = czas, Oś Y = liczba tweetów pozytywnych/negatywnych). Jak emocje pozytywne/negatywny się wznosiły i opadały? Alternatywnie: wybrać kilka okresów czasowych (np. przed trailerem filmu, po trailerze ale przed premierą filmu, po premierze filmu) i przeanalizować opinie z każdego okresu osobno.
- Interpretacja wyników: czy badania przyniosły jakieś zaskakujące wnioski? Czy wahania emocji w czasie wynikają z jakichś szczególnych zdarzeń?