7

ANALIZA I GENEROWANIE TEKSTU

BAG OF WORDS, WYDŹWIĘK, SIECI REKURENCYJNE

💻 ZADANIE 1: **BAG OF WORDS**

Zapoznaj się ze stronami NLTK: https://www.nltk.org oraz https://www.nltk.org/book/ .

Ponadto, 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/

Możesz poszukać innych podobnych tutoriali w internecie.

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.

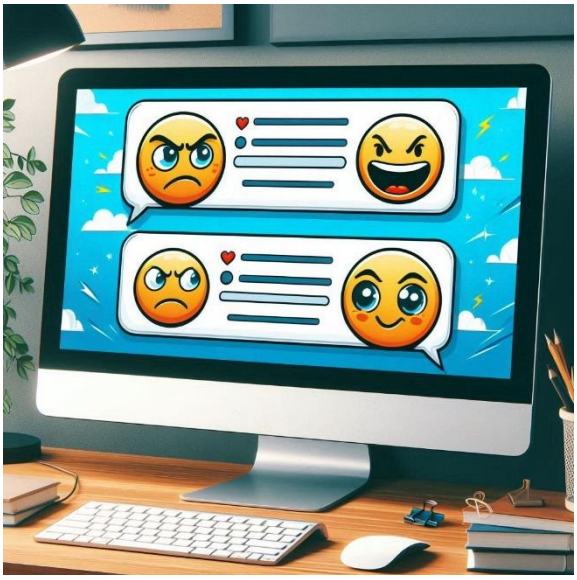
e) Dokonaj lematyzacji dokumentu. Jaki lematyzer został wybrany? Alternatywnie: możesz dokonać stemmingu. Podaj liczbę słów po tym etapie.

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/

Grzegorz Madejski
grzegorz.madejski@ug.edu.pl
Zakład Sztucznej Inteligencji,
Instytut Informatyki, Uniwersytet Gdański

💻 ZADANIE 2: **WYDŹWIĘK RECENZJI**



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 ją z innymi narzędziami wydobywającymi emocje.

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. Jeśli uważasz, że opinie są zbyt neutralne, możesz dodać parę słów nacechowanych emocjami.

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 inne narzędzie do wydobywania emocji. Możesz przetestować Text2Emotion, żeby sprawdzić jak obie opinie są tagowane wg pięciu emocji. Możesz też sięgnąć po nowsze deep-learningowe narzędzia np. Bert, który jest transformerem.

d) Czy wyniki testów są zgodne z oczekiwaniami?

💻 ZADANIE 3: **POBIERANIE POSTÓW**

Ważnym czynnikiem w analizie postów jest ich zdobywanie. Szczególnie interesujące są wszelakiej maści posty, informacje i opinie w mediach społecznościowych: Twitter/X, Facebook, Reddit. Gdyby udało nam się pobrać dużą bazę danych postów, to można je analizować ze względu na tematykę, emocje, itp. Szczególnie interesujący jest Twitter/X: ma krótkie posty, zasobne w informacje i opatrzone hasztagami.

Celem tego zadania jest wybranie jednego z portali społecznościowych (Twitter/X, Facebook, Reddit, lub inny) oraz wybranie narzędzia do automatycznego pobierania postów. Następnie trzeba pobrać 100 postów na wybrany przez Ciebie temat i pokazać zapisane posty w pliku.

Jakie narzędzie wybrać?

- W pierwszej połowie 2023 roku bardzo dobrym narzędziem do ściągania postów z Twittera i Reddita była paczka **snscrape**. Niestety po przejęciu Twittera przez E.Muska została ona zablokowana i radzi sobie już tylko z Redditem. Ludzie nadal zgłaszają, że nie działa dla Twittera, ale developer chyba nie chce jej poprawiać: https://github.com/JustAnotherArchivist/snscrape/issues
- Być może alternatywą dla Twittera będzie paczka **Nitter**. Warto sprawdzić czy zadziała.
- Pojawiają się też jakieś inne nowe narzędzia np. **twscrape** https://github.com/vladkens/twscrape

Jeśli nie uda się z Twitterem, wybierz inny portal np. Reddit.

Grzegorz Madejski
grzegorz.madejski@ug.edu.pl
Zakład Sztucznej Inteligencji,
Instytut Informatyki, Uniwersytet Gdański

💻 ZADANIE 4: **LSTM I GENEROWANIE TEKSTU**

Wykonaj następujące polecenia związane z sieciami rekurencyjnymi.

a) Uruchom i przeanalizuj pliki z wykładu o sieciach rekurencyjnych:
- rnn01.py (prosta demonstracja)
- rnn02.py (Badanie jak działa sieć rekurencyjna)
- rnn03.py (Przewidywanie liczby plam na słońcu w danym miesiącu -RNN)
- lstm01.py (Przewidywanie liczby plam na słońcu w danym miesiącu - LSTM)
- lstm02.py (Uczenie generowania tekstu przez LSTM litera po literze)
- lstm03.py (Uruchomienie generatora LSTM z poprzedniego zadania)
- lstm04.py (Uczenie generowania tekstu przez LSTM słowo po słowie)
- lstm05.py (Uruchomienie generatora LSTM z poprzedniego zadania)

Uwaga: w zadaniach lstm02.py i lstm04.py uczenie może trwać wiele godzin.

b) Znacznie zmniejsz liczbę epok, by trenowanie trwało parę minut.

c) Zademonstruj jak działają generatory tekstu (lstm03 i lstm05) po Twoim krótkim treningu.

d) Dotrenuj model z lstm02 i lstm04 o parę epok, wykorzystując jako bazę startową już wytrenowany przez Ciebie model zapisany w pliku hdf5. Następnie pokaż, czy generowany tekst jest trochę lepszy.

Grzegorz Madejski
grzegorz.madejski@ug.edu.pl
Zakład Sztucznej Inteligencji,
Instytut Informatyki, Uniwersytet Gdański