Anna Cieśniewska 2MUPJN

Projekt: Stworzyć własny korpus przy użyciu CategorizedPlainTextCorpusReader zawierający co najmniej po 5 tekstów z 4-5 kategorii i podać przykład użycia korpusu. Skorzystać z dokumentacji NLTK: http://www.nltk.org/howto/corpus.html

Data przygotowania: 11.01.2022.

Film review corpus

W trakcie projektu powstał korpus składający się z recenzji filmowych, napisanych w języku angielskim. Korpus zawiera 8 recenzji negatywnych oraz 8 recenzji pozytywnych, zapisanych w plikach typu txt, np. movie1_neg.txt.

Program napisano używając oprogramowania PyCharm, w języku Python 3. Głównym narzędziem stosowanym w tym projekcie był PlaintextCorpusReader, ze szczególnym wykorzystaniem klasy CategorizedPlaintextCorpusReader, przy którego użyciu stworzono korpus recenzji filmowych. Oprócz tego, kluczową paczką użytą przy tym projekcie była paczka nltk oraz stopwordsiso, z której zostały pobrane stopwords w języku angielskim. Dzięki niej z korpusu zostały usunięte tak zwane słowa nieznaczące, które często występują w tekście jednak nie są istotne podczas jego analizy. Obie paczki zostały zainstalowane używając wbudowanego dla środowiska języka Python 3 menadżera pakietów – pip.

1. Install NLTK: run `pip install --user -U nltk`

```
$ pip install stopwordsiso
```

Wszystkie użyte w korpusie recenzje są w języku angielskim. Teksty pochodzą z serwisu internetowego https://www.rottentomatoes.com. Na danej stronie można znaleźć informacje dotyczące filmów, jak i recenzje ich dotyczące, pisane przez prywatnych użytkowników oraz wykwalifikowanych krytyków filmowych. Teksty użyte w korpusie dotyczą różnych filmów, a

ich klasyfikacja do kategorii positive lub negative zależały od oceny wystawionej przez autora recenzji:

- Negative: 1-3 / 10
- Positive 7-10 / 10

Pierwszym krokiem podczas pisania kodu do projektu było przygotowanie danych do późniejszej analizy. Teksty zostały zapisane w formie plików txt, oraz zgrupowane w wspólnym pliku znajdującym się w pamięci komputera.

Następnym krokiem było przygotowanie korpusu nazwanego my_corpus. Używając CategorizedPlaintextCorpusReader program otrzymał dostęp do folderu z plikami tekstowymi zawierającymi recenzje. Następnie używając cat_pattern oraz przy pomocy wyrażeń regularnych pliki zostały skategoryzowane.

```
my_corpus = CategorizedPlaintextCorpusReader('/Users/annaciesniewska/Desktop/CORPUS_PROJEKT', r'movie._.*\.txt', cat_pattern=r'movie._(\w+)\.txt')
categories = my_corpus.categories()
```

Kolejnym istotnym krokiem było przygotowanie tekstu znajdującego się w plikach do analizy. Teksty zostały pozbawione znaków innych niż litery, w celu pozbycia się między innymi znaków interpunkcyjnych. Następnie wszystkie litery znajdujące się w tekstach zostały sprowadzone do małych liter. Z tekstów zostały również usunięte słowa utworzone z mniejszej ilości niż 2 znaków oraz stopwords.

```
def prepare_words(text, with_stopwords=True, min_length=1):
    words = []
    for word in text:
        if str.isalpha(word) and len(word)>=min_length:
            word_lower = str.lower(word)
            if with_stopwords or word_lower not in stopwords("en"):
                words.append(word_lower)
    return words
```

Korpus został użyty do wykonania paru analiz statystycznych. Pierwszą z nich są funkcje należące do sekcji „Words length" polegające na wyznaczeniu najdłuższych słów. Dzięki tej funkcji dowiadujemy się następujących informacji o najdłuższych słowach w całym korpusie jak i każdej z kategorii:

- Ilość znaków w najdłuższych słowach
- Jakie słowa są najdłuższe
- Ile razy występują w korpusie lub w danej kategorii

Następna funkcja polega na sprawdzeniu dziesięciu najczęściej występujących słów w każdej z kategorii. Dzięki niej użytkownik dowiaduje się jakie to słowa oraz ile razy wystąpiły w danej kategorii.

Program oferuje również opcję sprawdzenia, ile słów występuje w wybranym przez użytkownika tekście użytym w korpusie. Oprócz tego można sprawdzić takie podstawowe informacje jak:

- Na jakie kategorie korpus został podzielony
- Listę tekstów, z których składa się korpus
- Listę tekstów przyporządkowanych do każdej kategorii

Do umożliwienia łatwej nawigacji stworzono menu wyświetlające wszystkie funkcje dostępne w programie. Wybranie odpowiedniego numeru powoduje wywołanie wybranej opcji w programie.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
OPTIONS:
1 - CORPUS CATEGORIES
2 - LIST OF TEXTS
3 - LIST OF TEXTS USED IN EACH CATEGORY
4 - THE LONGEST WORDS
5 - THE MOST COMMON 10 WORDS
6 - NUMBER OF WORDS IN A GIVEN TEXT
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Choose your option:
```

Stworzony w ramach projektu korpus recenzji filmowych może zostać wykorzystany również na wiele innych sposobów. Jednym z nich może być sentyment analysis, oraz sprawdzenie negatywnego oraz pozytywnego zabarwienia tekstów przy wykorzystaniu list słów nacechowanych oboma sentymentami.