Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

**Analiza tekstu książki “Wojna i Pokój”**

Mateusz Tomczak

432561

Raport z projektu zaliczeniowego

Text Mining w języku Python

Warszawa, czerwiec 2023

**WSTĘP**

Głównym celem projektu jest analiza tekstu książki pt. „Wojna i Pokój” autorstwa Lwa Tołstoja. Książka wydana została w 1867 roku i składa się z 4 tomów podzielonych na 15 ksiąg i epilog złożony z 2 części. Do przeprowadzenia badania wykorzystano angielską wersję książki w nowoczesnym tłumaczeniu. Badanie ma na celu wychwycenie właściwości tekstu wspomnianej książki, zawartych w niej tematów, występujących słów oraz bohaterów jak i ogólnego nastawienia analizowanego tekstu. Analiza została przeprowadzona przy wykorzystaniu narzędzi i bibliotek języka Python.

**METODOLOGIA**

Do przeprowadzenia analizy wykorzystane zostały różne biblioteki języka python. W celu pobrania tekstu wykorzystano bibliotekę Selenium, co wymagało również wykorzystania biblioteki webdriver\_manager. W celu znalezienia odpowiednich fragmentów pobranego tekstu wykorzystano bibliotekę RegEx (re). Do odpowiedniego sformatowania analizowanego tekstu wykorzystano biblioteki NLTK oraz Genisim. Do liczenia występowań tokenów wykorzystano bibliotekę collections. Do stworzenia interaktywnej wizualizacji tematów posłużono się biblioteką pyLDAvis, natomiast do tworzenia wykresów wykorzystano bibliotekę matplotlib. Do przeprowadzenia analizy emocji i sentymentu wykorzystano bibliotekę transformers. Dodatkowo wykorzystano biblioteki tqdm (do tworzenia pasków postępu) oraz wordcloud (do tworzenia wizualizacji tematów).

Badanie rozpoczęto od tokenizacji oraz znalezienia bigramów. Następnie dane te zostały przefiltrowane w celu eliminacji skrajnych przypadków występujących słów. Po tym stworzony został model Word2Vec oraz LDA. Na koniec zbadane zostały emocje oraz nastawienie przekazywane w książce.

**DANE**

W pierwszej kolejności wymagane było pobranie tekstu analizowanej książki. Do tego celu wykorzystano webscraping przy użyciu biblioteki Selenium. Dzięki temu ze strony gutenberg.org pobrany został pełny tekst książki. Następnie tekst wymagał odpowiedniego sformatowania by nadawał się do analizy. W pierwszym kroku usunięte zostały znaki nowej linii i zastąpione zostały znakami spacji. W kolejnym kroku wymagane było wydzielenie faktycznej części książki poprzez usunięcie zbędnych informacji dotyczących wydania książki i jej aspektów technicznych, znajdujących się na jej początku oraz końcu. W tym celu wykorzystano bibliotekę re dzięki której możliwe było znalezienie wymaganej części tekstu używając wyrażeń regularnych. Po tym zabiegu książka została podzielona na księgi (books) i rozdziały (chapters). W tym celu ponownie wykorzystane zostały wyrażenia regularne.

Posiadając odpowiednio przygotowany tekst możliwe było rozpoczęcie działań mających na celu otrzymanie zbioru danych możliwych do analizy. W tym celu każdy z rozdziałów został podzielony na zdania oraz tokeny, przy wykorzystaniu biblioteki nltk. Następnie tokeny zostały ustandaryzowane oraz usunięte zostały tokeny niebędące słowami oraz stop-słowa (stopwords). Następnie do zbioru zostały wprowadzone bigramy, do których wyodrębnienia wykorzystano bibliotekę genisim. Kolejnym krokiem była lematyzacja, ponownie przy użyciu biblioteki nltk, której celem była standaryzacja odmian tych samych słów. Tak przygotowany zbiór wyniósł w sumie 269321 tokenów.

**WYNIKI**

Posiadając już odpowiednio przygotowane dane możliwe było przejście do właściwej analizy tekstu. W pierwszej kolejności sprawdzone zostały najczęściej występujące unigramy i bigramy w tekście. Jednak ze względu na najczęstsze występowanie powszechnych słów, takich jak „say”, „one” czy „go”, które nie odnosiły się bezpośrednio do fabuły, został stworzony odpowiedni słownik. Pozwolił on na usunięcie słów, które występowały w mniej niż 20 rozdziałach oraz w więcej niż połowie rozdziałów. Dzięki temu możliwe było wyodrębnienie najczęściej występujących słów związanych z fabułą analizowanej książki. Najczęstsza dwudziestka takich słów przedstawiona została na Rys. 1.

Jak można zauważyć trzy najczęściej występujące słowa (‘pierre’, ‘natasha’, ‘rostóv’) odnoszą się najprawdopodobniej do głównych bohaterów książki. Z kolei wykorzystując Word2Vec do budowy modelu i sprawdzając najbliższe wyrazy do słowa ‘count’ można zauważyć, że podobne do niego okazują się takie słowa jak 'vladímirovich', 'countess' czy 'cyril', co może wskazywać, że słowo to również wskazuje na jednego z bohaterów książki. Z kolei patrząc na częstotliwość występowania słowa ‘pierre’ można założyć, że jest to główna postać w analizowanej książce. Również widocznie wyższa częstość występowania słowa ‘natásha’ może wskazywać, że jest to postać istotna dla fabuły książki.

Obraz zawierający szkic, linia, Równolegle, Prostokąt

Opis wygenerowany automatycznieRys. 1. Dwadzieścia najczęściej występujących słów (po przefiltrowaniu)

*Źródło:* Opracowanie własne.

Kolejnym etapem analizy była próba znalezienia tematów występujących w książce. Wykorzystana do tego została biblioteka genisim. Na samym początku należało ustalić liczbę tematów, które byłyby wychwycone z fabuły książki. Aby znaleźć ich optymalną liczbę sprawdzono wartości spójności (coherence) dla liczby tematów do 5 do 14 zwiększając ich liczbę przy każdej iteracji o 1. Otrzymane wartości przedstawione zostały w Tabeli 1.

Tabela 1. Wartości spójności dla poszczególnych liczb tematów

|  |  |  |
| --- | --- | --- |
| **Lp** | **Liczba tematów** | **Spójność** |
| 1 | 5 | 0.336049 |
| 2 | 6 | 0.309366 |
| 3 | 7 | 0.331147 |
| 4 | 8 | 0.368873 |
| 5 | 9 | 0.354093 |
| 6 | 10 | 0.351057 |
| 7 | 11 | 0.367837 |
| 8 | 12 | 0.377011 |
| 9 | 13 | 0.355794 |
| 10 | 14 | 0.370263 |

*Źródło:* Opracowanie własne.

Rys. 2. Wizualizacja otrzymanych tematów

Obraz zawierający tekst, zrzut ekranu, diagram, krąg

Opis wygenerowany automatycznie

*Źródło:* Opracowanie własne.

Jak widać najwyższa wartość spójności odpowiada licznie 12 tematów. Przy użyciu biblioteki pyLDAvis stworzona została wizualizacja otrzymanych tematów przedstawiona na Rys. 2. Interaktywna wersja przedstawionej wizualizacji dostępna jest w pliku z kodem źródłowym wykorzystanym do przeprowadzenia analizy. Pierwsze z otrzymanych tematów prawdopodobnie odnoszą się do relacji między postaciami, ponieważ dominują w nich bohaterowie książki. Widać to w temacie 1 bazując na słowach ‘pierre’, ‘natásha’, ‘rostóv’, ‘sónya’, ‘princess’, ‘denísov’ oraz w temacie 3 poprzez słowa ‘pierre’, ‘rostóv’, ‘natásha’, ‘denísov’, ‘count’, ‘prince\_andrew’. W otrzymanych tematach zauważyć można również dużą ilość odwołań do aspektów wojennych, np. w temacie 5 występują takie słowa jak ‘emperor’, ‘army’, ‘officer’ lub w temacie 6: ‘napoleon’, ‘army’, ‘emperor’, ‘soldier’, ‘officer’. W temacie 8 zauważyć można nawiązanie do dużego wydarzenia przedstawionego w książce, na co wskazują słowa takie jak ‘event’, ‘crowd’, z kolei słowa ‘soldier’ oraz ‘movement’ mogą wskazywać, że chodzi w tym przypadku o duże ruchy wojsk. Bazując na pojawianiu się w tematach takich słów jak ‘napoleon’ czy ‘emperor’ ale i ‘russia’ można wysunąć wniosek, iż książka opowiada o wojnie francusko-rosyjskiej z czasów Napoleona.

Kolejny etap analizy polegał na zbadaniu sentymentu książki oraz występujących w niej emocji. Ze względu na duży rozmiar rozdziałów w książce analiza ta opierała się na zbadaniu wspomnianych aspektów dla każdego zdania występującego w książce. Do sprawdzenia emocji przekazywanych w zdaniach wykorzystano bibliotekę transformers oraz model o nazwie ‘jitesh/emotion-english’. Wykres emocji dla całej książki przedstawiony został na Rys. 3.

Obraz zawierający szkic, linia, rysowanie, design

Opis wygenerowany automatycznieRys. 3. Emocje występujące w książce

*Źródło:* Opracowanie własne.

Jak można zauważyć w książce przeważa strach (fear) oraz podejrzliwość (suspicious). Emocje te najprawdopodobniej powiązane są z częścią wojenną książki. Kolejne popularne emocje, zuchwałość (cheeky) oraz radość (joy) mogą odwoływać się do części pokojowej książki. W kolejnych występujących emocjach zauważyć można widoczny spadek ich częstotliwości występowania, zatem wspomniane cztery można uznać za główne emocje, które można zaobserwować w analizowanej książce. W przypadku podziale na księgi główną emocją jedynie w pięciu przypadkach nie jest ‘strach’. W księdze pierwszej dominuje ‘zuchwałość’, szóstej, ósmej oraz w pierwszym epilogu – ‘podejrzliwość’, natomiast w drugim epilogu dominuje emocja ‘empatyczny’.

Wyniki przedstawiające sentyment zdań występujących w książce przedstawione zostały w Tabeli 2. Badanie zostało przeprowadzone przy użyciu modelu 'distilbert-base-uncased-finetuned-sst-2-english'. Jak można zauważyć nie jest możliwe określenie ogólnego nastawienia książki jako pozytywne lub negatywne, ponieważ wartości te różnią się od siebie jedynie 536 zdaniami (2% wszystkich zdań) na korzyść sentymentu negatywnego. Można zatem uznać, że książka „Wojna i Pokój” ma neutralny sentyment.

Tabela 2. Sentyment zdań występujących w książce

|  |  |  |
| --- | --- | --- |
|  | **Sentyment** | |
| **Pozytywny** | **Negatywny** |
| **Liczba zdań** | 12969 | 13505 |

*Źródło:* Opracowanie własne.

**PODSUMOWANIE**

Wyniki analizy pozwoliły na ogólne przedstawienie tekstu książki „Wojna i Pokój” Lwa Tołstoja. Znalezione tematy pozwoliły na ogólne zarysowanie fabuły i występujących w niej postaci jak i potencjalnie relacji między nimi. Analiza emocji przekazywanych w książce wskazała na dominujące uczucia przekazywane przez książkę, natomiast analiza sentymentu pokazała, że ogólne nastawienie książki można uznać za neutralne. Niestety ograniczenia modeli analizujących emocje oraz sentyment nie pozwoliły na weryfikację tych danych dla całego tekstu czy tekstów poszczególnych rozdziałów, a jedynie na występujących zdaniach. Jednak mimo tego założone cele analizy można uznać za osiągnięte.