Faculty of Humanities

Applied Linguistics and
Text Analytics

Nizhniy Novgorod
2024

An NLP-BASED TEXT ANALYSIS OF JANE AUSTEN'S *SENSE AND SENSIBILITY*

Prepared by
Nwobi, Joy Uchenna

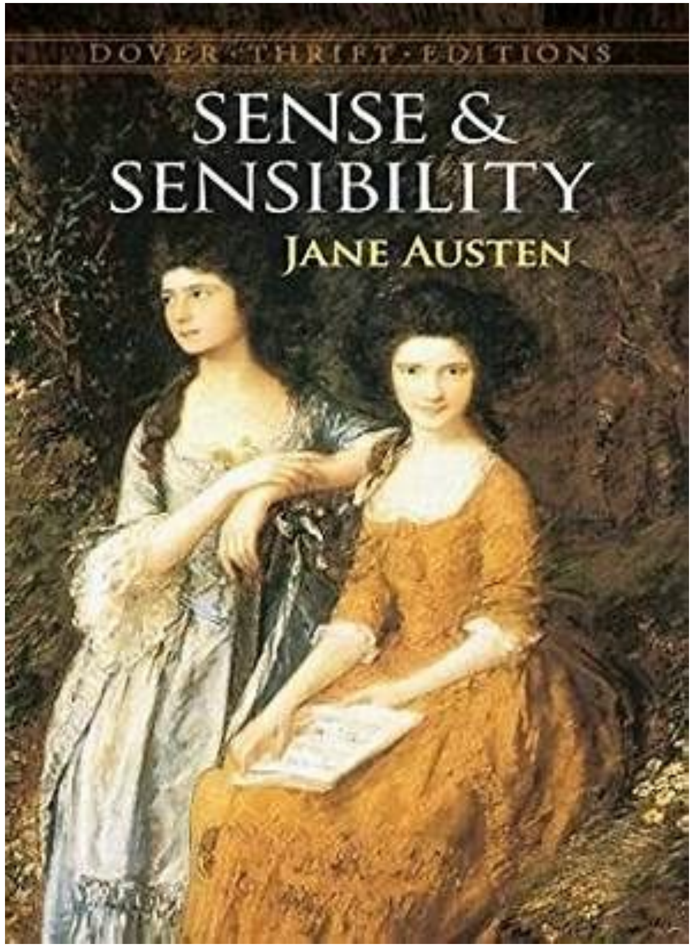
Outline:

1. Aim and Specific objectives of the study
2. Research Data;
3. Analytic tools;
4. Analysis procedure:
   - Lower Casing
   - Punctuation removal
   - Stop words removal
   - Parts of speech tagging (POS)
   - Tokenization
   - Stemming and lemmatization
   - Exploratory data analysis
   - N-grams
   - Word cloud
   - Term Frequency-Inverse Document Frequency (TF-IDF)
   - Topic modelling
   - Text similarity
   - Information extraction – NER – Entity recognition
   - Sentiment analysis

5. Conclusions

Aim and specific objectives

The aim of the study is to analyse the text ***Sense and Sensibility*** by Jane Austen using NLP libraries in Python.



Specifically, the study seeks to:

1) do text preprocessing;

2) describe the part of speech of the text and word and POS count

3) Find main themes in the text;

4. Identify the characters of the book;

5. Identify a similarity score between the **"sense"** and another text written by same author "and ***Persuasion***;

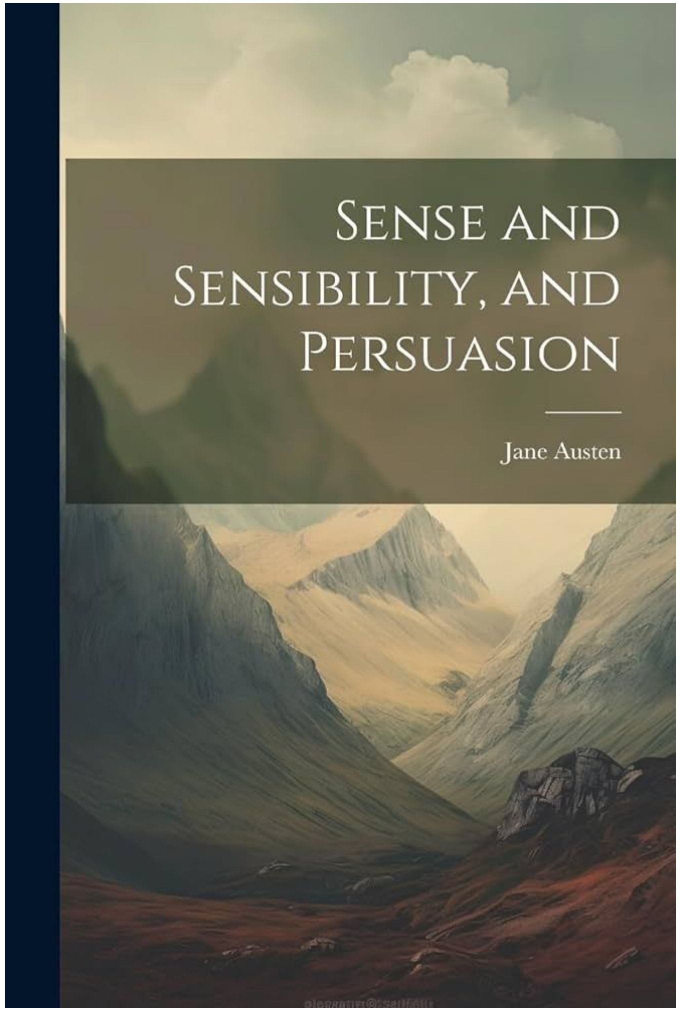
6) Determine the emotional polarities in the text.

Data for Analysis

The data for analysis is the Jane Austen's first novel ***Sense and Sensibility*** (format txt). ***Sense and Sensibility*** was written in **1811.**

To do text similarity, I will compare "sense" with and Austen's last novel ***Persuasion*** (format txt)

Persuasion *was written in* ***1817***



Analytical tools

Text analysis was carried out using **Python libraries**: nltk, spacy, re, gensim, sklearn, vaderSentiment.

Natural language processing techniques used:

- Converting text to lower case (lowercasing);
- Punctuation removal;
- Stop words removal;
- Parts of speech tagging;
- Tokenization;
- Stemming and lemmatization;
- Word cloud;
- N-grams;
- Term frequency-inverse document frequency or TF-IDF;
- Text similarity;
- Information extraction – NER – entity recognition;
- Topic modeling;
- Sentiment analysis.

Analysis procedure

- **The first step**
  - **installation and importation of libraries and lexical corpora (gutenberg)**

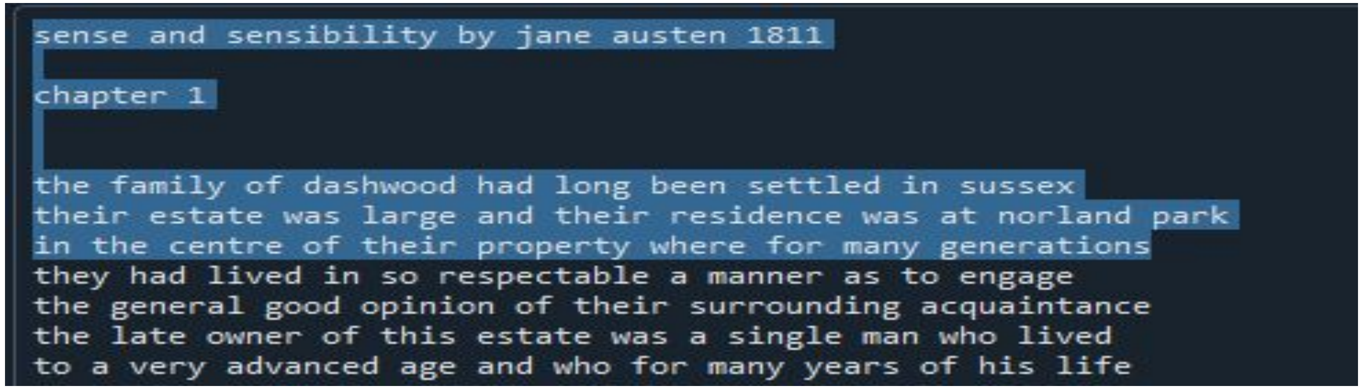
```
ip install textblob
import nltk
from nltk.corpus import gutenberg
from nltk.stem import WordNetLemmatizer, PorterStemmer
from nltk import word_tokenize, pos_tag
from nltk.corpus import stopwords
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.decomposition import NMF, LatentDirichletAllocation
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from gensim.models import Word2Vec
from sklearn.metrics.pairwise import cosine_similarity
from textblob import TextBlob
```

Analysis procedure

Second step
Data cleaning or text pre-processing

This is a necessary step in data analysis. Text preprocessing helps remove unnecessary elements from data, improve the quality of models, and speed up calculations. The text pre-processing includes

- **Lowercasing**
- **Punctuation removal**

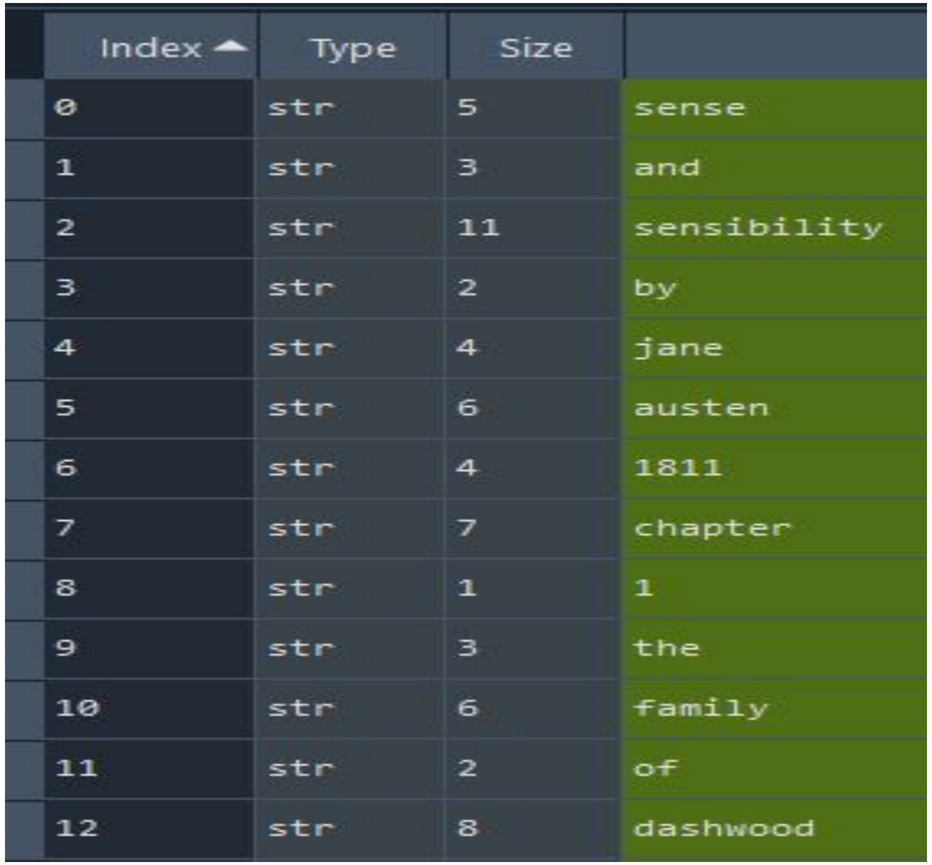


```
sense and sensibility by jane austen 1811

chapter 1


the family of dashwood had long been settled in sussex
their estate was large and their residence was at norland park
in the centre of their property where for many generations
they had lived in so respectable a manner as to engage
the general good opinion of their surrounding acquaintance
the late owner of this estate was a single man who lived
to a very advanced age and who for many years of his life
```

The above figures show “sense” without uppercase and punctuations.

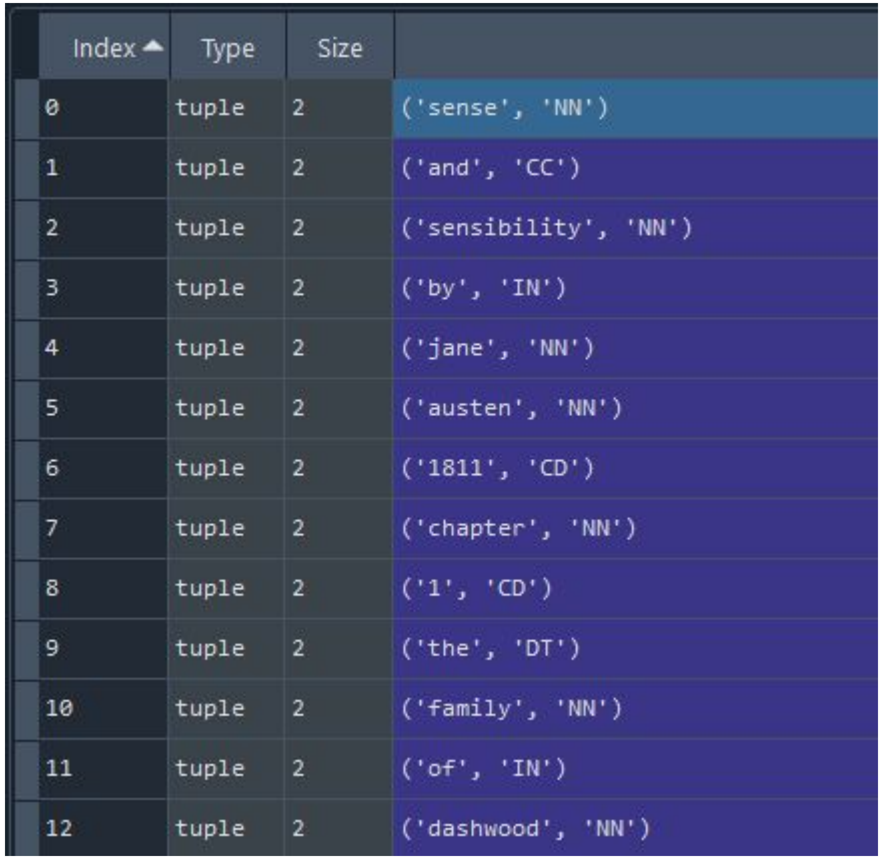
Analysis procedure

- **Tokenization**

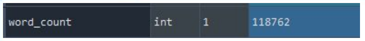


Index ▲	Type	Size	
0	str	5	sense
1	str	3	and
2	str	11	sensibility
3	str	2	by
4	str	4	jane
5	str	6	austen
6	str	4	1811
7	str	7	chapter
8	str	1	1
9	str	3	the
10	str	6	family
11	str	2	of
12	str	8	dashwood

POS Tagging



Index ▲	Type	Size	
0	tuple	2	('sense', 'NN')
1	tuple	2	('and', 'CC')
2	tuple	2	('sensibility', 'NN')
3	tuple	2	('by', 'IN')
4	tuple	2	('jane', 'NN')
5	tuple	2	('austen', 'NN')
6	tuple	2	('1811', 'CD')
7	tuple	2	('chapter', 'NN')
8	tuple	2	('1', 'CD')
9	tuple	2	('the', 'DT')
10	tuple	2	('family', 'NN')
11	tuple	2	('of', 'IN')
12	tuple	2	('dashwood', 'NN')



word_count	int	1	118762

Analysis procedure

Part of Speech tagging and count

Word count of Sense and Sensibility: 118,762

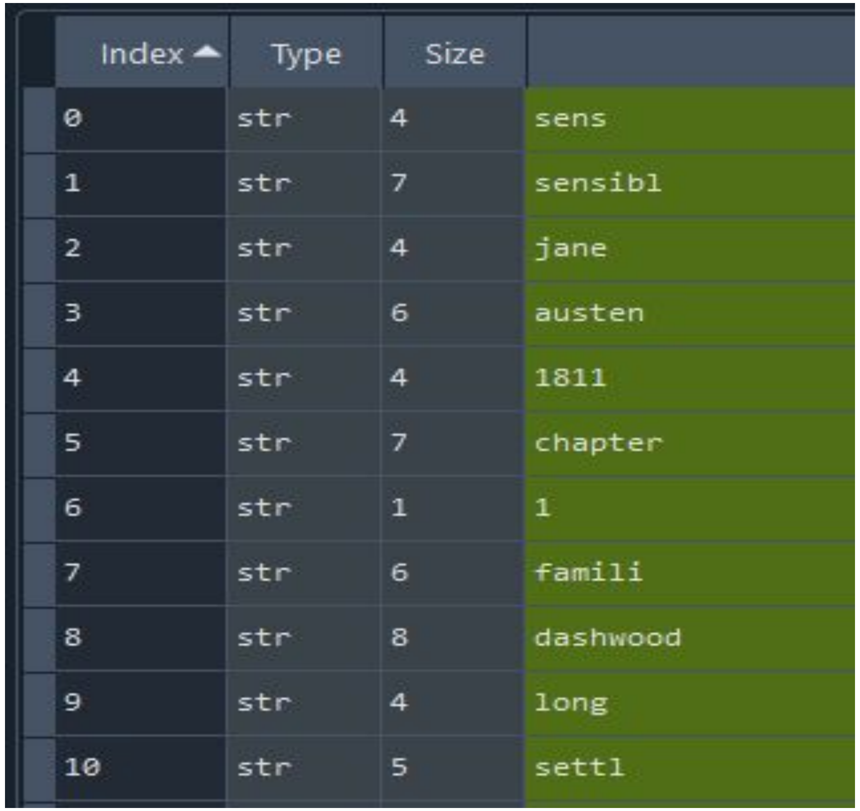
Average word length: 4.422559404523333

Part of speech tag counts: Counter({'NN': 19139, 'IN': 15444, 'DT': 9511, 'PRP': 8628, 'RB': 8610, 'JJ': 8516, 'VBD': 7232, 'VB': 6708, 'PRP$': 4932, 'CC': 4663, 'TO': 4086, 'NNS': 4001, 'VBN': 3643, 'MD': 2812, 'VBP': 2316, 'VBG': 2056, 'VBZ': 1433, 'WDT': 805, 'WP': 767, 'CD': 708, 'WRB': 672, 'JJR': 464, 'RBR': 315, 'RP': 315, 'JJS': 289, 'PDT': 283, 'EX': 177, 'RBS': 137, 'WP$': 48, 'FW': 29, 'UH': 17, 'NNP': 6})

Number of nouns: 19139
Number of adjectives: 8516
Number of verbs: 6708
Number of Adverbs: 8610

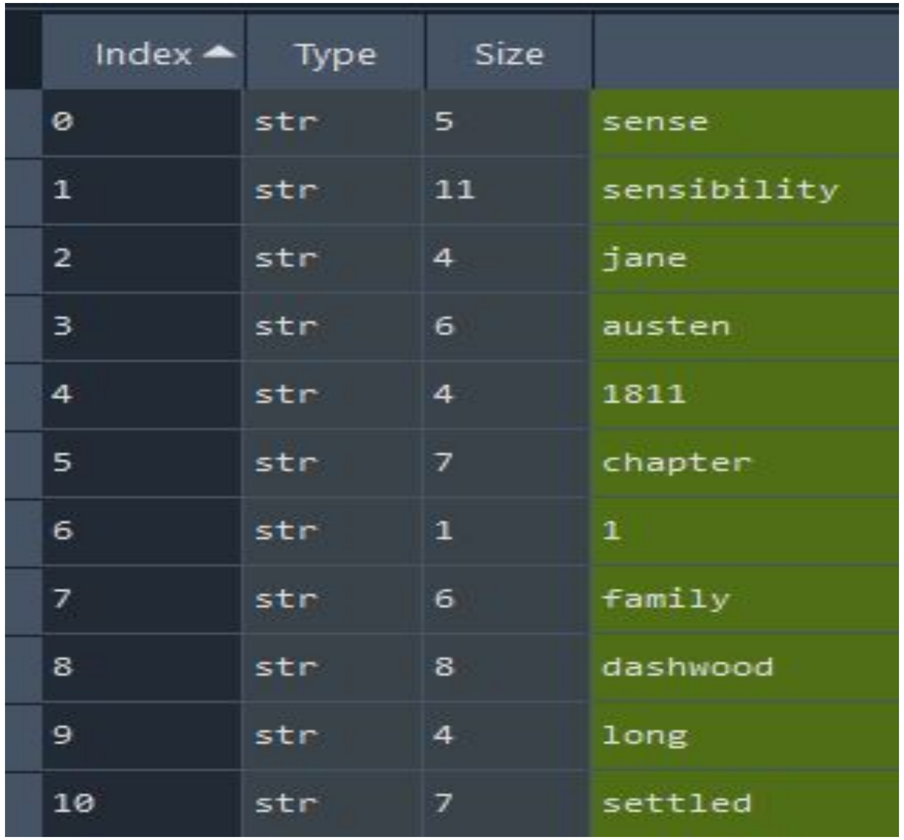
Analysis process

Stemmitisation



Index	Type	Size	
0	str	4	sens
1	str	7	sensibl
2	str	4	jane
3	str	6	austen
4	str	4	1811
5	str	7	chapter
6	str	1	1
7	str	6	famili
8	str	8	dashwood
9	str	4	long
10	str	5	settl

Lemmatization



Index	Type	Size	
0	str	5	sense
1	str	11	sensibility
2	str	4	jane
3	str	6	austen
4	str	4	1811
5	str	7	chapter
6	str	1	1
7	str	6	family
8	str	8	dashwood
9	str	4	long
10	str	7	settled

Exploratory data analysis

N-grams

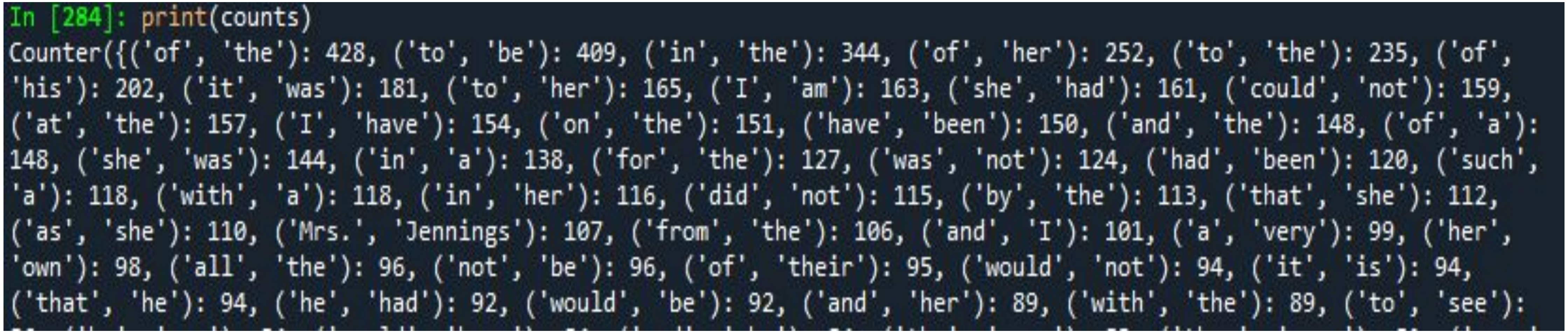
The TextBlob library is used to measure the frequency of bigrams (i.e. pairs of words) in a given text and prints out the bigrams. The bigrams are arranged in descending order

Step

Generate bigrams using ngrams

Count frequencies of each bigram using Counter

Print the Counter object to see bigram frequencies

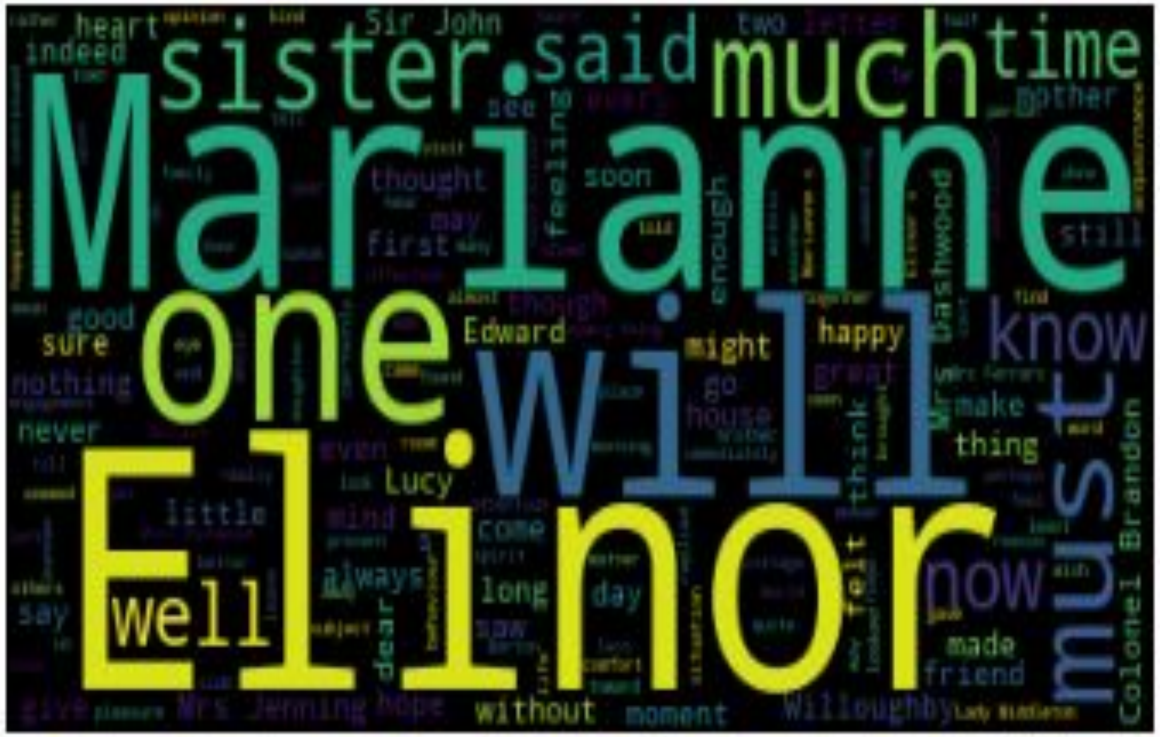


```
In [284]: print(counts)
Counter({('of', 'the'): 428, ('to', 'be'): 409, ('in', 'the'): 344, ('of', 'her'): 252, ('to', 'the'): 235, ('of',
'his'): 202, ('it', 'was'): 181, ('to', 'her'): 165, ('I', 'am'): 163, ('she', 'had'): 161, ('could', 'not'): 159,
('at', 'the'): 157, ('I', 'have'): 154, ('on', 'the'): 151, ('have', 'been'): 150, ('and', 'the'): 148, ('of', 'a'):
148, ('she', 'was'): 144, ('in', 'a'): 138, ('for', 'the'): 127, ('was', 'not'): 124, ('had', 'been'): 120, ('such',
'a'): 118, ('with', 'a'): 118, ('in', 'her'): 116, ('did', 'not'): 115, ('by', 'the'): 113, ('that', 'she'): 112,
('as', 'she'): 110, ('Mrs.', 'Jennings'): 107, ('from', 'the'): 106, ('and', 'I'): 101, ('a', 'very'): 99, ('her',
'own'): 98, ('all', 'the'): 96, ('not', 'be'): 96, ('of', 'their'): 95, ('would', 'not'): 94, ('it', 'is'): 94,
('that', 'he'): 94, ('he', 'had'): 92, ('would', 'be'): 92, ('and', 'her'): 89, ('with', 'the'): 89, ('to', 'see'):
```

Word cloud

In the context of the text, "***Sense and Sensibility***", the words "Marianne", "one", "will", "Elinor", "sister", and "much time" appearing in the word cloud suggest that these words are frequently used or hold significant importance in the text.

Marianne" and "Elinor" are the names of the two main characters, who are sisters. This explains why "sister" might also appear frequently.



Term Frequency-Inverse Document Frequency (TF-IDF)

Step:

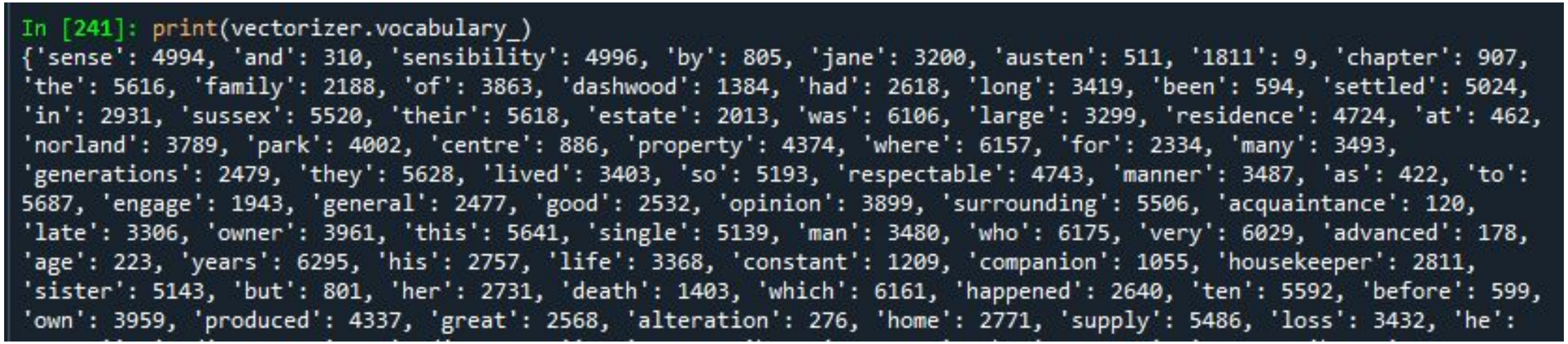
Fit vectorizer to learn vocabulary

Access learned attributes like vocab, idf

Transform text to TF-IDF encodings

Print results as needed

Result:



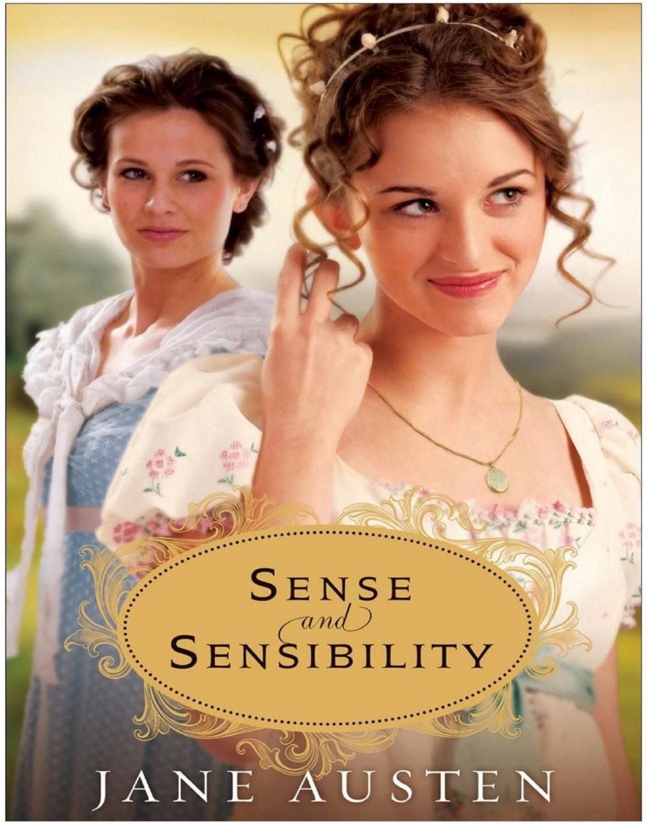
```
In [241]: print(vectorizer.vocabulary_)
{'sense': 4994, 'and': 310, 'sensibility': 4996, 'by': 805, 'jane': 3200, 'austen': 511, '1811': 9, 'chapter': 907,
'the': 5616, 'family': 2188, 'of': 3863, 'dashwood': 1384, 'had': 2618, 'long': 3419, 'been': 594, 'settled': 5024,
'in': 2931, 'sussex': 5520, 'their': 5618, 'estate': 2013, 'was': 6106, 'large': 3299, 'residence': 4724, 'at': 462,
'norland': 3789, 'park': 4002, 'centre': 886, 'property': 4374, 'where': 6157, 'for': 2334, 'many': 3493,
'generations': 2479, 'they': 5628, 'lived': 3403, 'so': 5193, 'respectable': 4743, 'manner': 3487, 'as': 422, 'to':
5687, 'engage': 1943, 'general': 2477, 'good': 2532, 'opinion': 3899, 'surrounding': 5506, 'acquaintance': 120,
'late': 3306, 'owner': 3961, 'this': 5641, 'single': 5139, 'man': 3480, 'who': 6175, 'very': 6029, 'advanced': 178,
'age': 223, 'years': 6295, 'his': 2757, 'life': 3368, 'constant': 1209, 'companion': 1055, 'housekeeper': 2811,
'sister': 5143, 'but': 801, 'her': 2731, 'death': 1403, 'which': 6161, 'happened': 2640, 'ten': 5592, 'before': 599,
'own': 3959, 'produced': 4337, 'great': 2568, 'alteration': 276, 'home': 2771, 'supply': 5486, 'loss': 3432, 'he':
```

Topic Modelling

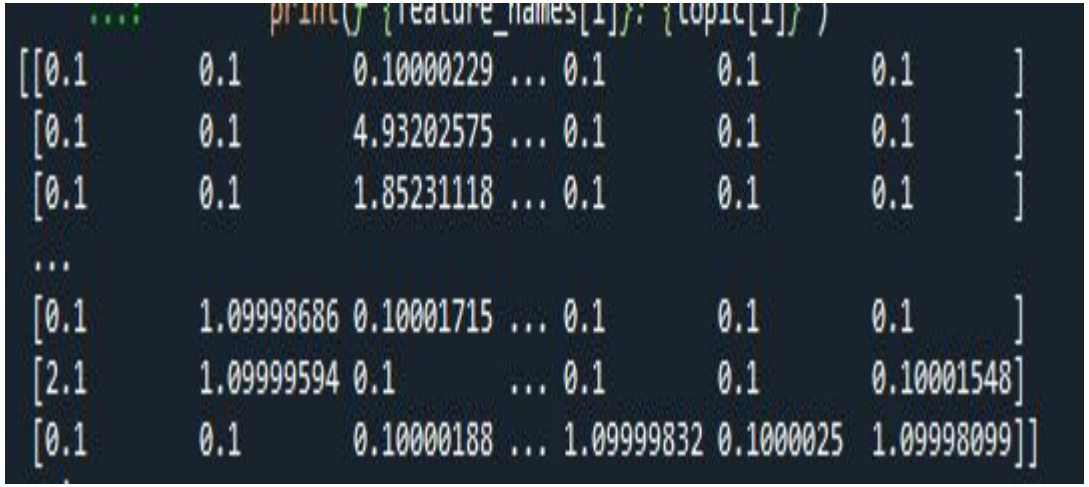
Using LDA

step

1. Initialize the LDA (Latent Dirichlet Allocation) model:
2. Fit the LDA model to the data:
3. Print the topic-word distributions:
4. Apply LDA with a random seed:
5. Print topics and keywords:



Topic Modelling



```
   ...:            print(f"{feature_names[i]}: {topic[i]}")
[[0.1        0.1        0.10000229 ... 0.1        0.1        0.1       ]
 [0.1        0.1        4.93202575 ... 0.1        0.1        0.1       ]
 [0.1        0.1        1.85231118 ... 0.1        0.1        0.1       ]
 ...
 [0.1        1.09998686 0.10001715 ... 0.1        0.1        0.1       ]
 [2.1        1.09999594 0.1        ... 0.1        0.1        0.10001548]
 [0.1        0.1        0.10000188 ... 1.09999832 0.1000025  1.09998099]]
```

Topic #1:

lord:	17.757922798288952
ah:	7.882919354601692
read:	5.087319327315791
good:	3.4455189298852686
williams:	3.123873609820211
excellent:	2.9180836203523066
smith:	2.581647787723462
smiled:	2.577330665810062
dance:	2.472496254462307
bless:	2.3046552762869124

Topic #2:

you:	79.82164419547817
is:	42.905227259632206
do:	36.246171881614664
it:	35.00022329583992
not:	34.1930165575266
what:	32.43241609594537
be:	30.11538826482029
said:	30.038164641860874
am:	27.799385050819204
he:	27.096866607959498

Topic #3:

laughing:	3.193833829468165
tomorrow:	2.0800922649033753
oxford:	2.0474180277339604
steele:	1.8115050181966348
choice:	1.7911297846797392
forty:	1.6723753590896915
opportunity:	1.6372977744688646
views:	1.5943055280969276
agree:	1.5657734439614173
uncle:	1.5634880164763092

Topic #4:

poor:	18.100173028185132
ill:	4.861482134629497
soul:	4.285390800356683
man:	4.1290641949337665
safe:	3.6031728076366147
is:	3.3195286430170223
edward:	2.9912671292240045
hearted:	2.6990669658484037
creature:	2.595279097129666
esteem:	2.5609844518702762

Topic Modelling

```
Topic #5:
familiarity: 51.9427952513129
anticipations: 40.92721975609714
contribute: 38.928473901912774
apparent: 34.981028127497034
gaily: 34.22171993138443
dispatch: 32.69868311509644
asserted: 32.487544137181125
declarations: 31.320584824931906
glimpse: 31.125576504163718
bye: 30.668888230593033
Topic #6:
dirt: 178.81267766656214
dealt: 131.94131534759933
discomposed: 115.20316794030867
five: 112.1939626946955
fairly: 105.37826462838224
discontent: 87.5938435369485
begun: 86.88307936810551
declining: 78.30734200848603
gently: 69.96603491418114
doubts: 62.135894122231974
```

```
Topic #7:
assigned: 52.948805816911026
declining: 50.412305056046186
discharged: 48.72854354767797
detail: 46.319117549206034
dim: 44.86086129508714
henceforward:
38.847085337058154
fix: 35.67396471299645
alicia: 32.55650860393048
god: 30.770668215382127
dependence: 26.877339491100035
Topic #8:
discontent: 333.46275978749156
dealings: 155.22360010089108
calmly: 101.8874918896314
begun: 98.85222007382485
clogged: 93.00905401227476
fellow: 83.47242588822472
concerns: 80.0311507925113
declarations: 59.70438972563594
gladly: 46.522179621384375
fairly: 45.751899369735256
```

```
Topic #9:
detail: 64.46148717930468
distress: 64.20791231245674
familiar: 53.10971168554107
calmly: 35.613513647924925
gladly: 33.33709820557767
bye: 27.24608136295366
climate: 23.126974483109958
children: 22.46654483747931
dined: 22.43272631968621
blindness: 21.996750321682164
Topic #10:
calmly: 181.78583795394047
announcing: 84.4924081390547
asserted: 82.36445584243039
bye: 71.75058449325184
detail: 68.18366038954925
god: 48.76930930626635
discharged: 47.53366677837826
blindness: 43.642159337843815
desperate: 41.74354115014123
fairly: 38.56776337804959
```

Text Similarity

Similarity score between “Sense” and Persuasion: 0.9655717468795175

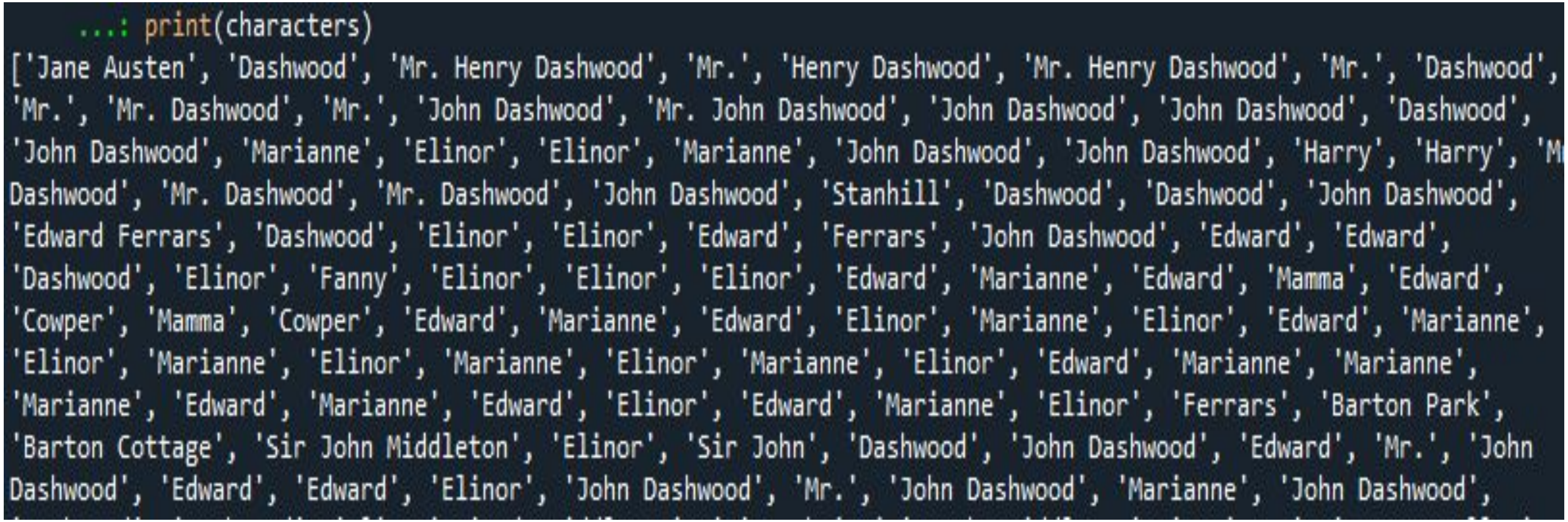
The similarity score of 0.9655717468795175 indicates a high degree of similarity between Jane Austen's "*Sense and Sensibility*" and "*Persuasion*" based on the cosine similarity measure.

Cosine similarity is a metric that measures the similarity between two vectors by calculating the cosine of the angle between them. The resulting similarity score ranges from 0 to 1, with 1 indicating identical vectors and 0 indicating no similarity. The higher the similarity score, the more similar the texts are considered to be.

"Sense" and "Persuasion" share a significant amount of commonality in terms of their word usage and distribution.
The reason for this high similarity may be because they were written by same person.

- **Information extraction – NER – Entity recognition**

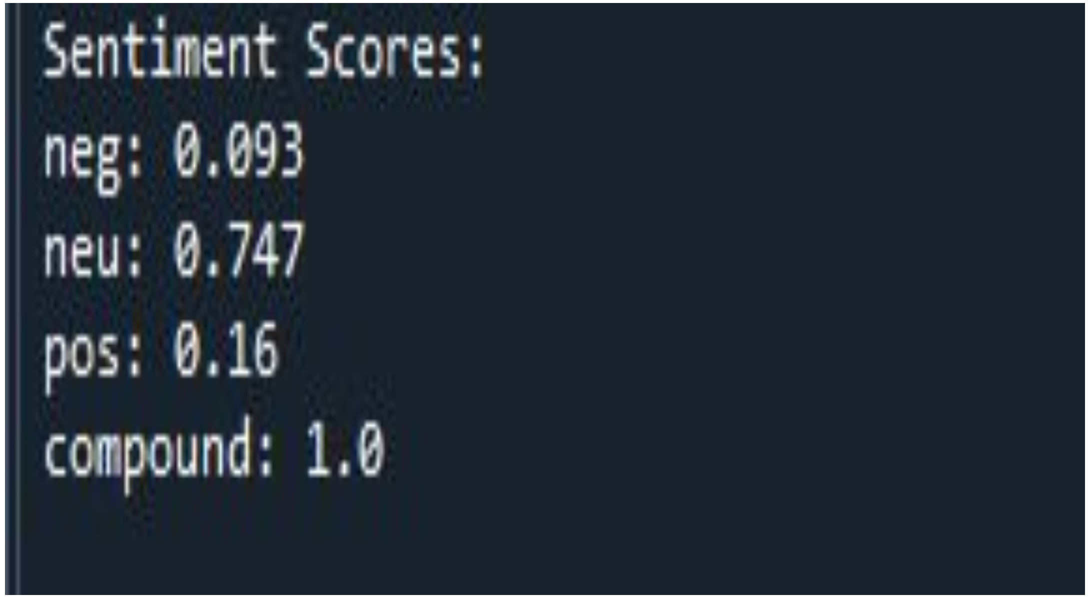
Using NLTK chunker



```
   ...: print(characters)
['Jane Austen', 'Dashwood', 'Mr. Henry Dashwood', 'Mr.', 'Henry Dashwood', 'Mr. Henry Dashwood', 'Mr.', 'Dashwood',
'Mr.', 'Mr. Dashwood', 'Mr.', 'John Dashwood', 'Mr. John Dashwood', 'John Dashwood', 'John Dashwood', 'Dashwood',
'John Dashwood', 'Marianne', 'Elinor', 'Elinor', 'Marianne', 'John Dashwood', 'John Dashwood', 'Harry', 'Harry', 'M
Dashwood', 'Mr. Dashwood', 'Mr. Dashwood', 'John Dashwood', 'Stanhill', 'Dashwood', 'Dashwood', 'John Dashwood',
'Edward Ferrars', 'Dashwood', 'Elinor', 'Elinor', 'Edward', 'Ferrars', 'John Dashwood', 'Edward', 'Edward',
'Dashwood', 'Elinor', 'Fanny', 'Elinor', 'Elinor', 'Elinor', 'Edward', 'Marianne', 'Edward', 'Mamma', 'Edward',
'Cowper', 'Mamma', 'Cowper', 'Edward', 'Marianne', 'Edward', 'Elinor', 'Marianne', 'Elinor', 'Edward', 'Marianne',
'Elinor', 'Marianne', 'Elinor', 'Marianne', 'Elinor', 'Marianne', 'Elinor', 'Edward', 'Marianne', 'Marianne',
'Marianne', 'Edward', 'Marianne', 'Edward', 'Elinor', 'Edward', 'Marianne', 'Elinor', 'Ferrars', 'Barton Park',
'Barton Cottage', 'Sir John Middleton', 'Elinor', 'Sir John', 'Dashwood', 'John Dashwood', 'Edward', 'Mr.', 'John
Dashwood', 'Edward', 'Edward', 'Elinor', 'John Dashwood', 'Mr.', 'John Dashwood', 'Marianne', 'John Dashwood',
```

- ## Sentiment analysis

neg: 0.093 indicates a relatively low level of negative sentiment in the text.
neu: 0.747 suggests that a significant portion of the text is considered neutral in terms of sentiment.
pos: 0.16 represents a moderate level of positive sentiment in the text.
compound: 1.0 ,a normalized and aggregated score, represents the overall sentiment of the text. A score of 1.0 indicates a highly positive sentiment.



```
Sentiment Scores:
neg: 0.093
neu: 0.747
pos: 0.16
compound: 1.0
```

Conclusions

This study has:

1. demonstrated the process of text pre-processing;

2) described the part of speech of the text and word/token and POS count

3) Found main themes in the text;

4. Identified the characters of the book;

5. Identified a similarity score between the ***"sense"*** and another

text written by same author "and ***Persuasion***;

6) and determined the emotional polarities in the text

In my future analysis, I will extensively compare the three texts of Jane Austen namely, *Sense and Sensibility, Pride and Prejudice* and *Persuasion*