ASSIGNMENT III

Sima Shafaei

1. Part 1

For this project, I have selected Five documents including one scientific text, one Screenplay, one novel story, one act, and one poem:

1) **Self-Organizing Systems**: This book is a collection of scientific essays and papers on the topic of self-organizing systems. The book covers various aspects of self-organization, including the mathematical foundations, the principles of organization and adaptation, the applications of self-organizing systems in biology, sociology, and psychology, as well as the implications for technology and engineering.

2) **Titanic screenplay**: The Titanic screenplay is a script written for the 1997 blockbuster movie "Titanic," directed by James Cameron. The movie is a fictionalized account of the sinking of the RMS Titanic in 1912 and the love story between the two main characters, Jack and Rose.

3) **Alice's Adventures in Wonderland**: This book has a series of adventures and encounters various characters, including the Cheshire Cat, the Mad Hatter, and the Queen of Hearts. Throughout her journey, Alice must navigate the nonsensical and unpredictable nature of Wonderland while trying to find her way back home. The book is often considered a classic of children's literature. Its narrative, structure, characters, and imagery have had a widespread influence on popular culture and literature, especially in the fantasy genre.

4) **The Prophet**: The book is a collection of 26 essays or "prose poems," each dealing with a different aspect of human experience and offering insights into life, love, marriage, work, joy, sorrow, and other topics. The essays are written in the form of conversations between a wise prophet named Almustafa and various people he encounters on his journey.

5) **Romeo and Juliet**: This is a tragedy act written by William Shakespeare and is about the romance between two Italian youths from feuding families.

Table 1: Genre and number of words in the selected text books

Book Title	Genre	Number of Words
Self-Organizing Systems	Scientific	25948
Titanic screenplay	historical fiction, drama, and romance	43358
Alice's Adventures in Wonderland	A unique blend of fantasy, children's literature, and literary fiction	26765
The Prophet	A combination of philosophical and spiritual literature, with elements of poetry, essay, and parable	12494
Romeo and Juliet	Tragedy	25997

2. Part 2

First, I performed the following preprocessing on each text:

1) Lowercased all text to ensure uniformity and avoid duplicates due to capitalization.

2) Removed numbers.

3) Removed special characters

4) Removed extra white space

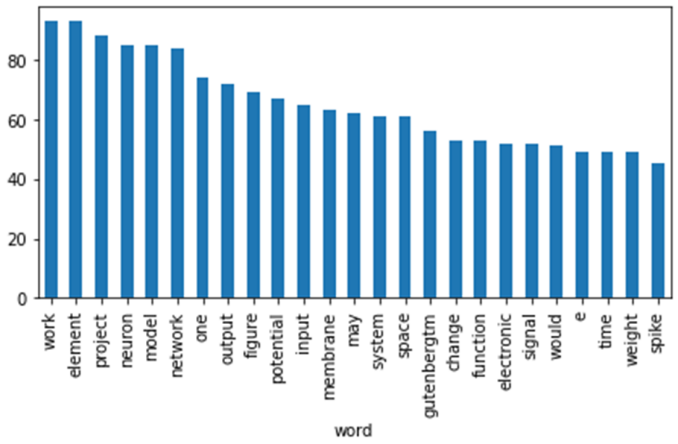
5) Removed stopwords

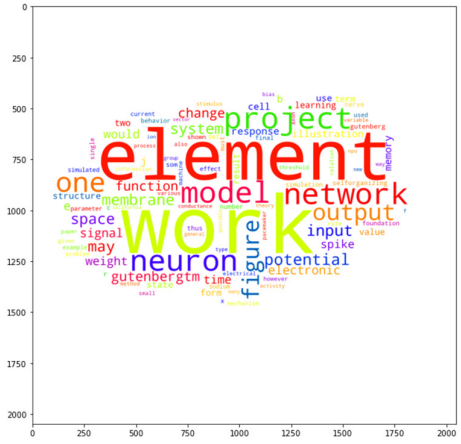
6) Lemmatized words

Then I created the bar chart of the 25 most frequent words and also the word cloud on preprocessed text.

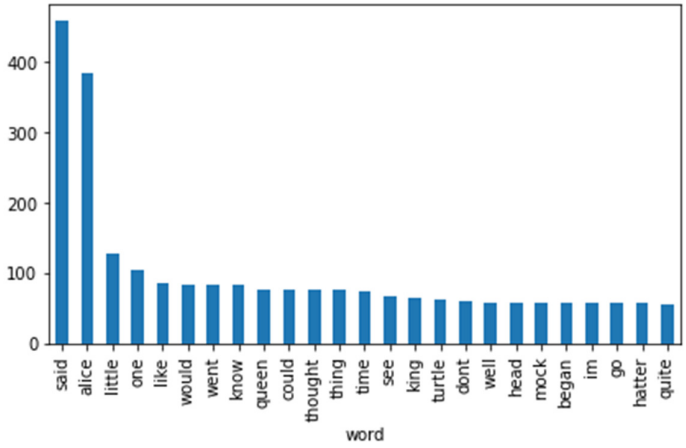
The obtained results for each textbook are shown below. As it is obvious from the pictures, the results of bar charts are compatible with the results of word clouds.

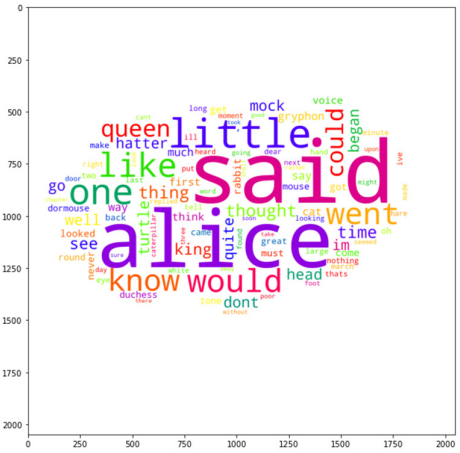
Bar chart of top 25 words and the cloud word of Self-Organizing Systems:



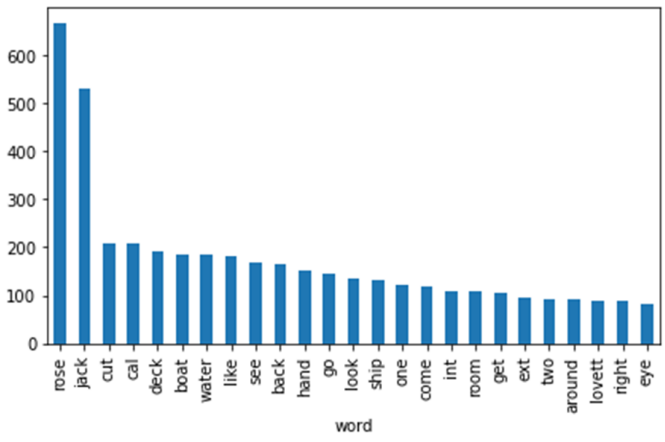


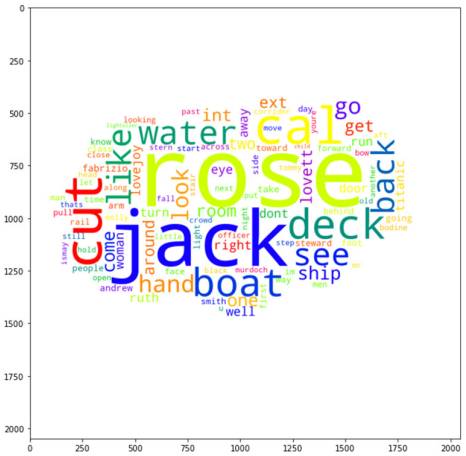
Bar chart of top 25 words and the cloud word of Alice's Adventures in Wonderland:



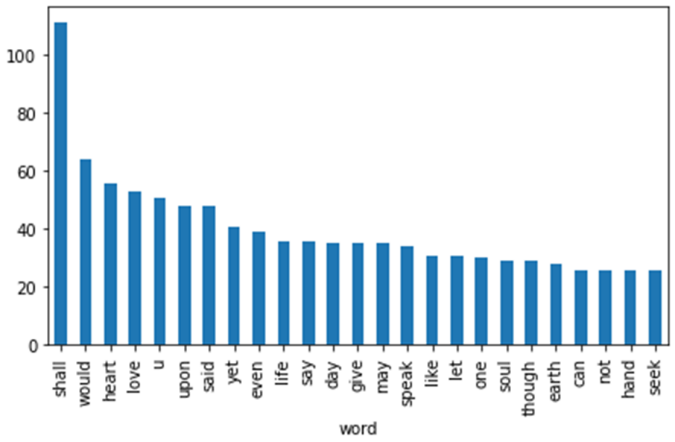


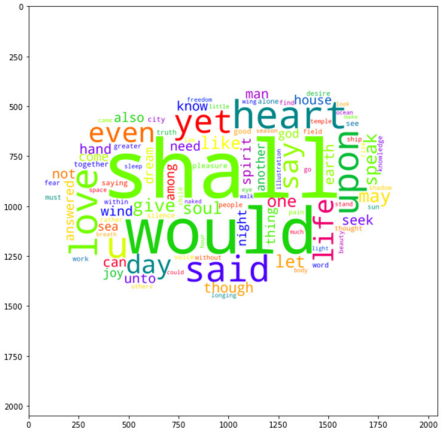
Bar chart of top 25 words and the cloud word of Titanic:



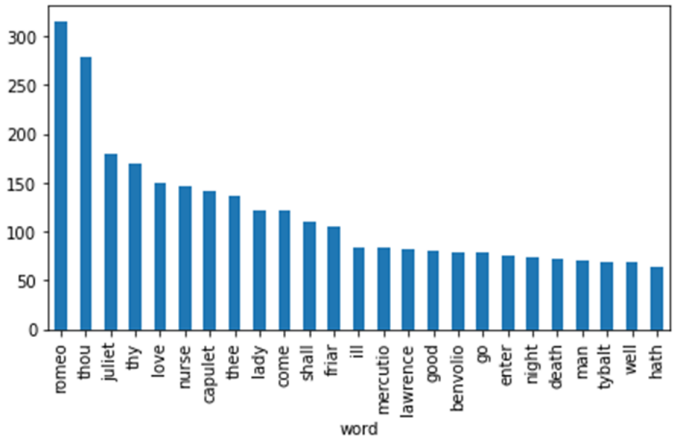


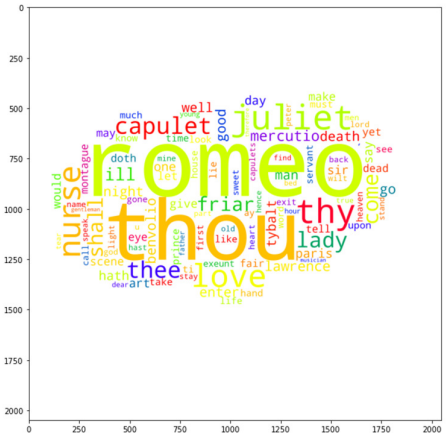
Bar chart of top 25 words and the cloud word of The Prophet:





Bar chart of top 25 words and cloud word of Romeo and Juliet:





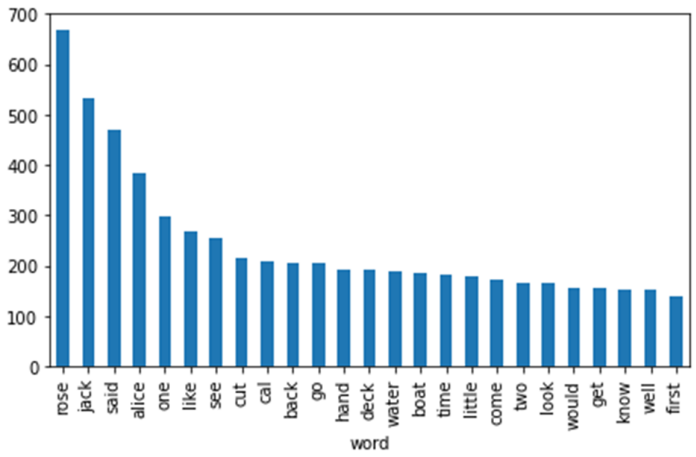
3. Part 3

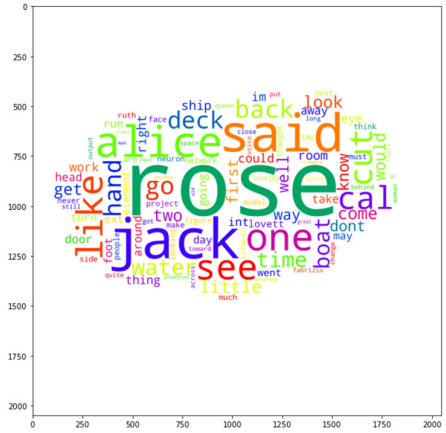
For simplifying analyses, this part is performed only on the three first textbooks including Self-Organizing Systems, Alice's Adventures in Wonderland, and Titanic.

It is worth noting that there are some words in the cumulative text that was not among the 25 most frequent words in any of the initial texts, such as "first" and "back." This could be due to the cumulative effect of combining multiple texts with different themes and styles, resulting in some new words appearing more frequently than expected. However, despite the inclusion of new words, "Rose" and "Jack" still stand out as the most repetitive words in the

cumulative text. This is likely due to the writing style of the Titanic screenplay, where the names of the main characters are frequently repeated throughout the script.

Bar chart of top 25 words and cloud word of the cumulative text:





The CSV file is created as follows:

	word	count
0	rose	669
1	jack	531
2	said	469
3	alice	385
4	one	298
5	like	268
6	see	254
7	cut	214
8	cal	208
9	back	206
10	go	204
11	hand	193
12	deck	193
13	water	190
14	boat	184
15	time	183
16	little	178
17	come	171
18	two	167
19	look	165
20	would	155
21	get	155
22	know	153
23	well	151
24	first	139
TotalCount		6083

4. Part 4

According to the readability scores, the Self-Organizing Systems book appears to be the most readable, while Titanic and Romeo and Juliet have the lowest scores. This outcome is not surprising since scientific texts often use simpler sentences and vocabulary, making them easier to comprehend. Conversely, literary works like plays and novels can be more challenging to read due to their complex language and structures.

Book Title	Average Readability Score
Self-Organizing Systems	11.15
Titanic screenplay	4.75
Alice's Adventures in Wonderland	6.54
The Prophet	6.77
Romeo and Juliet	4.77

5. Part 5

All similarity scores seem high to me. For example, Alice's Adventure in Wonderland which is a children's book in the genre of fantasy, and Romeo and Juliet which is a tragedy book with more literary and poetic words has a 0.975 similarity score!

However, it is notable that the Self-Organizing Systems book has the lowest similarity scores with the other books, particularly Romeo and Juliet, which is to be expected given its distinct subject matter and technical language.

In terms of the highest similarity score, it is interesting that Alice's Adventures in Wonderland and Titanic screenplay has a score of 0.979, indicating that they share some common elements or themes despite being from different genres. Although I expected the highest similarity to be between The Prophet and Romeo and Juliet because of their similarities in terms of language, and style (0.975), their scores are not very far from each other. and we should keep in mind that the quantitative measures do not always align with our expectations.

	Self-Organizing Systems	Titanic	Alice's Adventures in Wonderland	The Prophet	Romeo and Juliet
Self-Organizing Systems	1	0.92	0.912	0.938	0.897
Titanic	0.92	1	0. 979	0.975	0.97
Alice's Adventures in Wonderland	0.912	0. 979	1	0.976	0.975
The Prophet	0.912	0.975	0.976	1	0.975
Romeo and Juliet	0.897	0.97	0.975	0.975	1