EXPLORATORY TEXT ANALYSIS

Kaleb M. Shikur
School of Data Science
University of Virginia
Charlottesville, VA 15213
kms7cu2@virginia.edu

May 11, 2021

ABSTRACT

This project tries to explore 13 different books written by three well known African American writers of the past. The main goal of the project is to extract information about the books, the authors and the underlining social issues described in the books. comparative analysis is performed on the writing styles and perspectives of the three authors. To achieve these goals, I utilized different text analysis tools and techniques. Some of them are TF-IDF, LDA, word-embedding, topic modeling and different Visualization methods. I was able to observe interesting results about both the authors and their works. For instance, the different PCAs separated the works based on authors and other characteristics. The top TFIDF terms for each author also speak volume about the main focus area their works.

1 Introduction

Race is one of the major parts the American social dynamics. Race and other related social phenomenon are usually depicted through music, literature and other forms of art. These mediums usually serve as a rich and effective ways to communicate complex socio-cultural issues. On this paper I try to examine the works of three prominent African American writers whose works mainly focus on race related issues.

The three writers had diverse experiences and outlooks about the time they lived in and the future they looked forward to. Fredrick Douglass was an African American abolitionist, orator, publisher, and author who was born into slavery. He later gained his freedom and wrote a number of books about his life and the institution of slavery[1]. William Edward Burghardt Du Bois was an American sociologist, historian and civil rights activist. As an activist he led several organizations that promoted the advancement of black people in the united states and around the world. He also wrote a number of books about the history of black people. His work also includes two novels [2]. Like Fredrick Douglass, Booker T. Washington was born into slavery and become one of the most prominent African American leaders. His

work mainly focuses on education and business. His approach was at odd with that of Du Bois who was a strong opponent of Washington's Atlanta compromise on equality [3].

2 About The Data

This project is based on 13 books of 3 authors. The books are selected based on their popularity, their main focus area and their suitability for organizing according to OHCO levels. The corpus contains four books from each of the three authors and one book that contains the works of both Booker T. Washington and W.E.B. Du Bois. The main source of data is project Gutenberg website [4].

book_id	book_title	book_file	title	Author	genre
202	Project Gutenberg's My Bondage and My Freedom,...	./data_in\202-0.txt	Project Gutenberg's My Bondage and My Freedom,	Frederick Douglass	b
23	Narrative of the Life of Frederick Douglass, b...	./data_in\23-0.txt	Narrative of the Life of Frederick Douglass,	Frederick Douglass	b
408	The Souls of Black Folk, by W. E. B. Du Bois	./data_in\408-0.txt	The Souls of Black Folk,	W. E. B. Du Bois	n
59116	Why is the Negro Lynched?, by Frederick Douglass	./data_in\59116-0.txt	Why is the Negro Lynched?,	Frederick Douglass	n
62799	John Brown, by W. E. Burghardt Du Bois	./data_in\62799-0.txt	John Brown,	W. E. B. Du Bois	b

Figure 1: LIB table includes the book id from gutenberg and also the title and author as written in the document, the location of the document in the local directory, title of the book, author of the book and genre (weather it's a biography/autobiography or not)

The corpus is organized based on different criteria such as authors and genre. To make the genre fit into a binary class, I limited the categories into autobiography/biography and other genres. Except for one of the documents which is a transcript of a speech by Fredrick Douglass all the documents are well structured into different OHCO levels.

The acquisition of the raw text was fairly simple since the source is organized for such research purposes. The raw data contains additional contents on each ends of the main content. The first step was to remove these contents. After that the documents were organized into different tables such as LIB, DOC, VOCAB and TOKEN.

	book_id	chap_num	para_num	para_str
0	202	1	0	PLACE OF BIRTH - - CHARACTER OF THE DISTRICT ...
1	202	1	1	In Talbot county, Eastern Shore, Maryland, nea...
2	202	1	2	The name of this singularly unpromising and tr...
3	202	1	3	It was in this dull, flat, and unthrifty distr...
4	202	1	4	The reader will pardon so much about the place...

Figure 2: DOC table contains each document divided into paragraphs

term_str	term_id	n	num	stop	p_stem	df	idf	dfidf	tfidf_max4_sum
01	1	1	1	0	01	1.0	5.700440	5.700440	2.885860
1	2	34	1	0	1	8.0	2.700440	21.603518	23.013347
10	3	19	1	0	10	6.0	3.115477	18.692863	18.987982
100	4	11	1	0	100	NaN	NaN	NaN	12.307327
1000	5	8	1	0	1000	1.0	5.700440	5.700440	11.074494

Figure 3: VOCAB table, contains different information about all the terms in the corpus. number of times the term appear in the corpus (n), whether the term is a stopword or not (stop), document frequency (df), inverse document frequency(idf), dfidf and max4 tfidf sum

	book_id	chap_num	para_num	sent_num	token_num	pos_tuple	pos	token_str	term_str
0	202	1	0	0	0	('PLACE', 'NN')	NN	PLACE	place
1	202	1	0	0	1	('OF', 'NNP')	NNP	OF	of
2	202	1	0	0	2	('BIRTH', 'NNP')	NNP	BIRTH	birth
3	202	1	0	0	3	('-', ':')	:	-	nan
4	202	1	0	0	4	('-', ':')	:	-	nan

Figure 4: TOKEN table contains token level properties and tags most of them from NLTK library.

3 Top 10 important terms

The top 10 important terms were extracted based on their tfidf values. The bag used for the general corpus is books and for the individual authors' corpus the bag used is paragraphs. The first list of top 10 terms is for the entire corpus. This list will tell us what the overarching focus areas of the three writers are. The top 10 terms include: student, africa, john, brown, tuskegee, plantation, coloured, crime, labor and cotton. Almost all the terms are one of the top words anyone would immediately associate with slavery and American history. The name john brown who is also highly associated with slavery and abolition made it to this list because two of the 13 documents are mainly about him.

The other lists of top 10 terms were extracted for each individual authors. This helps us to know what each author focused on and also how their works contrast with one another. For the Fredrick Douglass corpus the top 10 most important terms are: Lloyds, mistress, negro, Lloyd, overseer, auld, charge, Michaels, covey, and st. It is interesting to see a number of proper nouns in this list. Since two out of four of his works are biographies, the proper nouns are names of people who are major part of his life such as his first slave holder Lloyds and his other "master" Thomas Auld. The name of his work place, st. Michaels is also part of this list [6]. For W.E.B.D. Bois the top 10 most important terms are: coast, boy, mr, shadow, laborers, music, veil, county, ideals, iron. None of the top 10 terms with the exception of the word laborers is related to the top 10 terms in the general corpus. This can be attributed to the fact that W.E.B.D. Bois's list unlike the other two authors contains a novel that is not directly related to african american history [7].

The top 10 terms for Booker T. Washington's works are very interesting. They seem to summarize the works that he is well known for. For example, He was known for his commitment to education and bussiness [3]. half of the top 10 terms such as acres, corn, soil, education, graduates are related to those areas of his work. The remaining five terms include. your, slaves, religious, etc, and cabin. The other important observation is related to the term Tuskegee. It is one of the most notable achievements of Booker T. Washington. It apears as one of the most important terms in the overall copus but didn't make the cut in the Washington's subcorpus. This can be due to the word being highly prevalent in all his works which will lower the IDF value and also the TFIDF value.

4 Similarity

similarity measures help to compare different documents by representing a text as a point in a vector space and the distance between them as a measure of similarity. Different distance measures are used to compare the documents in this project. some of them are: euclidean, cityblock, jaccard and dice.

		cityblock	euclidean	cosine	jaccard	dice	js	euclidean2
doc_a	doc_b							
7	8	3871.903540	71.143696	0.829831	0.757816	0.610067	0.718873	1.288278
5	7	4171.134430	70.359949	0.849994	0.830418	0.710012	0.738561	1.303836
4	7	4481.743594	70.328303	0.808411	0.826467	0.704256	0.723158	1.271543
7	9	3107.732850	69.995544	0.908027	0.801280	0.668446	0.762957	1.347611
6	7	3468.758209	69.444492	0.839954	0.766852	0.621866	0.722912	1.296113
7	12	3436.595949	69.119316	0.845347	0.802048	0.669516	0.728477	1.300267
2	7	4004.811955	68.936139	0.804262	0.822142	0.697997	0.718738	1.268276
7	10	3171.791262	68.866562	0.824225	0.739873	0.587141	0.716619	1.283920
0	7	3352.956533	68.673606	0.811534	0.756888	0.608866	0.709308	1.273997
3	7	3390.928208	68.658328	0.810000	0.779407	0.638548	0.708754	1.272792

Figure 5: Distance measure for all pairs of books.

The different clustering dendrograms show interesting results. The distance between document is calculated using the pdist function and using three different count measures such as binary count, raw count and probabilistic methods. Ward and complete linkage methods are used with the different distance measures. The two most interesting dendrograms are shown below.

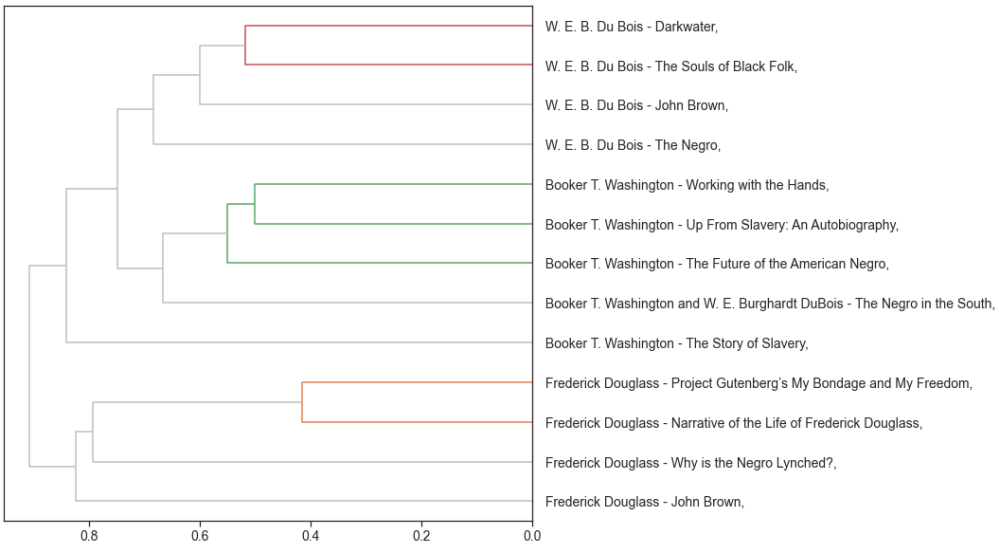

Figure 6: Hierarchical clustering dendrogram: using cosin distance measure and complete linkage. This dendrogram separates the authors successfully. It also places the book written by Washington and edited by Du Bois among the Washington cluster and closer to Du Bois than Fredrick Douglass. The other interesting observation is, the autobiographies of Fredrick Douglass are separated from his other books.

Most of the other Distance measures produce results similar to the above figure and other don't seem to communicate anything that can be noticed immediately. The HCA does a good job clustering the documents by author but fails to separate them by other criteria such as year and genre. We will employee PCA in the next section to see if we can get better results.

Figure 7: Hierarchical clustering dendrogram: using cityblock distance measure. This cluster to some extent separates each book by author but more importantly it isolates Fredrick Douglass's address of John brown separated from the other regular books. unlike the above cluster it puts the book 'The Negro in the south' among the other works of the editor(Du Bois) rather than the author (Washington)

5 Principal component analysis

In this section I try to see if there are features that can explain the different aspects of the corpus. PCA is manly used to surface latent properties.

Figure 8: PC0 and PC1 with label author: PC0 and PC1 do a good job separating the documents by authors. PC0 almost completely separate Booker T. Washington's books from the other two authors and PC1 separates Du Bois and Fredrick Douglas with little overlap. The Negro in the south is located in both it's Author's(Washington) and editors (Du Bois) cluster.

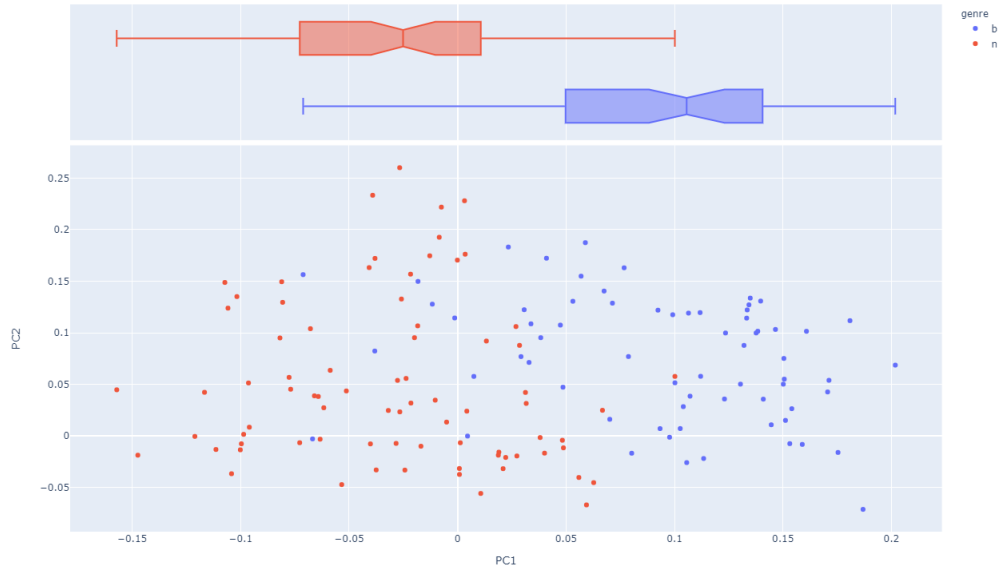

Figure 9: PC1 and PC2 with label genre: PC1 Separates the documents by genre with some overlap

If we further examine the overlap above, The three Authobiographies are almost completely separated from the non-biography books. The overlaping points belong to the biographies of John Brown by Du Bois. The Autobiographies of Fredrick Douglass and Washington are also clustered together which may be due to the similarity of their stories.

6 Topic Models

The other analysis performed on the corpus is LDA. sklearn's LDA tool is used to extract topic models from the corpus. The main BAG used during this analysis is paragraph BAG but other BAGs were also explored. From the corpus 40 topics were extracted for each genre.

genre	b	n	label
topic_id			
28	0.041771	0.120800	28 race, people, man, education, years, men, life, races, country, time
26	0.139161	0.104694	26 school, people, work, time, education, life, men, way, day, man
19	0.025761	0.096549	19 students, school, work, farm, student, training, building, year, value, room
22	0.001266	0.093347	22 trade, century, coast, culture, people, tribes, years, world, civilization, history
12	0.057909	0.081904	12 world, men, nan, man, life, eyes, thing, hands, face, soul
6	0.053921	0.075872	6 slavery, men, slaves, day, land, freedom, life, time, slave, nation
27	0.020477	0.059336	27 land, cotton, acres, man, crop, year, day, house, farmer, money
2	0.000685	0.054678	2 people, men, problem, mob, crime, man, law, charge, case, country
21	0.010611	0.037959	21 women, children, world, men, mother, life, woman, race, freedom, ideals
36	0.000041	0.035320	36 men, world, industry, today, government, democracy, women, labor, people, race

Figure 10: Top 10 topic ordered by the non-biography column.The most common topics on the non-biography documents are mainly related to race, people, education, school, training.

genre	b	n	label
topic_id			
16	0.216132	0.023867	16 slave, slaves, time, man, master, slavery, day, home, work, children
26	0.139161	0.104694	26 school, people, work, time, education, life, men, way, day, man
15	0.076228	0.004660	15 men, state, time, man, slaves, arms, house, slavery, way, party
23	0.070928	0.002742	23 slavery, man, slave, men, people, time, life, state, country, friends
12	0.057909	0.081904	12 world, men, nan, man, life, eyes, thing, hands, face, soul
6	0.053921	0.075872	6 slavery, men, slaves, day, land, freedom, life, time, slave, nation
33	0.044017	0.006579	33 slaves, master, man, slave, slaveholders, school, time, holidays, slaveholder, masters
37	0.042701	0.006696	37 slave, time, mistress, slavery, book, master, words, nature, knowledge, slaves
28	0.041771	0.120800	28 race, people, man, education, years, men, life, races, country, time
7	0.041642	0.005882	7 mother, children, child, house, plantation, father, master, slavery, home, life

Figure 11: Top 10 topic ordered by the biography column.The most common topics on the biography documents are mainly related to slavery. it's mentioned in almost all the top topics along with the words man and master. Topics such as education, school and life are also some of the top topics in this category.

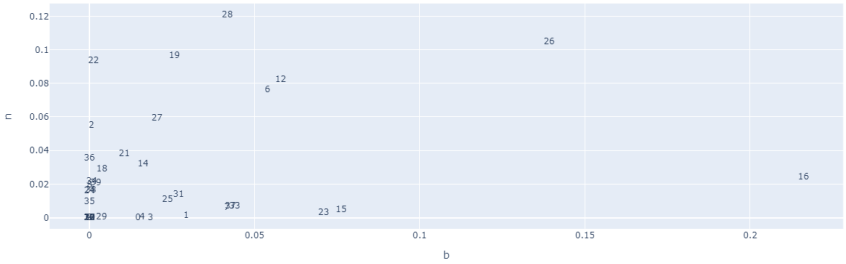

Figure 12: Scatter plot of top 40 topics on biography and non-boigraphy axis

There is predominance of gloomy words such as slavery and master (topic 19 and 15) in the biography documents which are replaced by brighter terms such as education, school and work in the non-biography documents. Can this be an indication of The hopeful outlook the authors have about the future of the country and black people in America despite their harsh circumstances they endured during their lives?

7 Word Embedding

This approach uses gensim's word2vec tool to represent words in a vector form. This helps to draw relationship between terms and surface important hidden insights about the corpus. The corpus is divided by authors to compare the findings among one another. The results from Booker T. Washington and Du Bois are shown below.

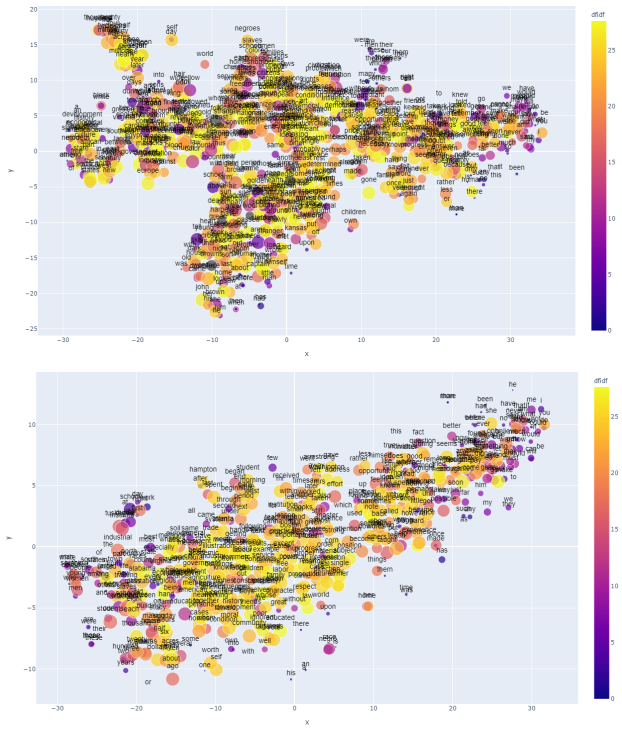

Figure 13: Word embedding for W.E.B. Du Bois (top) and Booker T. Washington (Bottom) corpus

There are a few interesting clusters in both cases. For example around (-20, 15) of the top diagram we find partially isolated cluster consisting of terms like, four, ten, hundred, thousand together with miles, acres, years, and age. This shows that measurable instances are clustered together with their different potential values. similar clusters also exist in the bottom diagram around the coordinate (-20,-7.5). there is also a cluster of terms such as man, white, black, coloured, young, women, men, southern. This may be due to people at that time being usually described using their color and gender attributes.

```
complete_analogy_f('slave', 'black', 'master', 3)
```

	term	sim
0	thomas	0.941147
1	after	0.938573
2	covey	0.937902

```
complete_analogy_w('slave', 'black', 'master', 3)
```

	term	sim
0	white	0.952731
1	southern	0.934256
2	man	0.922794

Figure 14: Top Three analogy for slave:black and master: from Fredrick Douglass(left) and Booker T. Washington(right) corpus.

From the right table we can see that if slave is to black master is to white, southern and man. The first word is the most expected outcome from this analogy but the other two are also interesting. especially the word man being analogous to master may be indicative of the common tradition of slave owners calling black men of all ages "boy" and refrain from acknowledging them as Men. on the left table for the same analogy we get the results: Thomas, after and Covey. These terms are not immediately noticeable for people who are not familiar with the corpus but upon further investigation I found that Thomas was the "master" of Fredrick Douglass. Covey was also a figure who temporarily take ownership of "trouble making" slaves and "break" them.

8 Sentiment Analysis

The other analysis performed on the corpus is sentiment analysis. I used NRC lexicon to detect 8 different emotions and polarity of the documents. This analysis is intended to make connections between the sentiment depicted in the different books with the sentiment the writers might have had about the situations they wrote about. Especially with the Autobiographies, does the different episodes of the writers' life (from slavery to becoming prominent figures) resemble with the different traces of emotions?

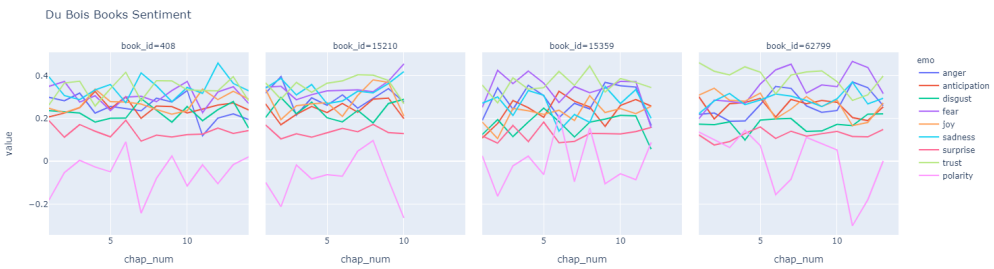

Figure 15: Sentiment and emotion Graph for Books By W.E.B.Du Bois. left to right (Souls of Black Folk, Darkwater, The Negro, John Brown)

In the figure below (left), the polarity for most of the works is negative. This is expected because the books in the corpus are mostly about slavery and other injustices in history. It was surprising to see that the highest positive polarity is observed on the book, "The story of slavery" by Booker T. Washington. The biographies are the books with the highest negative polarity. Among the eight different types of emotions Fear and anger are the predominant ones through our the corpus.

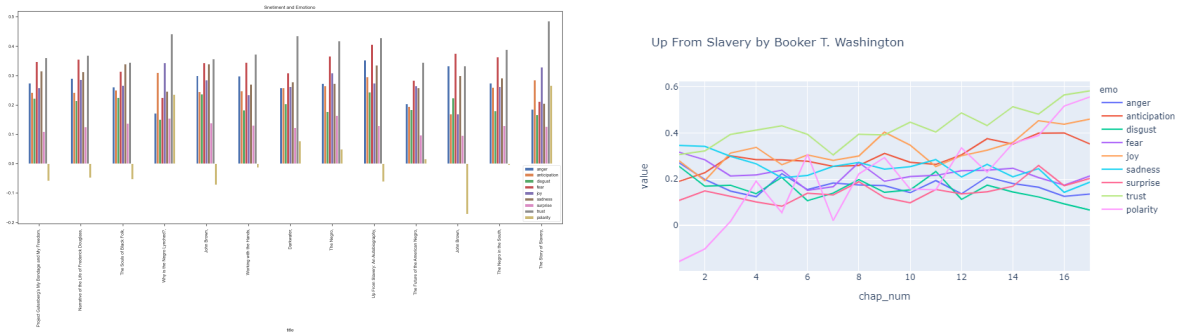

Figure 16: Sentiment bargraphs for each book (left) and Sentiment graph for the book "Up from Slavery by Booker T. Washington" (right)

Another interesting result is from the sentiment analysis of Booker T. Washington's autobiography, Up from slavery. If we see the polarity line it captures the essence of the life of the writer who was a slave and later become one of the most prominent public figures in America. The polarity line stars in the negative and slowly climbs up to positive values. Other emotions such as joy follow similar trends. When it comes to the negative emotions such as sadness, disgust and anger, they have a downward trend. This trend goes inline with the observations from the positive emotions

9 Summary

The above explorations revealed a number of interesting aspects of the life of the writers, the society they lived in and their outlook about the future. The term frequency analysis resembles with the main highlights of the authors' lives. The difference between the works of the the three authors was also explored. From the PCA we were able to see different principal components that distinguish the works based on author and genre. Topic models also reveal some interesting results about the difference between biography/autobiography and other works of the authors. Lastly, I tried sentiment analysis, especially on the autobiographies to see if the sentiments throughout the books can reveal something about the experiences of the authors.

10 Conclusion

The analysis performed through out this project gave us results that revealed interesting aspects of the books in the corpus, the styles of the authors and experiences the authors had. The work can be developed by including other works of the authors and it can be made even more representative by including other prominent African American writers. Including female African American writers can make the analysis even more interesting and help us see from another

unique perspective. Overall, exploratory text analysis is a powerful tool that can help us see into some hidden features of a text corpus. Employing this technique on such important corpus can lead to interesting results.

References

[1] Trent, Noelle Fredrick Douglass. In *Encyclopedia Britannica, 5 Apr. 2021, Accessed 10 May 2021.*

[2] Rudwick, Elliott. W.E.B. Du Bois. In *Encyclopedia Britannica, 23 Feb. 2021.*

[3] Britannica, The Editors of Encyclopaedia. Booker T. Washington. In *Encyclopedia Britannica, 6 May. 2021.*

[4] Project Gutenberg. In *https://www.gutenberg.org/*

[5] George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE, 2014.

[6] Frederick Douglass. Narrative of the life of Frederick Douglass, an American slave. In *ISBN: 9780486284996, May 1, 1845*

[7] W. E. B. Du Bois. Darkwater Voices From Within The Veil. In *https://www.gutenberg.org February 28, 2005 [EBook 15210]*

[8] Frederick Douglass. My bondage and my freedom .

[9] Frederick Douglass. Why is the Negro Lynched by Frederick Douglass.

[10] Frederick Douglass. John Brown.

[11] W.E.B.Du Bois. The Negro.

[12] W.E.B.Du Bois. John Brown.

[13] W.E.B.Du Bois. The Souls of Black Folks.

[14] Booker T. Washington. Working with the Hands.

[15] Booker T. Washington. Up from Slavery: An Autobiography.

[16] Booker T. Washington. The Future of the American Negro.

[17] Booker T. Washington. The Story of Slavery.

[18] Booker T. Washington and W.E.B.Du Bois. The Negro in the South by Booker T. Washington.