Milestone 2

Problem Description:

Device method for authorship attribution and compare them against state-of-the-art methods which can be supervised, semi supervised and unsupervised. Authorship attribution is basically the task of finding the author of a document. To achieve this purpose, one compares a query text with a model of the candidate author and determines the likelihood of the model for the query. A script from an unknown author is compared with all the authors' textual records for a match. We have a testing data set of 150 files and our aim is to determine whether they should be attributed to Daniel Defoe or not. According to various methods, the success rates dramatically change with different combinations, however, will take the best among them to compare with the new designed method.

Dataset Description:

Dataset was stored on Linux and then transferred to a Windows system. For the training, making use works of Daniel Defoe and his contemporaries such as Gutenberg collection.

For testing data set we have 150 text files with the same numbering system as Steig Hargevik's dissertation on Daniel Defoe's works. These 150 are just extracts from articles/books. Impeccable editions obtained from the various sources such as the British National Library, where Irving Rothman made a special trip to get 1st or earliest available editions of his works, since editors keep making changes over the years and the author's style may be obscured in this process.

Processing for 150 training files:

Reading all the files using Panda's library and performed the preprocessing steps. These steps include the removal the bullets, special characters, URL, and symbols from the text files.

Also removed the useless characters from starting and ending of the file. After performing these steps converted the files into a single data frame to process it further.

	File Name	Text
0	103D	<d-legio-m the memorial exami¹d> that o age ...
1	112D	remarks on the bill to prevent frauds commit...
2	113D	remark o the letter o the author o the state...
3	115D	this satyr had ever bee publishd, tho some o...
4	156D	advice to the electors o great britai; occas...
...	...	...
145	87D	cassandra answerd when a ma is big with his...
146	88D	to the kights, cttzzes ad burgesses i parlia...
147	92D	the writings of daniel defoe persecution ana...
148	95D	blackwell-hall a hint to the blackwell-hal...
149	99D	advice to all parties he that gives his advi...

150 rows × 2 columns

Calculated the length of text in each file:

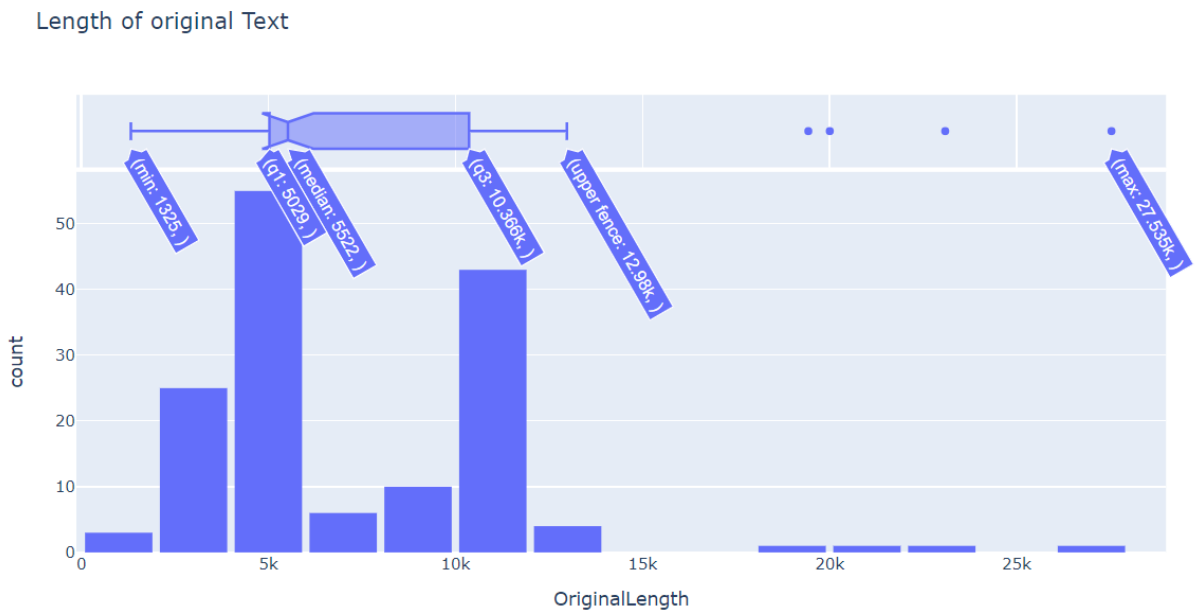
	File Name	Text	OriginalLength
0	103D	<d-legio-m the memorial exami¹d> that o age ...	5316
1	112D	remarks on the bill to prevent frauds commit...	5615
2	113D	remark o the letter o the author o the state...	11544
3	115D	this satyr had ever bee publishd, tho some o...	10315
4	156D	advice to the electors o great britai; occas...	1876

Calculated the statistics as per Original Length

```
test_df['OriginalLength'].describe()
```

```
count      150.000000
mean      7269.240000
std       4103.401267
min       1325.000000
25%       5029.500000
50%       5522.000000
75%      10361.250000
max      27535.000000
Name: OriginalLength, dtype: float64
```

Plotted the histogram using plotly for better understanding



Generated the word cloud:

Word Clouds are visual displays of text data – simple text analysis. Word Clouds display the most prominent or frequent words in a body of text (such as a State of the Union Address). Typically, a Word Cloud will ignore the most common words in the language.

Used the NLTK library to download the stop words and then removed the sop words from the data frame. Hence the length of each file got changed.

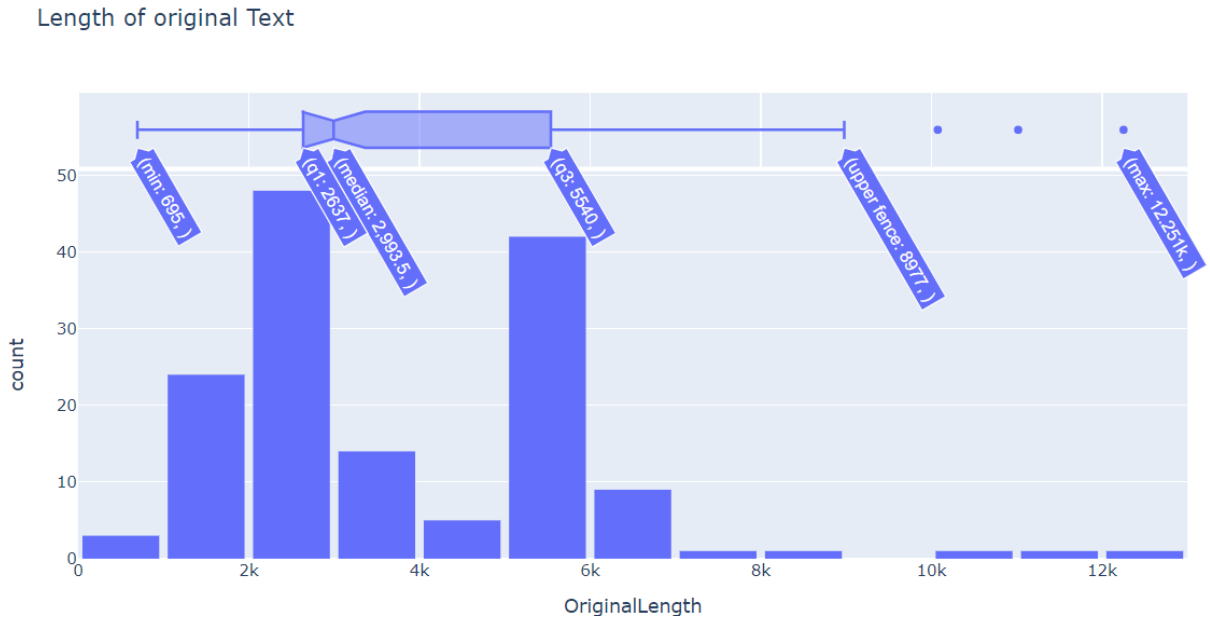
	File Name	Text	OriginalLength
0	103D	<d-legio-m memorial exami¹d> age ca parallel m...	2896
1	112D	remarks bill prevent frauds committed bankrupt...	3061
2	113D	remark letter author state-memorial stregth st...	6251
3	115D	satyr ever bee publishd, tho bee log time beig...	5452
4	156D	advice electors great britai; occasioed iteded...	977

Updated Statistics for Files:

```
test_df['OriginalLength'].describe()
```

```
count      150.000000
mean      3852.013333
std       2052.277981
min        695.000000
25%       2647.000000
50%       2993.500000
75%       5530.750000
max      12251.000000
Name: OriginalLength, dtype: float64
```

Updated Histogram:



Using the Gensim library created the Corpora and found that there are total of 1254 distinct words in all the 150 files after the stop word removal.

Things implemented in Milestone 3:

Topic Modelling:

Implemented topic modelling on all the 150 text files provided using LDA model from Gensim. It includes several steps:

- Stop word removal
- Tokenization

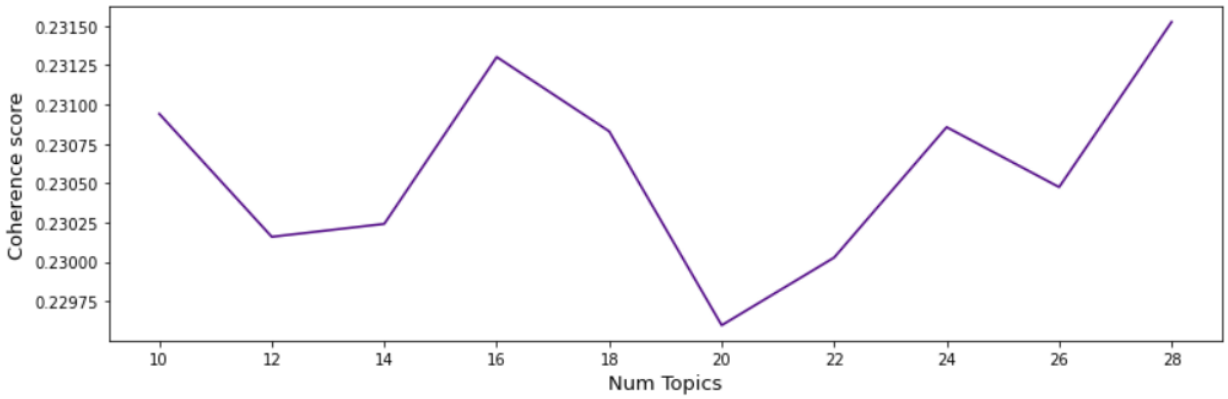
- Lemmatizing (Considered Lemmatization over stemming)
- Creating Dictionary, corpus, Term document frequency.

As a result, calculated the coherence score and perplexity which comes out to be.

```
Perplexity:  -8.269447631082693

Coherence Score:  0.2283013891483038
```

As number of topics increases the coherence score also increases after 20 topics. The plot clearly demonstrates the change of coherence score with respect to topics.




```
Num Topics = 10  has Coherence Value of 0.231
Num Topics = 12  has Coherence Value of 0.23
Num Topics = 14  has Coherence Value of 0.23
Num Topics = 16  has Coherence Value of 0.231
Num Topics = 18  has Coherence Value of 0.231
Num Topics = 20  has Coherence Value of 0.23
Num Topics = 22  has Coherence Value of 0.23
Num Topics = 24  has Coherence Value of 0.231
Num Topics = 26  has Coherence Value of 0.23
Num Topics = 28  has Coherence Value of 0.232
```

Used the below works for Model implementations from Gutenberg collection:

Works By Daniel Defoe:

1. The Life and Adventures of Robinson Crusoe by Daniel Defoe
2. The Further Adventures of Robinson Crusoe by Daniel Defoe

Works by Contemporaries of Daniel Defoe:

Author Name	Book
Alexander Pope	An Eassy on Criticism
Alfred Tennyson	The Princess
Elizabeth Barrett Browning	The Letters of Elizabeth Barrett Browning
Percy Bysshe Shelley	A Defence of Poetry and Other Essays
John Gay	The Beggers Opera

Implemented Classification Models

Model 1: Logistic Regression using TFIDF

Received an accuracy of 43%

Model 2: Multinomial Navie Bayes

Received an accuracy of 44%

Future Work:

Compare with stieg Hargevik's classification and improve the accuracy for the models created for milestone 3.

Related Work:

Different researchers have tried different machine learning algorithms for authorship attribution. Some of the main algorithms that I found in the papers that I consulted include K-nearest neighbors, Bayesian, Support Vector Machines (SVM), Feed Forward Multilayer Perceptron (MLP) and ensembles using combination of these algorithms.

In 2007 Bozkurt, Bağlıoğlu and Uyar found that 'Bag of Words' approach with SVM gave very high accuracy. In 2007 Stańczyk and Cyran used ANN and found that highest classification ratio is granted by the exploitation of syntactic textual features.

In 2014 Pratanwanich and Lio have used Supervised Author Topic (SAT) model that is based on probabilistic generative model and has exhibited same performance as Random Forests.

References:

[1] The Role of Linguistic Feature Categories in Authorship Verification, HossamAhmed, Procedia Computer Science, Volume 142, 2018, Pages 214-221.

[2] Code Authorship Attribution: Methods and Challenges, V Kalgutkar, R Kaur, H Gonzalez, ACM Computing Surveys Volume 52, 1 January 2020.

[3] Authorship attribution with thousands of candidate authors Moshe Koppel, Jonathan Schler, Shlomo Argamon, Eran Messeri , August 2006.

[4] Authorship Attribution Performance of various features and classification methods Đlker Nadi Bozkurt, Conference: Computer and information sciences, 2007.

[5] Shrestha, P., S. Sierra, F. Gonz´alez, P. Rosso, M. Montes-y G´omez, and T. Solorio (2017). Convolutional neural networks for authorship attribution of short texts. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain, 669–674.

[6] Brownlee, J. (2017). What are Word Embeddings for Text? https://machinelearningmastery.com/what-are-word-embeddings/.