DS5001: Exploratory Text Analytics (ETA) Final Project - Spring 2022

Manifest Document for Using Text Analytics to Explore the Works of Charles Dickens and Mark Twain

Cecily Wolfe (cew4pf)

Document Overview:
A description of each of the files found on this GitHub repository and/or on UVA Box. The cases in which files are hosted on only one or the other have been specified below. However, all of the elements necessary to recreate the original analysis are located on GitHub.

Provenance:
The 50 works of Charles Dickens (1812-1870) and 45 works of Mark Twain, i.e., Samuel Clemens (1835-1910), were compiled from Project Gutenberg, an online repository of ebooks in the public domain. The indices of the works for each author available on Project Gutenberg ([Dickens](https://www.gutenberg.org/ebooks/58157), [Twain](https://www.gutenberg.org/files/28803/28803-h/28803-h.htm)) were used to ensure that works were not duplicated and that, as needed, duplicated pieces in certain anthologies were removed from all sources except one.

Location:
Link to compressed source files on UVA Box: https://virginia.box.com/s/pkni5kmn9o8ngrz0qwfgqyu9arl45r31
Link to data tables (i.e., LIB, CORPUS, VOCAB, etc.) on UVA Box: https://virginia.box.com/s/ou44vr9j6t2i1uxxa8fogg814pxzcdqt

Description:
The general subject matter of the corpus is the works of Charles Dickens and Mark Twain, including novels, short stories, speeches, essays, letters, poetry, etc., with the purpose of comparing two of the most well-known and widely read authors in British and American literature. Although these two were not quite contemporaries, both had outsized influence on the literary profession during the lifetimes - that is, in the nineteenth century - and continue to enjoy widespread fame today. As such, the question as to whether or not the works of these authors share certain key attributes remains a relevant and interesting one.

Files:

cew4pf_DS_5001_Final_Paper.pdf: final report interpreting the results of the work completed with this corpus

Report_Notebook_PDFs.zip: compressed version of the final report and the PDF versions of Jupyter notebooks (i.e., cew4pf_DS_5001_Final_Paper.pdf and PDF_Files folder)

DS_5001_Project_Sources.pdf: complete list of sources (from class and online) used for coding and research purposes

Dickens_Twain_Corpus.zip: compressed version of the Dickens and Twain source files (i.e., the Dickens and Twain folders)

Dickens: folder containing 50 works of Dickens curated from [Project Gutenberg](https://www.gutenberg.org/ebooks/58157) in `.txt` format

Twain: folder containing 45 works of Twain curated from [Project Gutenberg](https://www.gutenberg.org/files/28803/28803-h/28803-h.htm) in `.txt` format

Notebooks: Jupyter notebooks for analysis
 * **dickens_preprocess.ipynb**: preprocessing Dickens works into LIB, CORPUS, VOCAB
 * **twain_preprocess.ipynb**: preprocessing Twain works into LIB, CORPUS, VOCAB
 * **dickens_analysis_M3-7.ipynb**: language models, word vector models, clustering, and PCA for Dickens works
 * **twain_analysis_M3-7.ipynb**: language models, word vector models, clustering, and PCA for Twain works
 * **full_analysis_M3-7.ipynb**: language models, word vector models, clustering, and PCA for all works (both Dickens and Twain)
 * **dickens_tmodel_wordem.ipynb**: topic models, word embeddings for Dickens
 * **twain_tmodel_wordem.ipynb**: topic models, word embeddings for Twain
 * **full_tmodel_wordem.ipynb**: topic models, word embeddings for all works
 * **sentiment_analysis.ipynb**: sentiment analysis for Dickens, Twain, and all works

ETA_Functions: python scripts with functions for text analysis
 * **bow_tfidf_pca.py**: functions to create BOW (Bag Of Words), TFIDF (Term Frequency Inverse Document Frequency) matrix, Principal Component Analysis (PCA) tables (LOADINGS, DCM (Document Count Matrix), COMPINF)
 * **hac2.py**: function to create HCA (Hierarchical Agglomerative Clustering) plots
 * **langmod.py**: functions to create a language model
 * **textparser.py**: functions to parse a list of works and associated regex patterns to process the works into chuncks by work, chapter, paragraph, sentence, token
 * **topicmodel.py**: function to create a topic model and produce the

associated tables (THETA, PHI, TOPIC)

HTML_Files: HTML versions of Jupyter notebooks

PDF_Files: PDF versions of Jupyter notebooks

salex: folder with CSVs containing various lexicons for sentiment analysis

Data_Tables: folder with CSVs generated from text analysis $ \rightarrow$ **NOTE THAT THIS IS HOSTED ON UVA BOX**

Dickens
* **dickens_pre_LIB.csv**: Dickens LIB table with work title, regex, year and decade published, number of chapters, etc.
* **dickens_pre_CORPUS.csv**: Dickens CORPUS table with OHCO book, chapter, paragraph, sentence, token and columns POS (Part of Speech), token_str, term_str
* **dickens_pre_VOCAB.csv**: Dickens VOCAB table with term_str as the index and cols term rank, number of characters, maximum part of speech, whether the term is a stop word, stems of the terms, etc.
* **dickens_BOW.csv**: Dickens BOW (Bag of Words)
* **dickens_TFIDF.csv**: Dickens TFIDF (Term Frequency Inverse Document Frequency)
* **dickens_LOADINGS.csv**: Dickens LOADINGS from PCA
* **dickens_DCM.csv**: Dickens DCM from PCA
* **dickens_COMPINF.csv**: Dickens COMPINF from PCA
* **dickens_THETA.csv**: Dickens THETA table from topic modeling
* **dickens_PHI.csv**: Dickens PHI table from topic modeling
* **dickens_TOPIC.csv**: Dickens TOPIC table from topic modeling
* **dickens_SENTS.csv**: Dickens sentences table from word embeddings
* **dickens_W2V.csv**: Dickens word2vec table from word embeddings
* **dickens_COORDS.csv**: Dickens COORDS table from word embeddings

Twain
* **twain_pre_LIB.csv**: Twain LIB table with work title, regex, year and decade published, number of chapters, etc.
* **twain_pre_CORPUS.csv**: Twain CORPUS table with OHCO book, chapter, paragraph, sentence, token and columns POS (Part of Speech), token_str, term_str
* **twain_pre_VOCAB.csv**: Twain VOCAB table with term_str as the index and cols term rank, number of characters, maximum part of speech, whether the term is a stop word, stems of the terms, etc.
* **twain_BOW.csv**: Twain BOW (Bag of Words)
* **twain_TFIDF.csv**: Twain TFIDF (Term Frequency Inverse Document Frequency)
* **twain_LOADINGS.csv**: Twain LOADINGS from PCA
* **twain_DCM.csv**: Twain DCM from PCA
* **twain_COMPINF.csv**: Twain COMPINF from PCA
* **twain_THETA.csv**: Twain THETA table from topic modeling

 * **twain_PHI.csv**: Twain PHI table from topic modeling
 * **twain_TOPIC.csv**: Twain TOPIC table from topic modeling
 * **Twain_SENTS.csv**: Twain sentences table from word embeddings
 * **twain_W2V.csv**: Twain word2vec table from word embeddings
 * **twain_COORDS.csv**: Twain COORDS table from word embeddings

Full Corpus
 * **full_LIB.csv**: full corpus LIB table with work title, regex, year and decade published, number of chapters, etc.
 * **full_CORPUS.csv**: full corpus CORPUS table with OHCO book, chapter, paragraph, sentence, token and columns POS (Part of Speech), token_str, term_str
 * **full_pre_VOCAB.csv**: full corpus VOCAB table with term_str as the index and cols term rank, number of characters, maximum part of speech, whether the term is a stop words, stems of the terms, etc.
 * **full_BOW.csv**: full corpus BOW (Bag of Words)
 * **full_TFIDF.csv**: full corpus TFIDF (Term Frequency Inverse Document Frequency)
 * **full_LOADINGS.csv**: full corpus LOADINGS from PCA
 * **full_DCM.csv**: full corpus DCM from PCA
 * **full_COMPINF.csv**: full corpus COMPINF from PCA
 * **full_THETA.csv**: full corpus THETA table from topic modeling
 * **full_PHI.csv**: full corpus PHI table from topic modeling
 * **full_TOPIC.csv**: full corpus TOPIC table from topic modeling
 * **full_SENTS.csv**: full corpus sentences table from word embeddings
 * **full_W2V.csv**: full corpus word2vec table from word embeddings
 * **full_COORDS.csv**: full corpus COORDS table from word embeddings
 * **full_COMBO.csv**: full corpus joined CORPUS and sentiment lexicon
 * **full_vocab_sentiment**.csv: full corpus joined VOCAB and sentiment lexicon
 * **full_books.csv**: full corpus sentiment by book
 * **full_SENT_SENTIMENT.csv**: full corpus sentiment by sentence using `VADER`