DS 5001: Exploratory Text Analytics – Final Project Sources

Cecily Wolfe (cew4pf)

Spring 2022

I. CODING SOURCES

Class Resources (courtesy of Professor Rafael Alvardo, PhD)

- M02_01_Importing-Persuasion.ipynb
- M02_02_TokenizingWithSciKitLearn.ipynb
- M03_02_LanguageModels.ipynb
- M03_03_Entropy-and-Perplexity.ipynb
- M03_04-Entropy-and-Term-Length.ipynb
- M04_00_NLTK_Intro.ipynb
- M04_01_Pipeline.ipynb
- M05_01_BOW_TFIDF.ipynb
- M06_01_SimilarityMeasures.ipynb
- M06_02_On_Clustering.ipynb
- M07_01_PCA.ipynb
- M08_02_LDASciKitLearn.ipynb
- M08_02a_LDASciKitLearn.ipynb
- M08_03_UseTopicModelLib.ipynb
- M08_03b_PrepNOVELS.ipynb
- M08_03c_PrepOKCUPID.ipynb
- M09_01_GloVe.ipynb
- M09_04_word2vec.ipynb
- M10_01_GeneralInquirer.ipynb
- M10_02_CombineLexicons.ipynb
- M10_03_Novels.ipynb
- M10_04_AustenMelville.ipynb
- SALEX lexicon
- langmod.py
- textparser.py (includes adaptations to fit specific corpus)
- hac2.py
- hw07.py (renamed bow_tfidf_pca.py includes adaptations to fit specific corpus)
- topicmodel.py

Online Sources

- **Text Parsing**
  - **Corpus background**
    - Wikipedia bibliography of Charles Dickens: https://en.wikipedia.org/wiki/Charles_Dickens_bibliography#Novels_and_novellas

  - *Miscellaneous Papers* publication approximate date: https://digitalcollections.nypl.org/items/30af3110-7ba8-0131-723a-58d385a7b928
  - Wikipedia bibliography of Mark Twain: https://en.wikipedia.org/wiki/Mark_Twain_bibliography
  - *Alonzo Fitz and Other Stories* approximate publication date: https://www.theatlantic.com/magazine/archive/1878/03/the-loves-of-alonzo-fitz-clarence-and-rosannah-ethelton/538638/
  - *In Defense of Harriet Shelley* approximate publication date: https://www.loc.gov/item/18010587/
- **Web scraping**
  - "Web Scraping Using BeautifulSoup" from *Surfing the Data Pipeline with Python* by Jonathan Kropko: https://jkropko.github.io/surfing-the-data-pipeline/ch5.html#id1
  - re.sub() to replace regex with given str: https://stackoverflow.com/questions/12453580/how-to-concatenate-items-in-a-list-to-a-single-string
  - Beautiful Soup Using Regex to Find Tags: https://stackoverflow.com/questions/24748445/beautiful-soup-using-regex-to-find-tags
  - Regex for links: https://stackoverflow.com/questions/11331982/how-to-remove-any-url-within-a-string-in-python
  - Regex to match text between square brackets: https://stackoverflow.com/questions/2403122/regular-expression-to-extract-text-between-square-brackets
  - Extract class name from tag beautifulsoup python: https://stackoverflow.com/questions/21592012/extract-class-name-from-tag-beautifulsoup-python
  - Need to add index when creating dataframe: https://stackoverflow.com/questions/17839973/constructing-pandas-dataframe-from-values-in-variables-gives-valueerror-if-usi
  - Using regex to search attributes of tags in html with BeautifulSoup: https://stackoverflow.com/questions/24748445/beautiful-soup-using-regex-to-find-tags
  - How to scrape multiple pages with BeautifulSoup: https://data36.com/scrape-multiple-web-pages-beautiful-soup-tutorial/
  - Position of tag in BeautifulSoup with .find() and .sourceline: https://www.skytowner.com/explore/getting_the_position_of_a_tag_in_beautiful_soup
- **Terminal commands to remove duplicate works ncluded in multiple anthologies**
  - Find lines containing a string in Linux with grep: https://stackoverflow.com/questions/11797730/how-to-find-lines-containing-a-string-in-linux
  - File split at a given line number: https://stackoverflow.com/questions/3066948/how-to-file-split-at-a-line-number
  - Delete specific line numbers from a text file with sed: https://stackoverflow.com/questions/2112469/delete-specific-line-numbers-from-a-text-file-using-sed
  - Get the line number while using grep: https://stackoverflow.com/questions/3213748/get-line-number-while-using-grep
  - Use grep to report back only line numbers: https://stackoverflow.com/questions/6958841/use-grep-to-report-back-only-line-numbers
  - Using sed to remove lines using numeric variables: https://unix.stackexchange.com/questions/462857/using-sed-with-an-integer-variable

- **Language Models, Vector Space Models, Similarity and Clustering, Principal Component Analysis (PCA)**
  - pandas.DataFrame.droplevel to drop one or more levels of a MultiIndex: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.droplevel.html
  - How to retrieve specific combinations of MultiIndex levels: https://stackoverflow.com/questions/52798386/pandas-dataframe-how-to-retrieve-specific-combinations-of-multiindex-levels
  - Accessing data in a MultiIndex: https://towardsdatascience.com/accessing-data-in-a-multiindex-dataframe-in-pandas-569e8767201d
  - sklearn.metrics.silhouette_score: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
  - Append a row as a list to a dataframe: https://sparkbyexamples.com/pandas/pandas-append-list-as-a-row-to-dataframe/
  - pandas background_gradient: https://pandas.pydata.org/docs/reference/api/pandas.io.formats.style.Styler.background_gradient.html
  - Deal with SettingWithCopyWarning in pandas: https://stackoverflow.com/questions/20625582/how-to-deal-with-settingwithcopywarning-in-pandas

- **Topic Modeling and Word Embeddings**
  - Dropping multiple columns by name starting with drop and loc: https://www.geeksforgeeks.org/how-to-drop-one-or-multiple-columns-in-pandas-dataframe/
  - Adding a new index level from the columns of a dataframe: https://stackoverflow.com/questions/14744068/prepend-a-level-to-a-pandas-multiindex
  - Setting pandas df column width with pd.set_option(display.max_colwidth', None) to prevent truncating column values: https://pandas.pydata.org/docs/user_guide/options.html
  - Reset width to default: https://pandas.pydata.org/docs/user_guide/options.html

- **Sentiment Analysis**
  - Sentiment analysis using VADER in nltk: https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664
  - How to fix matplotlib .title() TypeError: ‘Text’ object is not callable with .set_title(): https://techoverflow.net/2021/04/04/how-to-fix-matplotlib-title-typeerror-text-object-is-not-callable/
  - Géron, Aurélien, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media, 2019) (for plotting images)

II. Research Paper Sources

1. Collins, Philip. “Charles Dickens.” *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., 3 Feb. 2022, https://www.britannica.com/biography/Charles-Dickens-British-novelist. Accessed 28 April 2022.

2. Quirk, Thomas V. “Mark Twain” *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., 17 Apr. 2022, https://www.britannica.com/biography/Mark-Twain. Accessed 28 April 2022.

3. Gardner, Joseph H. “Mark Twain and Dickens.” *Publications of the Modern Language Association*, vol. 84, no. 1, pp. 90-101, 1969,

https://www.jstor.org/stable/1261160?saml_data=eyJzYW1sVG9rZW4iOiIzYzFjMTc3Ny0xYjk4LTRhNTUtOGIyNC1mOGRlNmY4ZjI5MTkiLCJlbWFpbCI6ImNldzRwZkB2aXJnaW5pYS5lZHUiLCJpbnN0aXR1dGlvbklkcyI6WyJmOGFmN2FjOS01MWVjLTQ0YzgtODFhOS0wNzlhNzQzMjgzOTMiXX0&seq=1. Accessed 28 April 2022.

4. Dawidziak, Mark. “Charles Dickens and Mark Twain: Separated at Birth?” *WordPress.com*, On the road with Gerald Dickens, https://geralddickens.wordpress.com/2021/03/12/guest-blog-mark-twain-and-charles-dickens/. Accessed 28 April 2022.

5. Yuan, Siyu. “A Comparative Analysis of Charles Dickens and Mark Twain.” *Academic Journal of Humanities and Social Sciences*, vol. 3, no. 7, 2016, https://francis-press.com/uploads/papers/JWqvK6RY0I3SK3KiHrC5SW6YIg0ns2yXWDDmHGqg.pdf. Accessed 28 April 2022.

6. Blair, Walter. “The French Revolution and ‘Huckleberry Finn.’” *Modern Philology*, vol. 55, no. 1, pp. 21-35, 1957, https://www.jstor.org/stable/pdf/435269.pdf?casa_token=xS_hf6CBXxEAAAAA:qwt0tRAzfBH9nxq4qLWziNndltHojhtwYWt9pWwndtwUTDAMo0MwDEbT3G3EjG24rgnJd4n9RhvIRWNNN_VdScR8MXTNHkamg6cNkU_TeOR3oUx4n_k. Accessed 28 April 2022.

7. Widger, David. “Index of the Project Gutenberg Works by Charles Dickens.” *Project Gutenberg*, Project Gutenberg, 2018. https://www.gutenberg.org/ebooks/58157. Accessed 28 April 2022.

8. Widger, David. “The Works of Mark Twain.” *Project Gutenberg*, Project Gutenberg, 2019. https://www.gutenberg.org/files/28803/28803-h/28803-h.htm. Accessed 28 April 2022.

9. dickens_analysis_M3-7.ipynb → section M03: Language Models, subsection: Trigram table

10. twain_analysis_M3-7.ipynb → section M03: Language Models, subsection: Trigram table

11. full_analysis_M3-7.ipynb → section M03: Language Models, subsection: Trigram table

12. “Maximum Tf Normalization.” *Maximum TF Normalization*, Cambridge University Press, 4 July 2009, https://nlp.stanford.edu/IR-book/html/htmledition/maximum-tf-normalization-1.html. Accessed 28 April 2022.

13. dickens_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Top 20 nouns by DFIDF, sorted in descending order (including plural nouns but not proper nouns)

14. twain_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Top 20 nouns by DFIDF, sorted in descending order (including plural nouns but not proper nouns)

15. full_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Top 20 nouns by DFIDF, sorted in descending order (including plural nouns but not proper nouns)

16. dickens_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Hierarchical agglomerative cluster diagrams for the distance measures

17. twain_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Hierarchical agglomerative cluster diagrams for the distance measures

18. full_analysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: Hierarchical agglomerative cluster diagrams for the distance measures

19. full_ananlysis_M3-7.ipynb → section: M06: Similarity and Clustering, subsection: K-Means

20. full_ananlysis_M3-7.ipynb → section: M07: Principal Component Analysis, subsection: Manual PCA Methods with Only 10000 Most Significant Terms (excluding proper nouns)

21. full_ananlysis_M3-7.ipynb → section: M07: Principal Component Analysis, subsection: Prince PCA with Outliers Removed

22. dickens_tmodel_wordem.ipynb → section: M08: Topic Models, subsection: Works and Top Terms Associated with Each Topic

23. twain_tmodel_wordem.ipynb → section: M08: Topic Models, subsection: Works and Top Terms Associated with Each Topic

24. full_tmodel_wordem.ipynb → section: M08: Topic Models, subsection: Works and Top Terms Associated with Each Topic

25. full_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Noun tSNE plot

26. dickens_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Noun tSNE plot

27. twain_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Noun tSNE plot

28. full_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Similarities

29. dickens_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Similarities

30. twain_tmodel_wordem.ipynb → section: M09: Word Embeddings, subsection: Similarities

31. sentiment_analysis.ipynb → section: Sentiment by Book

32. Gardner, Joseph. *Dickens in America: Twain, Howells, James, and Norris*. E-book, Routledge, 1988, https://books.google.com/books?id=2UdnDwAAQBAJ&pg=PT77&lpg=PT77&dq=twain%27s+most+similar+book+to+dickens&source=bl&ots=iesmzfcUWo&sig=ACfU3U3OtqVOreabkL4HHt1BE_Lt2tEvWA&hl=en&sa=X&ved=2ahUKEwjqlIqrvab3AhX4knIEHeyEDgsQ6AF6BAg2EAM#v=onepage&q=twain's%20most%20similar%20book%20to%20dickens&f=false. Accessed 28 April 2022.

33. Chesterton, G.K. *Martin Chuzzlewit - Introduction by G.K. Chesterton*, American Literature, https://americanliterature.com/author/charles-dickens/book/martin-chuzzlewit/introduction-by-gk-chesterton. Accessed 28 April 2022.

34. Burke, Jackson. “History of Scams: Nothing New under the Sun.” *CNBC*, CNBC, 17 Feb. 2015, https://www.cnbc.com/2015/02/17/scams-hacking-spanish-prisoner.html. Accessed 28 April 2022.

35. Messent, Peter. “The American Claimant Review.” *Goodreads*, Goodreads, https://www.goodreads.com/book/show/2010710.The_American_Claimant. Accessed 28 April 2022.