DS 5001: Exploratory Text Analytics
Maureen O'Shea (mo2cr@virginia.edu)
06 May 2022
Text Analysis of Shakespeare Plays

Manifest

The data consist of 37 Shakespeare plays downloaded as XML files from The Folger Shakespeare Library. The Folger Library is the world's largest Shakespeare collection and includes resources such as an API that provides extracted information for each play. (The Folger Shakespeare. n.d.) This analysis focuses on the spoken text in Shakespeare's plays. A description of the Jupyter notebooks, source files, and output data files are included in this document. The original author of all Jupyter notebook files is R. C. Alvarado. (https://github.com/ontoligent/DS5001-2022-01) The notebooks were altered for this analysis.

Collection: The Folger Shakespeare Play collection in XML and TXT was downloaded from https://shakespeare.folger.edu/download-the-folger-shakespeare-complete-set/

Source files and Output files:

- folger/data: data source and output files
- folger/lib: python functions
- folger/XML: Shakespeare Plays XML Format
- folger/TXT: Shakespeare Plays TXT Format

https://virginia.box.com/s/9xwfwl4x75mqsf22fs25otwdvfkk1605

Tools:

- lxml.etree, scikit-learn, nltk, topicmodel, genism, vader

ETA Model Pipeline

0 - Source Documents and Metadata

Notebook ID: 00-FolgerAPI.ipynb
Title: Convert Folger XML format to CSV
Description: Source documents and metadata about 37 Shakespearean plays. Register a LIB table with play title and play code and an API table to hold API functions.
Data Source:

- 37 names and play codes for Folger XML formatted plays from https://shakespeare.folger.edu/download-the-folger-shakespeare-complete-set/
- API Functions https://shakespeare.folger.edu/the-folger-shakespeare-api/

Register Data:

- LIB: index = ['play_code'], columns = ['play_title', 'play_id']
- API: index = ['func_id'], columns = ['func_key', 'func_desc']

Data Output:

- folger-LIB.csv
- folger-API.csv

1 - Standard Machine Learning Corpus Format

Notebook ID: 01-Folger2CSV.ipynb
Title: A Client for the Folger API
Description: Import and parse XML files. Establish OHCO and register a TOKEN table with extracted token strings, lemma and part of speech annotation.
Data Source:

- folger-LIB.csv
- XML formatted Shakespearean plays (37)

Register Data:

- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- TOKEN: index=[OHCO], columns = ['token_str', 'lemma', 'pos']

Data Output:

- folger-TOKEN.csv

2 & 3 - Data Model: Documents, Tokens, Terms, Labels, add NLP annotations

Notebook ID: 02-FolgerF2andF3Pipeline.ipynb
Title: Pipeline for LIB, TOKEN, VOCAB and CORPUS
Description: Import a collection of texts and convert to F2. Then we annotate the collection to create an F3-level model. Created LIB, TOKEN, VOCAB and CORPUS tables.
Data Source:

- folger-LIB.csv
- folger-TOKEN2.csv

Register Data:

- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- LIB: index = [‘play_code’], columns = ['play_title', 'year', 'source_file_path', 'genre', 'play_id', 'play_len', 'n_acts', 'n_scenes', 'n_speeches', 'label']
- VOCAB: index =['term_str'], columns = ['n', 'n_chars', 'p', 'i', 'h', 's', 'max_pos', 'n_pos', 'cat_pos', 'stop']
- CORPUS/TOKEN2: index = [OCHO], columns = ['token_str', 'lemma', 'pos', 'term_str']

Data Output:

- folger-LIB2.csv
- folger-VOCAB.csv
- folger-CORPUS.csv
- folger-TOKEN2.csv

Notebook ID: 03-FolgerExploration
Title: Entropy and Term Length
Description: Explores the relationship between term length and entropy in corpus
Data Source:

- folger-LIB2.csv
- folger-VOCAB.csv
- folger-TOKEN2.csv

Register Data:

- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- LIB: index = [‘play_code’], columns = ['play_title', 'year', 'source_file_path', 'genre', 'play_id', 'play_len', 'n_acts', 'n_scenes', 'n_speeches', 'label']
- VOCAB: index =['term_str'], columns = ['n', 'n_chars', 'p', 'i', 'h', 's', 'max_pos', 'n_pos', 'cat_pos', 'stop']
- TOKEN2: index = [OCHO], columns = ['token_str', 'lemma', 'pos', 'term_str']

Data Output:

- None

4 - Vector Space Models

Notebook ID: 04-FolgerVectorizationTFIDF.ipynb
Title: Vectorization, TFIDF and BOW
Description: Vectorize corpus with SciKit Learn and create TFIDF and BOW. Register a BOW at OHCO level of ACTS with “English” stop words and proper nouns removed. Register a VOCAB2 table of 4000 words.

Data Source:

- folger-LIB2.csv
- folger-VOCAB.csv
- folger-CORPUS.csv

Register Data:

- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- BAG OHCO = ['play_code', 'act_num', 'term_str']
- BOW: index = [BAG_OHCO], columns =['n', 'tfidf']
- VOCAB2: index = ['term_str'], columns=['term_rank', 'n', 'n_chars', 'n_tokens', 'tfidf_mean', 'df', 'dfidf', 'max_pos', 'n_pos', 'cat_pos', 'term_rank2']
- CORPUS2: index = [OHCO], columns=['token_str', 'pos', 'term_str']

Data Output:

- folger-BOW.csv: BOW created at OHCO[:2] level. Stopwords and Proper nouns removed
- folger-VOCAB2.csv: Reduced VOCAB Table (4000 rows)
- folger-CORPUS2.csv

5 & 6 - Models and Visualizations

Model-1: **STADM with Analytical Models**
Notebook ID: 05-FolgerSimilarityMeasures.ipynb
Title: Similarity and Distance Measures
Description: Compute cosine similarity with ward clustering and distance measures by working with a larger BOW (OHCO[:1]). K-means clustering is included.

Data Source:

- folger-BOW.csv
- folger-VOCAB2.csv
- folger-LIB2.csv

Register Data:

- BOW_REDUCED_OHCO:['play_code', 'term_str']
- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- BOW_REDUCED: index =[BOW_REDUCED_OHCO], columns = ['tfidf', 'binary', 'tfidf_l1', 'tfidf_l2']

Data Output:

- folger-BOW_REDUCED.csv: BOW grouped by OHCO[:1] to visualize clusters

Model-2: **PCA**

Notebook ID: 06-FolgerPCA.ipynb

Title: PCA with Interactive Visualization

Description: Implement PCA from scratch using eigendecomposition of the term covariance matrix and explore. Create DCM scatterplots grouped by play and genre. Explore components by play with bar charts and dendrograms (cosine and euclidean distance metric).

Data Source:

- folger-BOW.csv
- folger-VOCAB2.csv
- folger-LIB2.csv

Register Data:

- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- BAG OHCO = ['play_code', 'act_num', 'term_str']
- PCA_COMPS: index=[0, 1, 2, 3, 4, 5], columns=['pos', 'neg', 'eig_val', 'exp_var']
- PCA_THETA: index=['term_str'], columns = [0, 1, 2, 3, 4, 5]
- PCA_DCM: index = ['play_code', 'act_num'], columns=[0, 1, 2, 3, 4, 5]
- PCA_DOC: index=['play_code', 'act_num'], columns=['play_title', 'genre', 'label', 'year', 'mean_tfidf', 'n_tokens', 0, 1, 2, 3, 4, 5]

Data Output:

- folger-PCA_COMPS.csv: COMP Table
- folger-PCA_THETA.csv: LOADINGS/TERM-COMP
- folger-PCA_DCM.csv: DOC-COMP Matrix
- Folger-PCA_DOC.csv: DOC Table with PCA

Model-3: **Topic Model using TopicModel**

Notebook ID: 07-FolgerUseTopicModelLib.ipynb

Title: Using TopicModel with Interactive Visualization

Description: Generate Topic Model for Corpus using TopicModel. Explore topics by play, genre and year.

Data Source:

- folger-BOW.csv
- folger-VOCAB2.csv
- folger-LIB2.csv

- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- BAG OHCO = ['play_code', 'act_num', 'term_str']

Data Output:

- None

Model-4: **LDA Topic Models with Sci-Kit Learn**

Notebook ID: 08-FolgerLDASciKitLearn.ipynb

Title: LDA Topic Models with Sci-Kit Learn and Interactive Visualization

Description: Create LDA Topic Models for corpus and separate topic models for genres within corpus using Sci-Kit Learn wrapped in TopicModel. Visualize mean topics per play from the overall topic model. Visualize mean topics per play using genre topic models. Create topic dendrograms. Create heat maps of topics within acts of plays using overall topic model and genre topic models.

Data Source:

- folger-BOW.csv
- folger-VOCAB2.csv
- folger-LIB2.csv

Register Data:

- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- BAG OHCO = ['play_code', 'act_num', 'term_str']
- LDA_TOPIC*: index = [RangeIndex(start=0, stop=20, step=1, name='topic_id')], columns=['phi_sum', 'theta_sum', 'h', 'top_terms_rel', 'top_terms', 'label']
- LDA_THETA*: index=['play_code', 'act_num'], columns=[RangeIndex(start=0, stop=20, step=1, name='topic_id')]
- LDA_PHI*: index=[RangeIndex(start=0, stop=20, step=1, name='topic_id')], columns=[`term_str`]

Data Output:

Topic model for corpus

- folger-LDA_TOPIC-20.csv
- folger-LDA_THETA-20.csv
- folger-LDA_PHI-20.csv

Topic model for History genre

- folger-LDA_TOPIC_HISTORY-20.csv
- folger-LDA_THETA_HISTORY-20.csv
- folger-LDA_PHI_HISTORY-20.csv

Topic Model for Tragedy Genre

- folger-LDA_TOPIC_TRAGEDY-20.csv
- folger-LDA_THETA_TRAGEDY-20.csv
- folger-LDA_PHI_TRAGEDY-20.csv

Topic Model for Comedy Genre

- folger-LDA_TOPIC_COMEDY-20.csv
- folger-LDA_THETA_COMEDY-20.csv
- folger-LDA_PHI_COMEDY-20.csv

Topic Model for Romance Genre

- folger-LDA_TOPIC_ROMANCE-20.csv
- folger-LDA_THETA_ROMANCE-20.csv

- folger-LDA_PHI_ROMANCE-20.csv

Model-5: **Word Embeddings**
Notebook ID: 09-FolgerWordEmbeddingword2vec.ipynb
Title: Word2Vec Word Embeddings using word2vec with Interactive Visualization
Description: Create word embeddings of play data using word2vec, perform semantic algebra and visualize results with tSNE.
Data Source:
- folger-TOKEN2.csv

Register Data:
- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- BAG OHCO = ['play_code', 'act_num', 'term_str']
- W2V: index=['term_str'], columns=[RangeIndex(start=0, stop=100, step=1)]
- W2V_VOCAB: index=['term_str'], columns=['n', 'pos_max', 'pos_group', 'df', 'dfidf']
- GENSIM_DOCS: list of sentences for GENSIM

Data Output:
- folger-W2V.csv
- folger-W2V_VOCAB.csv
- folger-GENSIM_DOCS.csv

Model-6: **Sentiment Analysis**
Notebook ID: 10-FolgerSentiment.ipynb
Title: Sentiment Analysis of Plays
Description: Use the NRC lexicon and VADER to explore sentiment in plays and in genres.
Data Source:
- salex_csv
- emo_cols = "anger anticipation disgust fear joy sadness surprise trust polarity"
- folger-CORPUS2.csv
- folger-LIB2.csv

Register Data:
- OHCO: ['play_code', 'act_num', 'scene_num', 'speech_id', 'speaker', 'line_num']
- VADER_DOC: index =[OHCO], columns= ['anger', 'anticipation', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'trust', 'polarity', 'html_str', 'sent_str', 'vader_neg', 'vader_neu', 'vader_pos', 'vader_compound']

Data Output:
- folger-VADER_DOC.csv

References:

Folger Shakespeare Library. (n.d.) *Shakespeare's Plays, Sonnets and Poems* from The Folger Shakespeare. Retrieved from https://shakespeare.folger.edu