Final Project Notebook

DS 5001 Text as Data | Spring 2025

Metadata

- Full Name:
- Userid:
- GitHub Repo URL:
- UVA Box URL:

Overview

The goal of the final project is for you to create a **digital analytical edition** of a corpus using the tools, practices, and perspectives you've learning in this course. You will select a corpus that has already been digitized and transcribed, parse that into an F-compliant set of tables, and then generate and visualize the results of a series of fitted models. You will also draw some tentative conclusions regarding the linguistic, cultural, psychological, or historical features represented by your corpus. The point of the exercise is to have you work with a corpus through the entire pipeline from ingestion to interpretation.

Specifically, you will acquire a collection of long-form texts and perform the following operations:

- **Convert** the collection from their source formats (F0) into a set of tables that conform to the Standard Text Analytic Data Model (F2).
- **Annotate** these tables with statistical and linguistic features using NLP libraries such as NLTK (F3).
- **Produce** a vector representation of the corpus to generate TFIDF values to add to the TOKEN (aka CORPUS) and VOCAB tables (F4).
- **Model** the annotated and vectorized model with tables and features derived from the application of unsupervised methods, including PCA, LDA, and word2vec (F5).
- **Explore** your results using statistical and visual methods.
- **Present** conclusions about patterns observed in the corpus by means of these operations.

When you are finished, you will make the results of your work available in GitHub (for code) and UVA Box (for data). You will submit to Gradescope (via Canvas) a PDF version of a Jupyter notebook that contains the information listed below.

Some Details

- Please fill out your answers in each task below by editing the markdown cell.
- Replace text that asks you to insert something with the thing, i.e. replace `(INSERT IMAGE HERE)` with an image element, e.g. `![](image.png)`.
- For URLs, just paste the raw URL directly into the text area. Don't worry about providing link labels using `[label](link)`.
- Please do not alter the structure of the document or cell, i.e. the bulleted lists.
- You may add explanatory paragraphs below the bulleted lists.
- Please name your tables as they are named in each task below.
- Tasks are indicated by headers with point values in parentheses.

Raw Data

Source Description (1)

Provide a brief description of your source material, including its provenance and content. Tell us where you found it and what kind of content it contains.

(INSERT DESCRIPTION HERE)

Source Features (1)

Add values for the following items. (Do this for all following bulleted lists.)

- Source URL:
- UVA Box URL:
- Number of raw documents:
- Total size of raw documents (e.g. in MB):
- File format(s), e.g. XML, plaintext, etc.:

Source Document Structure (1)

Provide a brief description of the internal structure of each document. That, describe the typical elements found in document and their relation to each other. For example, a corpus of letters might be described as having a date, an addressee, a salutation, a set of content paragraphs, and closing. If they are various structures, state that.

(INSERT DESCRIPTION HERE)

Parsed and Annotated Data

Parse the raw data into the three core tables of your addition: the `LIB`, `CORPUS`, and `VOCAB` tables.

These tables will be stored as CSV files with header rows.

You may consider using `|` as a delimitter.

Provide the following information for each.

LIB (2)

The source documents the corpus comprises. These may be books, plays, newspaper articles, abstracts, blog posts, etc.

Note that these are *not* documents in the sense used to describe a bag-of-words representation of a text, e.g. chapter.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:
- Number of observations:
- List of features, including at least three that may be used for model summarization (e.g. date, author, etc.):
- Average length of each document in characters:

CORPUS (2)

The sequence of word tokens in the corpus, indexed by their location in the corpus and document structures.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:
- Number of observations Between (should be >= 500,000 and <= 2,000,000 observations.):
- OHCO Structure (as delimitted column names):
- Columns (as delimitted column names, including `token_str`, `term_str`, `pos`, and `pos_group`):

VOCAB (2)

The unique word types (terms) in the corpus.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:
- Number of observations:
- Columns (as delimitted names, including `n`, `p`', `i`, `dfidf`, `porter_stem`, `max_pos` and `max_pos_group`, `stop`):
- Note: Your VOCAB may contain ngrams. If so, add a feature for `ngram_length`.
- List the top 20 significant words in the corpus by DFIDF.

(INSERT LIST HERE)

Derived Tables

BOW (3)

A bag-of-words representation of the CORPUS.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:
- Bag (expressed in terms of OHCO levels):
- Number of observations:
- Columns (as delimitted names, including `n`, `tfidf`):

DTM (3)

A represenation of the BOW as a sparse count matrix.

- UVA Box URL:
- UVA Box URL of BOW used to generate (if applicable):
- GitHub URL for notebook used to create:
- Delimitter:
- Bag (expressed in terms of OHCO levels):

TFIDF (3)

A Document-Term matrix with TFIDF values.

- UVA Box URL:
- UVA Box URL of DTM or BOW used to create:

- GitHub URL for notebook used to create:
- Delimitter:
- Description of TFIDIF formula ($\LaTeX$ OK):

Reduced and Normalized TFIDF_L2 (3)

A Document-Term matrix with L2 normalized TFIDF values.

- UVA Box URL:
- UVA Box URL of source TFIDF table:
- GitHub URL for notebook used to create:
- Delimitter:
- Number of features (i.e. significant words):
- Principle of significant word selection:

Models

PCA Components (4)

- UVA Box URL:
- UVA Box URL of the source TFIDF_L2 table:
- GitHub URL for notebook used to create:
- Delimitter:
- Number of components:
- Library used to generate:
- Top 5 positive terms for first component:
- Top 5 negative terms for second component:

PCA DCM (4)

The document-component matrix generated.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:

PCA Loadings (4)

The component-term matrix generated.

- UVA Box URL:

- GitHub URL for notebook used to create:
- Delimitter:

PCA Visualization 1 (4)

Include a scatterplot of documents in the space created by the first two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)

(INSERT IMAGE HERE)

(INSERT IMAGE HERE)

Briefly describe the nature of the polarity you see in the first component:

(INSERT DESCRIPTION HERE)

PCA Visualization 2 (4)

Include a scatterplot of documents in the space created by the second two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)

(INSERT IMAGE HERE)

(INSERT IMAGE HERE)

Briefly describe the nature of the polarity you see in the second component:

(INSERT DESCRIPTION HERE)

LDA TOPIC (4)

- UVA Box URL:
- UVA Box URL of count matrix used to create:
- GitHub URL for notebook used to create:
- Delimitter:
- Libary used to compute:
- A description of any filtering, e.g. POS (Nouns and Verbs only):
- Number of components:

- Any other parameters used:
- Top 5 words and best-guess labels for topic five topics by mean document weight:
  - T00:
  - T01:
  - T02:
  - T03:
  - T04:

LDA THETA (4)

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:

LDA PHI (4)

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:

LDA + PCA Visualization (4)

Apply PCA to the PHI table and plot the topics in the space opened by the first two components.

Size the points based on the mean document weight of each topic (using the THETA table).

Color the points basd on a metadata feature from the LIB table.

Provide a brief interpretation of what you see.

(INSERT IMAGE HERE)

(INSERT INTERPRETATION HERE)

Sentiment VOCAB_SENT (4)

Sentiment values associated with a subset of the VOCAB from a curated sentiment lexicon.

- UVA Box URL:
- UVA Box URL for source lexicon:

- GitHub URL for notebook used to create:
- Delimitter:

Sentiment BOW_SENT (4)

Sentiment values from VOCAB_SENT mapped onto BOW.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:

Sentiment DOC_SENT (4)

Computed sentiment per bag computed from BOW_SENT.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:
- Document bag expressed in terms of OHCO levels:

Sentiment Plot (4)

Plot sentiment over some metric space, such as time.

If you don't have a metric metadata features, plot sentiment over a feature of your choice.

You may use a bar chart or a line graph.

(INSERT IMAGE HERE)

VOCAB_W2V (4)

A table of word2vec features associated with terms in the VOCAB table.

- UVA Box URL:
- GitHub URL for notebook used to create:
- Delimitter:
- Document bag expressed in terms of OHCO levels:
- Number of features generated:
- The library used to generate the embeddings:

Word2vec tSNE Plot (4)

Plot word embedding featues in two-dimensions using t-SNE.

Describe a cluster in the plot that captures your attention.

(INSERT IMAGE HERE)

(INSERT DESCRIPTION HERE)

Riffs

Provde at least three visualizations that combine the preceding model data in interesting ways.

These should provide insight into how features in the LIB table are related.

The nature of this relationship is left open to you -- it may be correlation, or mutual information, or something less well defined.

In doing so, consider the following visualization types:

- Hierarchical cluster diagrams
- Heatmaps
- Scatter plots
- KDE plots
- Dispersion plots
- t-SNE plots
- etc.

Riff 1 (5)

(INSERT IMAGE HERE)

(INSERT INTERPRETATION HERE)

Riff 2 (5)

(INSERT IMAGE HERE)

(INSERT INTERPRETATION HERE)

Riff 3 (5)

(INSERT IMAGE HERE)

(INSERT INTERPRETATION HERE)

Interpretation (4)

Describe something interesting about your corpus that you discovered during the process of completing this assignment.

At a minumum, use 250 words, but you may use more. You may also add images if you'd like.

(INSERT INTERPRETATION HERE)

In [ ]: