Final Project

100 Possible Points

5/2/2023

Attempt 1

IN PROGRESS
Next Up: Submit Assignment

Add Comment

Unlimited Attempts Allowed
2/28/2023 to 5/5/2023

Details

Overview

The goal of your final project is to apply what you have learned in this course to create a digital analytical edition of a corpus that will support exploration of the social, historical, or cultural contents of that corpus. These contents are broadly conceived—they may be about language use, social events, cultural categories, sentiments, identity, taste, etc., and these may be described synchronically or diachronically, i.e. as structures or as trends over time.

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.

Deliverables

To receive full credit for the assignment, you will produce a digital analytical edition of a corpus, which will include a written report and be hosted on a dedicated GitHub repository.

This edition should include the following deliverables.

Data Files

(https://canvas.its.virginia.edu/courses/58239/modules/items/349279)

Submit Assignment

A collection of **data files**, each in CSV format, containing the F2 through F5 data you extracted from the corpus. These files should include, at a minimum, the following core tables:

- `LIB.csv` — Metadata for the source files.
- `CORPUS.csv` — This is a tokens table annotated with statistical and linguistic features, such as TFIDF. It should include and index that represents the OHCO of the documents in your corpus.
- `VOCAB.csv` — Annotated with statistical and linguistic features, such as DFIDF.

In addition, you should include the following data sets, either as features in the appropriate core table or as separate tables. Note that all tables should have an appropriate index and, where appropriate, an OCHO index.

Principal Components (PCA)

- Table of documents and components.
- Table of components and word counts (i.e., the "loadings"), either added to the VOCAB table or as a separate table with a shared index with the VOCAB table.

Topic Models (LDA)

- Table of document and topic concentrations.
- Table of topics and term counts, either added to the VOCAB table or as a separate table with a shared index with the VOCAB table.

Word Embeddings (word2vec)

- Terms and embeddings, either added to the VOCAB table or as a separate table with a shared index with the VOCAB table.

Sentiment Analysis

- Sentiment and emotion values as features in VOCAB or as a separate table with a shared index with the VOCAB table.
- Sentiment polarity and emotions for each document.

Code Files

The **Jupyter notebooks** used to perform all operations that produced the data in your tables.

Any Jupyter notebooks used to **explore** and **visualize** the data in preparation for your final report.

Any **Python files** (e.g., `.py` files) you wrote to support your work.

(https://canvas.its.virginia.edu/courses/58239/modules/items/349279)

Submit Assignment

Report Document

A Jupyter notebook called `FINAL_REPORT.ipynb` describing your work and interpreting its results along with links to all the files listed above. This report should be written using Markdown text cells and embedded graphics from your other notebooks to illustrate points. Do not reference images that are not listed in the notebook. You may use images to show images in the notebook if you don't want to include the code there. Include citations for any references made in the notebook.

This notebook should contain the following four sections:

1. **Introduction**. Describe the nature of your corpus and the question(s) you've asked of the data.
2. **Source Data**. Provide a description of all relativant source files and describe the following features for each source file:
    1. *Provenance*: Where did they come from? Describe the website or other source and provide relevant URLs.
    2. *Location*: Provide a link to the source files in UVA Box.
    3. *Description*: What is the general subject matter of the corpus? How many observations are there? What is the average document length?
    4. *Format*: A description of both the file formats of the source files, e.g., plaintext, XML, CSV, etc., and the internal structure where applicable. For - example, if XML then specify document type (e.g., TEI or XHTML).
3. **Data Model**. Describe the analytical tables you generated in the process of tokenization, annotation, and analysis of your corpus. You provide a list of tables with field names and their definition, along with URLs to each associated CSV file.
4. **Exploration**. Describe each of your explorations, such as PCA and topic models. For each, include the relevant parameters and hyperparemeters used to generate each model and visualization. For your visualizations, you should use at least three (but likely more) of the following visualization types:
    - Hierarchical cluster diagrams
    - Heatmaps showing correlations
    - Scatter plots
    - KDE plots
    - Dispersion plots
    - t-SNE plots
5. **Interpretation**. Provide your interpretation of the results of exploration, and any conclusion if you

›

(https://canvas.its.virginia.edu/courses/58239/modules/items/349279)

Submit Assignment

Format and Style

Any non-data files you produce, such as a Jupyter notebook or a Python program, should contain a header stating your name and email address, the name of this class (DS 5001), and the date. It should look something like this (depending on the document):

Rafael Alvarado (rca2t@virginia.edu) DS 5001 Spring 2023

Jupyter notebooks should be properly outlined with headers and explanatory text where necessary to follow what is happening.

Submission

All of your content, except for any data files too large to upload to GitHub, should be in a dedicated repo that will be linked to in your homework submission in Canvas.

About Group Work

Students may work in groups. Ideally, these may be composed of up to three students. In these groups, students may collaborate on the work of acquiring, consolidating, and modifying the source data. In addition, students in groups are free to share code and ideas. However, each student is responsible for their own deliverables. In addition, the observations made in the final report must be unique to each student. There may be some overlap of ideas, but students must demonstrate that they have engaged individually with the material by writing up their own conclusions and expressing them in their own language.

Rubric

Given that the focus of this course is on method and not domain knowledge per se (although we have covered a good bit of that), you may be liberal in your interpretations. That is, do not worry about whether they will meet high scholarly and scientific standards. Remember, the purpose of ETA is to open up texts so that you may explore them and extract possibly significant patterns from them. However, you are expected to present your conclusions in a coherent and compelling manner. And, if you do find that you have discovered something interesting about your data beyond the requirements of the assignment, by all means consider pursuing it beyond this course.

Grading is divided evenly among the quality of your deliverables -- the completeness of your data files, the manifest, etc. -- and that of your report.

Item	Percent	Criteria

Submit Assignment

(https://canvas.its.virginia.edu/courses/58239/modules/items/349279)

Item	Percent	Criteria
Final Report	50%	Full points if the essay makes use of the analyses to provide a reasonable interpretation of the data. Points will be taken off anything that detracts from you presentation, such as incomplete sentences and poor graphics.

Appendix 1: Example Projects

> Note that these projects do not conform to the current instructions. They are provided here to give you a sense of the kinds of questions and answers you may give to your data and get from your exploratory analyses.

- https://www.dropbox.com/s/6zwh70eehj90fyl/EXAMPLE01.pdf?dl=0 (https://www.dropbox.com/s/6zwh70eehj90fyl/EXAMPLE01.pdf?dl=0)
- https://www.dropbox.com/s/qcf9x9zcz20g4di/EXAMPLE02.pdf?dl=0 (https://www.dropbox.com/s/qcf9x9zcz20g4di/EXAMPLE02.pdf?dl=0)
- https://www.dropbox.com/s/y1m4t1qaorckvh3/EXAMPLE03.pdf?dl=0 (https://www.dropbox.com/s/y1m4t1qaorckvh3/EXAMPLE03.pdf?dl=0)
- https://www.dropbox.com/s/dklx30f1o5niuk2/EXAMPLE04.pdf?dl=0 (https://www.dropbox.com/s/dklx30f1o5niuk2/EXAMPLE04.pdf?dl=0)
- https://www.dropbox.com/s/qa7fp2318wuccc0/EXAMPLE05.pdf?dl=0 (https://www.dropbox.com/s/qa7fp2318wuccc0/EXAMPLE05.pdf?dl=0)
- https://www.dropbox.com/s/q32cxz91s1i32jl/EXAMPLE06.pdf?dl=0 (https://www.dropbox.com/s/q32cxz91s1i32jl/EXAMPLE06.pdf?dl=0)

Appendix 2: Some Data Sources

- https://virginia.box.com/s/bj8f1khrkfd6thm9umq35m6xp2an4zej (https://virginia.box.com/s/bj8f1khrkfd6thm9umq35m6xp2an4zej)
- https://docs.google.com/document/d/1--qMm_7gmSSExV8vocvljwlAr1jHjErsuVBngEugwF4/edit (https://docs.google.com/document/d/1--qMm_7gmSSExV8vocvljwlAr1jHjErsuVBngEugwF4/edit)

Appendix 3: Forms of Text Data

Form Level	Description
F0	Source Format. The initial source format of a text, which varies by collection, e.g. XML (e.g. TEI and RSS), HTML, plain text (e.g. Gutenberg), JSON, and CSV.
F1	Machine Learning Corpus Format (MLCF). Ideally a table of minimum discursive units indexed by document content hierarchy.

›

(https://canvas.its.virginia.edu/courses/58239/modules/items/349279)

Submit Assignment

Form Level	Description
F3	NLP Annotated STADM. STADM with annotations added to token and term records indicating stopwords, parts-of-speech, stems and lemmas, named entities, grammatical dependencies, sentiments, etc.
F4	STADM with Vector Space models. Vector space representations of TOKEN data and resulting statistical data, such as term frequency and TFIDF.
F5	STADM with analytical models. STADM with columns and tables added for outputs of fitting and transforming models with the data.
F6	STADM converted into interactive visualization. STADM represented as a database-driven application with interactive visualization, .e.g. Jupyter notebooks and web applications.

Enter Web URL

http://

(https://canvas.its.virginia.edu/courses/58239/modules/items/349279)

Submit Assignment