Course Assignment for Text Analysis Course

Max Planck Summer School 2017

August 4, 2017

Provide code, answers, output, and explanations, as appropriate. Your answers here can be submited instead of the written exam. If you are taking the written exam, you can do this assignment in small groups (up to three people). A passing grade will complete at least 50% of questions in each section.

1 Corpora

Download a corpus of documents with some associated social/behavioral/economic metadata.

1. Describe why you chose this corpus and discuss the research question(s) one could ask with the resulting features, X.
   (a) Identify outcome variable(s) Y that can be predicted with your features: $Y = f(X)$.
   (b) Identify treatment variable(s) Z that might have an impact on your text features: $X = g(Z)$.
   (c) Identify a grouping (e.g., time t) that could have interesting patterns.
2. Describe how you will organize and label the documents, and outline how you will merge them with the associated metadata.
3. Explain what counts as a "document" in your corpus, and how you will split documents up.
4. Inspect your corpus and explain what cleaning steps are needed to remove uninformative material.

2 Featurization

1. Split the corpus into sentences.

2. Split sentences into tokens and tag parts of speech.

3. Discuss whether these features are important in your corpus or should you remove them: capitalization, punctuation, stopwords, numbers, rare words, word stems, short (one-letter or two-letter) words, very long words, symbols (e.g. currency symbols).

4. Compute number of sentences, words, and letters for each document. Can you check for any interesting patterns across time (or across some other grouping)?

5. Compute number of nouns, verbs, and adjectives in each document. Check for any patterns over time and interpret.

6. Produce document frequency and term frequency counts over tokens (after removing capitalization, punctuation, and word stems, if desired). Produce summary statistics on most-frequent and least-frequent words, to decide on a minimum (and possible maximum) threshold for inclusion of tokens. Make a word cloud of the 100 most frequent words, weighted by frequency.

7. Are there any single words whose frequency across documents might be interesting or informative? Plot those over time (or other grouping) and check for interesting patterns.

8. Choose two categories from LIWC and plot the relative frequencies over time. Discuss any interesting patterns.

9. Create n-grams up to length 3.

10. Create POS-tagged N-grams up to length 3 and filter out words except nouns, verbs, adjectives, and adverbs.

11. Create document frequencies, term frequencies, relative frequencies, and TF-IDF frequencies for your preferred set of N-grams. Report summary statistics on the highest-frequency features and discuss. Are there big differences in the top features across the frequency metrics? Discuss which you think will be most useful for your research question. Make a word cloud of the 100 most frequent bigrams and trigrams, weighted by frequency.

12. Are there any phrases whose frequency across documents might be interesting or informative? Plot those over time (or other grouping) and check for interesting patterns.

13. Run a syntactic dependency parser on your corpus. What are the most frequent subjects, direct objects, subject-verb pairs, and subject-verb-object tuples in your corpus?

14. Are there any dependencies, subject-verb-object structures, entity relations, etc. that you could extract that you expect to be informative about or predictive for your research question? Compute these frequencies, plot over time, and discuss.

3 Topic Models

1. Train LDA on your corpus with 12 topics.
2. Produce word clouds for the topics.
3. Experiment with the following feature variants: 1) filter on nouns, adjectives, and verbs; 2) bigrams and trigrams; 3) bigrams and trigrams, but filtered on nouns, adjectives, and verbs, 4) syntactic dependencies: subject-verbs and subject-verb-objects.
4. Experiment with number of topics to find a topic count that results in intuitive/interpretable topics.

4 Word Embeddings

1. Train Word2Vec on your corpus, once with window=2 and once with window=8. What do you expect to change for the different window sizes?
2. Report sample output on most-similar words, and on analogies, that your corpus may have specific informativeness about. Compare to the similarity scores using the pre-trained word vectors in spaCy.
3. Discuss how you might use a word or set of words as a supervised classifier in your corpus.
4. Use the spaCy word vectors to implement the WEAT test for a pair of social groups.

5 Corpus Similarity

1. Discuss how you might use an auxiliary corpus to rank your documents along some interesting metric.
2. Divide your documents up by time period, and compute TF-IDF similarity between the documents in a year. Take the average similarity in each year. Plot the similarity over time.
3. Use k-means clustering to assign your documents to clusters of about 100 documents per cluster.
4. Use k-means clustering to assign your word embeddings to clusters of about 30 words per cluster.

6 Regression

1. Write out a linear regression model for a real-valued Y variable you would like to predict with your text features. Describe the terms and how you would interpret the coefficients.
2. Discuss why the text features might be predictive of Y.
3. Discuss what options you have for residualizing out fixed effects.
4. Produce t-statistics for univariate OLS regressions of x^k on Y, with or without fixed effects. Make a histogram of t-statistics. Make a scatter plot of the t-statistics, x-axis without FE's, y-axis with FE's.
5. Produce word clouds for most predictive features, separately for positive and negative t-statistics.
6. Create a new set of predictors with t-statistic above 2 or below -2. What is the change in number of predictors?
7. Predict Y with your full set of text features. Compare out-of-sample prediction performance for elastic net, partial least squares, and random forests.
8. Regress your LDA topic proportions on the outcome and discuss the strongest correlations.
9. Optional, for R users: train a structural topic model and report summary statistics.
10. Isolate the vector direction in Word2Vec space that is most predictive of your outcome. Discuss how you could use this in your research.

7 Classification

1. Identify a multi-valued un-ordered label in your metadata. Predict this label using logistic regression, random forest classifier, and xgboost. Report accuracy and AUC for the held-out test sample.

8 Causal Inference Methods

1. Identify a source of (arguably) exogenous variation in your outcome, your text features, or both.
2. Look at the impact of a treatment on an outcome Y, as well as the effect on text-predicted outcome $\hat{Y}$. Use time and group fixed effects as available.