Spring 2018
Iowa State University

LING 410X-Language as Data

Final Project Example Ideas

Submission Deadlines: (Total: 25% of your grade)

1. Forming groups + writing initial report: 7 April 2018 (5%)
2. Classroom Presentations: April 24, 26 (5%)
3. Code and report submission: May 3 (15%)
   (report up to 5 pages long, no double spacing)

Instructions: Here are some final projects you can consider working on in groups of 2 people. You are encouraged to come up with alternate ideas if you want, but I should approve them before you start on them. The project carries 25% of your grade. I expect to see three deliverables related to this project and the deadlines are mentioned above. Your project will be evaluated holistically -i.e., beating existing solutions is not the sole criteria to get a good grade. So, do not panic if your well-executed new idea turns out in the end to be not good as you thought it will be. If you already have a domain specific dataset and some problem to be solved in mind, please start expanding on those ideas! I split the ideas into four categories: corpus analysis, text classification, topic modeling, and visualization

Idea 1: Twitter analytics

keywords: visualization, corpus analysis, topic modeling
Description: Analyze and visualize the trends in twitter for a topic of your choice based on geographic location and keywords, and create visualizations that you think are relevant to talk about your dataset. There are several other interesting things you can mine from twitter. Read a bit and figure out what you want to work on otherwise.

Idea 2: Topic modeling with newspapers or research articles or blogs (or any such data you can find)

keywords: topic modeling, corpus analysis
Description: Pick a newspaper of your choice (nytimes, guardian are good choices, as we discussed about them), analyze somewhere between 200-500

articles from recent past (figure out how to get these!) and create topic models and their visualizations. Alternatively, you can also pick research articles from your discipline over a certain span of time or Literary pieces from Gutenberg etc.

Idea 3: Sentiment classification on movie reviews

Keywords: text classification
Description: Have a look at this website and the data it provides: `https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data` Your task is to use the dataset and develop a text classification approach to classify movie reviews into positive (scores 3,4 in the data), neutral (2) and negative (0,1). You should explore multiple approaches, evaluate and compare them and present your findings. Visualizations are not mandatory, but recommended. This is not the only text classification dataset out there. You can come up with anything relevant to your area of research (or talk to me to figure out) and work on text classification and then visualizing the models.

Idea 4: Stylistic analysis

Keywords: corpus analysis, visualization
Description: Take any two favorite authors of yours whose works are accessible online. Compare their writings in terms of their writing style, and take a third author who you believe has a similar writing style to one of these authors and analyze the similarity. You can make use of Stylo R library as a starting point, use different visualizations it supports, along with ngram analysis.

Other ideas

You can take the basic question as one of: text classification, clustering, topic modeling; and work on corpus analyses and visualizations for any textual dataset you have access to.