Carmen Easterwood

Tuesday 6:30pm PST

**Project Proposal**

**What:** Classify news articles according to topic

**Why:** Classifying news articles by topic helps us understand what journalists are writing about and can make searches more efficient. However, manual topic coding is quite time-consuming, so automating this process can save a lot of time and money for news organizations.

**Dataset:** I am using the New York Times Annotated Corpus, which contains 1.8 million articles from 1987-2007, labeled with topics, subtopics, and newspaper sections. These articles come in individual XML files, so I have written a Python program that selects N random files, parses those files, and saves the results to a CSV file that I can import to a Jupyter notebook. I will experiment with how large I can make N before I run into memory problems in the notebook.

**Algorithms:** References #1 and #2 have good information on preprocessing steps for news articles, such as lemmatization, using nouns only, using the lead paragraph, or using the headline. I plan to test all these preprocessing methods. However, both of those papers are dealing with unsupervised learning (and some WordNet, which I want to avoid). References #3 and #4 do supervised topic classification for political texts, using CNN, SVM, Naïve Bayes, and logistic regression (one-vs-one and one-vs-rest). I plan to use Naïve Bayes as a simple baseline and compare it to more complicated algorithms.

**References:**

1. Fiona Martin and Mark Johnson. 2015. More Efficient Topic Modelling Through a Noun Only Approach. <http://www.aclweb.org/anthology/U15-1013>.
2. Stefan Wermter and Chihli Hung. 2002. Selforganizing classification on the Reuters news corpus. <http://www.aclweb.org/anthology/C02-1087>.
3. Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-Lingual Classification of Topics in Political Texts. <http://www.aclweb.org/anthology/W17-2906>.
4. Mladen Karan, Jan Šnajder, Daniela Širinić, and Goran Glavaš. 2016. Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts. <http://www.aclweb.org/anthology/W16-2102>.