Frameworks Final Project

Goal

Our goal is to use existing data on NPM Frameworks, including npmjs.com download counts and dates, github.com issues and comments, and stackoverflow.com questions and answers, to answer a question of our own. We are seeking a mathematical method of determining the most effective dimensionality K of an LDA topic model.

Data

We scraped data regarding NPM Frameworks from a variety of sources. From GitHub (github.com), we scraped Issues data with their respective comments, retaining dates of original issue posting. From Stack Overflow (stackoverflow.com), we scraped questions and answers. From NPM (npmjs.com) we scraped downloads of frameworks.

Predictions from Modeling Actual Data

Our primary modeling technique is topic modeling with latent Dirichlet allocation (LDA). Our proposal indicates running topic models of the GitHub Issues and comments data, as well as the Stack Overflow Q&A, with models of a variety of topic dimensions. We plan to run models of 10-100 topics (by tens), to compare the predictive effectiveness of the various dimensions of the models.

Using the topic frequencies per framework-month as features, we plan to predict growth in popularity of frameworks in a long short-term memory (LSTM) recurrent neural net time series model. The effectiveness of growth prediction within our given data set is the measure of the effectiveness of the topic model for a given dimensionality K.

Modeling of Simulated Corpuses

We also plan to create several corpuses of simulated data, consisting solely of 0s and 1s. These will be simulated using probability distributions described in the hypothetical creation of corpuses by latent Dirichlet allocation. The steps include:

1. Pre-determine the number of topics (10-100) to model;
2. The number of words for each document is selected from a Poisson distribution;
3. A vector (length K) of probabilities per topic for each document is selected from a Dirichlet distribution, of parameter alpha;
4. Within each document, the nth word is selected using a process:
   1. Select a topic Zn from multinomial(1, φ);
   2. Select a Vx1 vector ø from Dirichlet(ϐZn), where V is the size of the vocabulary;
   3. Select a word wn from Multinomial(1, ø)

As with the modeled actual data, we will use the simulated data to compare models of different dimensionality.

Here, the dimensionality K is predetermined when we create the corpus. We will compare models of varying K using existing methods of model effectiveness, including perplexity, change in perplexity, and Goodness of Fit (GOF). We will compare the degree to which these methods select for the predetermined K in the corpus.

Completed Work and Obstacles

Thus far, we ran a single topic model on a sample of our GitHub Issues and comments data. We used 300 Issues with their respective comments from each Framework (i.e., N=1800); we did not separate these Issues by month for time series analysis yet.

We were unable to model the full data (N = 40,333) due to freezing of the R software and the desktop OS X computer running the analysis.

Plans for Future Work

We plan to continue our proposal as planned during spring semester, as a project for COSC 594 Evidence Engineering. To overcome the R limitations, we plan to use python for topic modeling. This will include dividing data into training and testing sets so that predictive effectiveness may be measured.