EE126 Project Proposal

Don Han, Chris Jeng, Giulio Zhou

Overview

For our project, we will apply PageRank to create a question answering program. In order to gather data from the optimal sources, we run PageRank across Reddit and Quora and then generate an answer using Markov chains, where Markov chain distributions from higher ranked pages will be more highly weighted.

Problem Formulation

First, let's discuss how to formulate the Reddit and Quora entries for importance ranking via PageRank. Recall that PageRank operates by approximating the invariant distribution of a Markov Chain by performing a random walk with random restarts. How do we choose the nodes in our Markov Chain? In Quora, users ask questions, which are in turn answered by other users (and upvoted as well). Users who are highly upvoted are considered to be more important, especially if they are upvoted by other important users. Additionally, the importance of each question and its corresponding answers will be weighted by their similarity to the provided question (determined using similarity via a bag of words or other NLP technique).

Upvotes in Reddit work differently in that they are anonymous. Using a slightly different model that considers friend relations, we will also crawl Reddit for useful answers.

After attaching weights to our data sources, we will generate answers from the text using a Markov Chain sentence model.