Topic Summarization

PURE Spring 2015 Mid-Semester Progress Report

Mentees: Kevin K. Lin & Harshay R. Shah
Mentor: Ismini Lourentzou
Motivation

Topic modeling facilitates textual analysis in today’s explosion of untapped text-based information.

Most of today’s topic algorithms model text using lists of keywords. Although concise, keyword lists lose a great deal of information crucial to a more meaningful understanding.

Thus, our research hopes to improve the outcome of topic modeling techniques to preserve document semantics while also providing audiences with human-comprehensible descriptions of topics.
Strategy

Since linguistic meaning relies so heavily on the relation between words, as opposed to just the words themselves, our approach accounts for the word order when creating a directed weighted graph.

Such a model will allow us to chain different keywords together. Through a bit of theory and experimentation, we can potentially develop models that produce accurate, relevant, and complete phrases, as well as explore similarities and differences between two topics.
Progress

To get up to speed, we familiarized ourselves with a wide range of topics including graphical modeling, latent semantic indexing, and topical phrase mining. We still have several more readings to review with our PURE mentor.

We are also reviewing the implementation for creating graphs from documents and determining weights from topic distributions before we can start pushing code to GitHub.
Next Steps

In order to compare different topic models, we need to implement graph similarity analysis, which will likely involve ranking and matching algorithms.

We then need to implement walks on these graphs so that we can appropriately chain together keywords for phrase mining.

We can use these generated phrases in the final poster presentations and create a possible conference submission afterwards.