PROJECT PROPOSAL: TEXT MINING TOOLKIT FOR DIGITAL HUMANITIES

OLGA SCRIVNER, IRINA TRAPIDO, JAY LEE
OLGAS2@ILLINOIS.EDU,TRAPIDO2@ILLINOIS.EDU,LEE9@ILLINOIS.EDU

1. Goals

Access to large digitized collections present a new opportunity to digital humanities researchers. At the same time, traditional methodology (e.g. close reading or keyword search) becomes very inefficient for literary data analysis. On the other hand, most available mining tools are built for different purposes and often require programming skills (even for software installation). Our project addresses these needs by building a Shiny web application ITMS that would allow researchers to explore visually and interactively their data collection. In addition, this application does not require any installation. While the core of this tool has been built, a lot of tuning and evaluation is needed, similar to META tool, in which the user is able to select parameters. Finally, the current tool uses various topic modeling algorithms, each with a different parameter setting, thus providing their evaluation and guidance to a non-technical user is necessary.

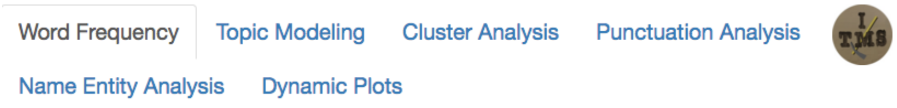

FIGURE 1. Interactive Text Mining Suite - Design

2. Research Outlines

2.1. **Functionalities.** i) interactive text pre-processing, stopwords selection, stemming, ii) interactive data selection, iii) frequency analysis, iv) topic modeling and model evaluation, v) cluster analysis, vi) name entities relation network

2.2. **Existing Tools.** Voyant and Zotero pluggin Paper-Machine. Voyant does not provide a topic modeling and cluster analyses. Paper-Machine is based on Zotero reference manager and is mainly built for bibliography. In both tools users cannot tune analyses.

2.3. **Users.** Researchers in Digital Humanities and Linguistics fields

2.4. **Techniques and Resources.** We will implement the following techniques: clustering, topic modeling, name entity network modeling. These techniques will be performed by using several text mining packages and topic modeling packages in R, such as *tm*, *lda*, *topicmodels*, *NLP*, and *stm*.

3. Timeline of the Project

Nov 7 Parameters for topic modeling: adding more parameter options to existing functions

Nov 14 Interactive data selection by chapters or other segments (term and abstract extraction is already implemented): creating a function that takes users input and segments data accordingly

Nov 21 Cluster analysis: tuning and parameters evaluation. Review literature on clustering visualization.

Nov 28 Topic modeling evaluation. Review literature on various types of evaluation (e.g. best K topic, best parameters etc). Currently, only one evaluation is included, namely a best K topic analysis based on Log-likelihood ratio. This evaluation suggests how many topics users should choose.