Course Project Report
Text Clustering/Classification Plugin for Web Server

Team: Abot, botelho3@illinois.edu

Progress made thus far

As of writing this report the first couple of tasks have been completed or are in progress. The dataset has been ingested and cleaned. The basic webserver setup and build scripts have been completed. Some basic web pages are being served but need to be hooked up to the dataset views. Preliminary research on the Clustering Algorithm has been done based on the Literature Review. The Gensim library will be used to build the corpus and tokenize/filter the dataset. Gensim's built in topic modeling algorithms will be evaluated, if they do not achieve the expected level of accuracy Labelled Latent Dirichlet Analysis (LLDA) may be used. Documentation hasn't been started and will be done at the end of the project.

Task	**Hours Required**	**Progress (Hrs)**	**Finished**
Collection + Cleaning of Dataset	6	6	Y
Design of Classification or Clustering Algorithm	10	2	N
Coding of Plugin/Framework Implementing Algorithm in Flask	10	3	N
Designing basic web-interface to Demo results on Example dataset.	6	3	N
Documentation	2	0	N

Remaining Tasks

Task	Hours Required	Progress (Hrs)	Remaining
Design of Classification or Clustering Algorithm	10	2	8
Coding of Plugin/Framework Implementing Algorithm in Flask	10	3	7
Designing basic web-interface to Demo results on Example dataset.	6	3	3
Documentation	2	0	2

Any challenges/issues being faced

A few risks have been identified at this point, related to the dataset quality and topic modeling algorithm.

The dataset initially proposed as part of the project is a Kaggle dataset of recipe reviews. This dataset has been cleaned and ingested and is easy to work with in Pandas + Gensim. Unfortunately the volume of text in the recipe instructions and recipe description may not be detailed enough to produce high-quality clustering via LSA/LDA. If the desired result can't be achieved it may be possible to use a more advanced topic modeling algorithm e.g. Labelled Latent Dirichlet Analysis (LLDA). Alternatively the topic modeling algorithm could be applied to a dataset with a higher volume of text e.g. the BBC News Article dataset. This will require more work as the data ingestion component of the project will need to be rewritten.

The topic modeling algorithm research has been done. Ideally LSA or LDA will be sufficient. Libraries exist implementing both algorithms that appear to be performant. However, based on the research, topic modeling algorithms on large corpuses can be computationally and memory intensive. Frameworks claim to implement tunable versions in which the hidden variables, and topic matrices can be computed with a fixed amount of resources. If not due to the hardware limitations of the developers in the group a creative solution may need to be found to reduce the system requirements. There are a number of workarounds but they all come with a penalty of time and complexity. For example comput constraints could be removed by renting compute time from Amazon Web Services or DigitalOcean, the corpus could be randomly sampled or batched to reduce the size of the matrices, or the libraries implemented LSA/LDA can be re-written or customized for better performance.