Project Specification

### Project Description and Methodology

This project, Hidden Topic Discovery Using Latent Semantic Analysis and k-means Clustering is aimed at extracting hidden topics from documents, so that the term-document, term-term and document-document relationships can be presented. Some further operations like document clustering and article recommendation system can be implemented based on this mathematic model.

Basically the topic model can be built using LSA, which is based on Singular Value Decomposition, then the documents are clustered using K-means algorithm. After training the mathematic model, it will be tested using new data sets (an approach to add new data to this model should also be developed).

### Project Tasks and Milestones

Week 1~4

The first part is data acquisition. Abundant data (documents) is going to be collected in this stage, then a document – term matrix is constructed. This stage requires automation tools like crawler and word counting programs. In addition, to use other people’s data sets, XML parsers (written in python) would be needed. -This stage may need 4 weeks.

Week 5~8

The second part is implement LSA and clustering upon the data sets obtained in 1st state, then the performance of the model would be analyzed. In this part, the main task is doing SVD and choosing proper dimension to improve the performance of clustering. This stage may take 4 weeks.

Week 9~14

The third part would be investigating more advanced approached to figure out how the SVD model can be improved (better performance, more rational explanation of the model). Then improved models will be compared with the original model.

Week 15+

#### Project Deliverables

By the end of this project, an improved version of clustering algorithm may be obtained which includes some more advanced features than traditional LSA and K-means clustering. For example, supervised clustering can inform the machine of the theme of articles, so human’s work can be reduced. A more robust model can reduce the impact of noise such as synonymous, homonymy and useless words.