Course : CSC 630 Independent Study Proposal
Name: Nisarg Gandhi
Unity ID: ndgandh2
Research Supervisor: Prof. Nagiza Samatova

Project Title: Text Analytics using Word Representation Learning Techniques

Project Description:
The goal of the project is to generate meaningful analytics based on the monthly notes provided by Laboratory of Analytic Sciences using various Natural Language Processing techniques like:

- Named Entity Recognition
- Automatic Summarization
- Relationship Extraction
- Topic Segmentation & Recognition
- Information Extraction
- Machine Learning

The main objective of this project is to build a user profile based on the monthly notes written by the user and provide recommendations on next project, team, tasks, etc. for each user. For achieving this goal we intend to use machine learning algorithms like *doc2vec* and *word2vec* which are based on *neural networks*, commonly referred to as Deep Learning. Using large amounts of unannotated plain text, *word2vec* learns relationships between words automatically. The output are vectors, one vector per word, with remarkable linear relationships that allow us to do things like:

- vec("king") – vec("man") + vec("woman") =~ vec("queen"), or
- vec("Montreal Canadiens") – vec("Montreal") + vec("Toronto")resembles the vector for "Toronto Maple Leafs".

word2vec optimizes through its async stochastic gradient backpropagation algorithm, and neatly connected it to the well-established field of matrix factorizations. In short, word2vec ultimately learns word vectors and word context vectors. These can be viewed as two 2D matrices (of floats), of size #words * #dim each.

We will be exploring various word2vec models, including the original continuous bag-of-word (CBOW) and skip-gram models, as well as advanced tricks, hierarchical soft-max and negative sampling.