L3S Research Center
Temporal Information Retrieval, WiSe 2016-17
Dr. Avishek Anand
Programming 4, due: **09 January 2017**

Please complete the programming assignment using Java. If you are extremely unfamiliar with Java you may use another language that has a Lucene port. The evaluation script will be in Java and requires you to have the latest JVM installed.

Problem 1.
Build word embeddings for the text corpus used in the previous assignments. Generate embeddings for different parameter settings and evlauate against the most expressive model. The most expressive model is 500 dimensions for the hidden layer with a window size of 5. Take the top 10 words as your ground truth for any given word. Then use this to calculate precision@5, recall, ndcg@10 and MAP metrics for various models with different window sizes and hidden dimensions. (For ndcg, assume the top 5 are very relevant and the next 5 are slightly relevant. The rest are irrelevant.)
You will need one script to generate new embeddings according to the window size, cbow/skip-gram and the number of dimensions. Your next script will take as input a single word and a given word embedding and as output return the top 10 similar words and the metrics you compute against the ground truth from the most expressive model.
You will get a set of words and parameters before the next lecture for which you will email your results (metrics and top 10 words) to Jaspreet.

1. Script to derive embeddings for the given document corpus depending on the parameter settings – 15 points

2. Script to compute metrics based on top 10 words for a given model – 35 points