**Data Preparation**

The change I did in the data preparation:

1. Use the simple Porter stemming algorithm to stem the words.
2. Clean out the common but not useful words:

·       Thousand(s)

·       Euro(s)

·       Million(s)

·       Billion(s)

·       Year(s)

·       Company (companies)

·       Fiscal

·       “financial” as a standalone word (but not in phrases)

·       Consolidated

Also, remove 1- and 2-word phrases that are in the top 0.01% most common phrases for the full set of 1grams and 2grams to be more systematic.

In this project, the corpus got from training data contains 35707, thus the top 0.01% most common phrases are the top 3 most frequency words. These three words are shown in more than 5983 documents. Therefore, I wrote the code to remove all n-grams which are shown in more than 5983 documents.

**Build the sLDA**

The work I did in the sLDA process:

1. Get the top words for each topics
2. Get the OUT-OF-SAMPLE prediction for Y
3. Calculate Predictive R square
4. Get the topic loadings (for training dataset and test dataset)
5. Get the mean and median of topics loadings ( for training dataset and test dataset)

**Predictive R-squared**

Calculate predictive R-squared for different models (model with lemmatization, Porter stemmed model with different topic numbers). Then we selected best model with highest predictive R-squared. In the end, I find that the Porter stemmed model with 150 topics and 200 topics have highest predictive R-Square, so we decide to use these two models for the future application.

**Descriptive statistics on the topic loading**

 Generate descriptive statistics on the topic loading for each topic. Calculate the mean and median topic loading for topics 1-150 in the 150-topic model, and for topics 1-200 in the 200-topic model.

**Run a regression on topics loadings from sLDA and LDA**

Run a regression where local advantage (the outcome variable from the sLDA and LDA procedure) is the dependent variable and topic loading of topic N is the independent variable. Generate R-squared and coefficients/t-statistics from each of these models.

**Generate set of “representative paragraphs”**

Identifies the set of sentences (representative paragraph) that reflect those topics. Please read the explanation in the file “ Work Description” in folder“ Representative Paragraph”.

**Ninety thousand documents**

Take the 150- and 200-topic sLDA models trained on the training set (which is about 6,000 documents) and apply it out of sample to estimate the topic loadings on the full set of textual documents, which is about 90,000 documents. The set of 6,000 documents have the local\_adv (local advantage) variable available, but the remaining documents do not.