**Peer Review for project 1**

From QIU Zhenyu, Dept. of Mathematics

**Paper 1: NIPS Conference Papers 1987-2015 Data Set**

**by LI Haobo; CHEN Zixin; TENG fei; SHENG Rui**

* Summary of the report:
  + They first applied a Bert pre-train model to obtain the words’ embeddings, then used K-means to cluster the words. Then, they use HMM model to predict words occurrence probability distribution in 2015 based on the words’ distribution in previous papers from 1987 to 2014. They also use t-SNE to visualize the clustering results and use a sunburst chart to report the frequent words prediction result.
* Strengths of the report:
  + The visualization is very good-looking.
  + Using Bert pre-train model to obtain the words embeddings may be a good idea to figure out the sparse relationship between words and papers.
* Weaknesses of the report:
  + Need more theoretical analysis for the model and results.
* Evaluation on Clarity and quality of writing (1-5):
  + 3/5
  + Clearly written and well organized. Good figures. Good style and grammar. And report the results on a public accessible website.
  + Suggestion:
    - The poster is well enough to illustrate the idea of the project. But I think it would be better if presenting more details about the methodology part, such as the models of Bert and HMM.
* Evaluation on Technical Quality (1-5):
  + 4/5
  + Reproducible codes provided. Convincing Results. Well sited.
  + Suggestion:
    - Preprocessing the dataset may be effective. For example, filter the stop words.
    - The validation process may be naïve. I think it would be better to perform the prediction with a rolling window basis.
* Overall rating (1-5):
  + 5- The best report
* Confidence on your assessment (1-3)
  + 2- I just browse the paper without checking the details

**Paper 2: Statistical Analysis on Authors and Word Trend of NIPS Papers from 1987 to 2017**

**by DUNDA, Gerry Windiarto Mohamad**

* Summary of the report:
  + Based on the NIPS papers dataset, after performing preprocessing, they first applied TopicRank algorithm to extract the keywords for each paper. Then, they used Louvain and Leiden community detector to identify the communities of researchers’ networks and performed linear regression analysis on each community size and number of publications of each community. They also used MDS to reduce the dimension and visualize the document clusters on a two-dimensional graph.
* Strengths of the report:
  + Interesting research questions, such as detecting the communities in the authors and relationship between community size and community productivity.
  + Good preprocess the data, such as removing stop words, stemming, and lemmatizing
* Weaknesses of the report:
  + The visualization of the result is not sharp enough.
* Evaluation on Clarity and quality of writing (1-5):
  + 3/5
  + Clearly written and well organized. Many figures. Good style and grammar.
  + Suggestion: I think it would be better if presenting more details about the methodology part, such as adding an introduction about the TopicRank algorithm and community detector algorithms. Just a simple algorithm illustration is good enough.
* Evaluation on Technical Quality (1-5):
  + 4/5
  + Reproducible codes provided. Convincing Results. Well sited.
  + Suggestion:
    - May try to use K-means or other clustering algorithms to cluster the papers.
    - May try to use more dimension reduction methods, such as manifold learning, to acquire the coordinates for papers. Because the distance metrics may just have locally relationship
* Overall rating (1-5):
  + 4- A good one
* Confidence on your assessment (1-3)
  + 2- I just browse the paper without checking the details

**Paper 3: Visualization and Forecasting on Finance Data**

**by HAO, Yifan; Yueying Hu; Yakun Li; Yonglin Liu**

* Summary of the report:
  + Based on the S&P finance dataset, they performed different dimension reduction methods to reduce dimension of each stock and visualize the clustering of stocks. Also, they considered a forecasting model based on different dimension reduction methods and reported its prediction accuracy in different industries.
* Strengths of the report:
  + Good preprocessing the data: focus on the daily returns of each stock
  + Used many different dimension reduction methods to obtain results.
  + Designed a latent factor model to predict the stocks’ returns
* Weaknesses of the report:
  + The daily return is heavy tailed, thus the empirical return distribution fits the normal distribution not very well.
  + The Figure 2 would be better if normalizing the initial prices as a relative starting price, such as 1.
  + In financial market, the PC1 will be the whole market performance. To study the returns of different industries, I think it would be better to extract and exclude the whole market performance in advance.
* Evaluation on Clarity and quality of writing (1-5):
  + 3/5
  + Clearly written and well organized. Many figures. Good style and grammar.
  + Suggestion:
    - The Figure 2 would be better if normalizing the initial prices as a relative starting price, such as 1.
* Evaluation on Technical Quality (1-5):
  + 3/5
  + Reproducible codes provided. Convincing Results. Well sited.
  + Suggestion:
    - To study the returns of different industries, I think it would be better to extract and exclude the whole market performance in advance.
* Overall rating (1-5):
  + 3- An average one
* Confidence on your assessment (1-3)
  + 2- I just browse the paper without checking the details

**Paper 4: Application of Shrinkage in Portfolio Optimization**

**by ZHANG Fa; XIA Ruizhe**

* Summary of the report:
  + They tested several shrinkage approaches on portfolio optimization based on S&P finance dataset. They showed the shrinkage of precision matrix does mitigate the noisy and unstable estimation and lead to a better out-of-sample portfolio performance.
* Strengths of the report:
  + A clear introduction and formula illustration about the shrinkage estimators
  + Convincing results
* Weaknesses of the report:
  + May be better to try different look-back length (n) for the experiment, and study the shrinkage effects on different look-back length.
* Evaluation on Clarity and quality of writing (1-5):
  + 5/5
  + Clearly written and well organized. Many figures. Clear methodology. Good style and grammar.
* Evaluation on Technical Quality (1-5):
  + 5/5
  + Reproducible codes provided. Convincing Results. Well sited.
  + Suggestion:
    - May be better to try different look-back length (n) for the experiment, and study the shrinkage effects on different look-back length.
* Overall rating (1-5):
  + 5- My vote as the best-report
* Confidence on your assessment (1-3)
  + 2- I just browse the paper without checking the details

**Paper 5: Visualization and Dimensionality Reduction Techniques for US Crime Data**

**by XU Yingxue; Jiaxin ZHUANG; Fengtao ZHOU**

* Summary of the report:
  + Based on the US crime data, they performed different dimension reduction techniques to reduce the factors of cities for visualization. Also, they visualize the cities with crime-level grouping in the space of PC1 and PC2.
* Strengths of the report:
  + Tried different reduction techniques for visualization
* Weaknesses of the report:
  + Need more clarify on the reasoning of data preprocessing
* Evaluation on Clarity and quality of writing (1-5):
  + 4/5
  + Clearly written and well organized. Clear methodology. Good style and grammar.
    - I have a question: Why does the total number of crimes needs to be categorized on a four-point scale? Why would you choose 4 categories?
* Evaluation on Technical Quality (1-5):
  + 4/5
  + Reproducible codes provided. Convincing Results. Well sited.
  + Suggestion:
    - I think it would be better to perform a Horn’s parallel analysis on the dataset to empirically examinate the noise in the dataset.
* Overall rating (1-5):
  + 3- An average one
* Confidence on your assessment (1-3)
  + 2- I just browse the paper without checking the details