From: **HU Yueying** yhucn@connect.ust.hk
Subject: CSIC 5011: Project 1 Review (HU Yueying)
Date: 21 April 2023 at 6:35 PM
To: datascience.hw@gmail.com

HY

HU, Yueying 20908648

Statistical Analysis on Authors and Word Trend of NIPS Papers from 1987 to 2017.

• Summary of the report.
The authors use TopicRank to extract some important keywords and use Louvain and Leiden algorithms to identify the communities of researchers, and finally use MDS to do the clustering.

• Describe the strengths of the report.
The chosen topic is great, and the visualization of these three parts is all well done. Moreover, the analyses are very detailed and contain a lot of explanations.

• Describe the weaknesses of the report.
It might be better to contain other dimension reduction methods and conduct the comparison between them in the last part. More deep analyses are also encouraging.

• Evaluation on Clarity and quality of writing (1-5): 3
There are many typos in the report. For example, "…the clustering are…" should be "…the clustering is…", "Figure shows the result of …" should be "Figure 4 shows the result of …". Also, the explanation of Figure 4 is lacking.

• Evaluation on Technical Quality (1-5): 5

• Overall rating: 4

• Confidence on your assessment: 3

NIPS Conference Papers 1987-2015 Data Set.

• Summary of the report.
To explore the evolution of word usage, the authors use Bert pre-train model to do the word embedding, and then use K-means to do the clustering. After that, they use HMM model and t-SNE to do the prediction and visualization respectively.

• Describe the strengths of the report.
The authors use a flow diagram to explain the methods they use, which is very clear and straightforward. Also, the structure of the whole poster is also very clear.

• Describe the weaknesses of the report.
Firstly, the format of poster is not symmetric, and the word in figures cannot be seen clearly. Also, as it is a poster, representative figures and details should be put on it, instead of putting a link on the top of the figure. In addition, more deep analyses are encouraging.

• Evaluation on Clarity and quality of writing (1-5): 3

• Evaluation on Technical Quality (1-5): 5

• Overall rating: 4

• Confidence on your assessment: 3

Topic Modeling for NIPS Words.

• Summary of the report.

The authors first used LDA to extract some topics from NIPS paper datasets and then performed different dimensional reduction methods and clustering methods to visualize the results.

• Describe the strengths of the report.

The logic of this report is clear, and the visualization is very effective. And it is great to give formulas and use graphical model representation to explain LDA.

• Describe the weaknesses of the report.

Several methods (LDA, K-means, MDS, and t-SNE) are contained in the report but the authors only introduce LDA, and the content of LDA is too much. Also, the explanation of figures is not deep and enough.

• Evaluation on Clarity and quality of writing (1-5): 4

• Evaluation on Technical Quality (1-5): 4

• Overall rating: 4

• Confidence on your assessment: 3

Ancestry Prediction via Dimensionality Reduction Techniques on SNPs Data.

• Summary of the report.

The authors use PCA, MDS, and Random Projection to explore the genetic variation with geographic variation. And Random Forest, Extra Trees are also used to do the prediction of 5 top important SNPs. The visualization of these methods is well done.

• Describe the strengths of the report.

The structure of the whole poster is also very clear. When conducting PCA and MDS, the authors draw the diagram in both 2D and 3D, which is more straightforward. Also, a special case is also studied, which gives more insights for geographical variation of human gene.

• Describe the weaknesses of the report.

There are some wrong formats in the poster. Firstly, figures cannot be seen clearly. Secondly, the figures of case study don't have the chart names and number, such as "shown as bottom fig. in the middle column".

• Evaluation on Clarity and quality of writing (1-5): 4

• Evaluation on Technical Quality (1-5): 5

• Overall rating: 5

• Confidence on your assessment: 3

Visualization and Dimensionality Reduction Techniques for US Crime Data.

• Summary of the report.

The authors first conduct data preprocessing and then use PCA, SPCA, MDS, ISOMAP, LLE to analyze the relationship between region-related parameters and the total number of crimes, and finally visualize the result.

• Describe the strengths of the report.

The structure of the whole poster is also very clear. Several dimension reduction methods are used, and the visualization, which is in both 2D and 3D, is straightforward.

• Describe the weaknesses of the report.

The word in figures cannot be seen clearly. Also, in the part of visualization results, the number of figures is on the figure, which is not preferred. In addition, more deep analyses are encouraging.

• Evaluation on Clarity and quality of writing (1-5): 4

• Evaluation on Technical Quality (1-5): 4

• Overall rating: 4

• Confidence on your assessment: 3