Math5473 Mini-Project 1: NIPS Conference Papers 1987-2015 Data Set

Li Haobo, CHEN Zixin, TENG fei, SHENG Rui {hliem, zchendf, fteng, rshengac}@connect.ust.hk
LI Haobo and TENG Fei: Analysis and Poster writing, SHENG Rui: Methodology Coding, CHEN Zixin: Result visualization

1. Introduction

The impact of academic writing style on paper acceptance is a major concern for researchers. There has been a noticeable shift in language usage over the past few decades, making it essential for researchers to stay up-to-date with these changes and adapt their writing style accordingly. By doing so, researchers can ensure that their work is relevant and appealing to the contemporary academic community, increasing their chances of success in the publishing process. However, the sheer volume of research papers across various areas makes it impractical to analyze the progression of word usage. To address this issue, we propose a visual analytics approach aimed at helping researchers explore the evolution of word usage across time in accepted papers.

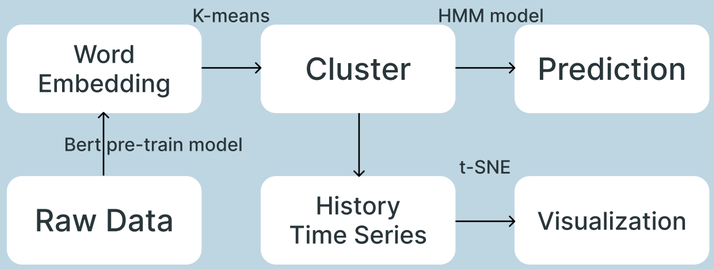
2. Dataset

Data challenge: The dataset is a co-occurrence matrix of 11463 words and 5811 NIPS conference papers from 1987 to 2015. There exists plenty of words where quite a few are rare resulting in a sparsity of dataset. The long-time interval also enlarged the search space for targeting frequent pattern variations.

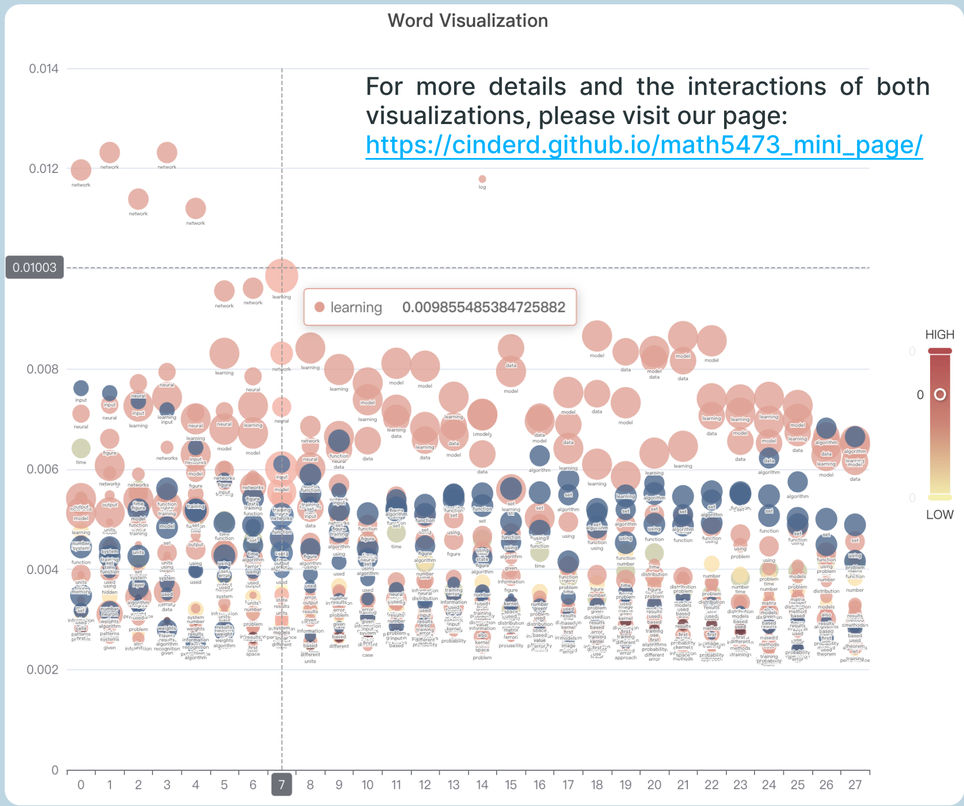
3. Task

We visualize the progression of word distribution in academic writing which can be applied as a tool to predict the possibly frequent words or phrases. We conducted an experiment to estimate the probability of word occurrences of conference papers in 2015 based on the previous words distribution of papers from 1987 to 2014. The performance is evaluated by the MAE score between predicted word distribution and ground-truth distribution.

4. Methodology



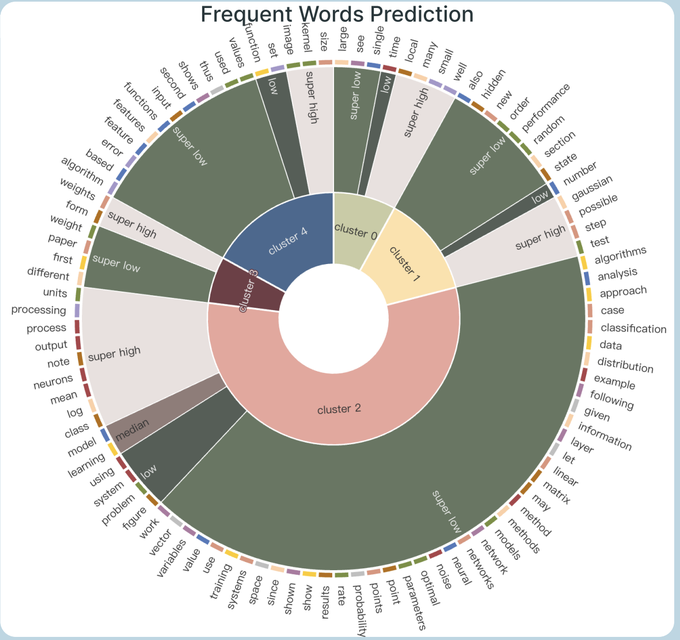
5. History Time Series Visualization



The visualization diagram illustrates the frequent words change as time goes on, where the x-axis represents the time from 1987 to 2014 and the y-axis stands for the frequency of words. Each word is represented as a circle where the size means frequency and color means the cluster that it belongs to. We can see that words from cluster 2 and cluster 4 tend to be more frequent which is actually more correlated to computer science, especially the machine learning area. In papers in the last century, the most frequent words include network, neural and input while in recent years the most frequent keywords have been changed into model, learning, etc. Since we select the NIPS paper dataset, most of the frequent words from 1987 to 2014 are relevant to the machine learning area. From the perspective of writing styles, we can deduce that researchers tend to name neural networks as models or learning-based models in recent decades. We can also summarize from the diagram that the word data has a relatively high probability across the entire time interval since data has always been an important factor in the machine learning area.

6. Word Prediction

The above chart reports the top 100 possible words. We apply a sunburst chart to organize the result, where the inner circle shows the cluster that each word belongs to and the outer circle implies the probability for each word. The 5 clusters are computed by the K-Means algorithm from each word's BERT embedding. The 5 groups of probability, super low, low, medium, high, and super high, generated by 5 equal parts of the prediction probability, imply the probability from super low to super high. From the chart, we can see that the words like output, unit, neurons, and kernel tend to have a high probability in 2015 NIPS accepted papers, which means the emerging Neural Network technology in recent decades. On the contrary, words like random algorithms tend to have low probability in distribution since traditional methods like random algorithms were used less in 2015. Our MAE score also proves that our model predicts a convincible result.



8. Reference

Kwon B C, Anand V, Severson K A, et al. DPVis: Visual analytics with hidden markov models for disease progression pathways[J]. IEEE transactions on visualization and computer graphics, 2020, 27(9): 3685-3700.