**IBM Unsupervised Learning – NIPS Conference Papers Clustering Projects**

This study looks into using clustering techniques in order to cluster words that are used in similar papers.

**1 Introduction**

This dataset can be accessed at <https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015> [1] and contains the distribution of words in the full text of the NIPS conference papers published from 1987 to 2015. The data contains 11463 unique words (words that appeared less than 50 times were not included) and spans over 5811 NIPS conference papers. The number of appearances of a word in a paper is counted.

**2 Data Exploration and Cleaning**

The dataset was found to have no missing or N/A values and was fully numerical (excluding the labelling of the words).

The distribution of word counts was found in order to determine whether it was necessary to scale the data – using the distance metric of Euclidean distance, it would only make sense to use the original scaling given all of the papers were around the same length.

|  |
| --- |
|  |
| Figure 1: Distribution of word counts |

Given the wide range of (key) word counts, it became clear that a scaling method was required whereby papers with high quantities of as word would be closer together whereas those with low quantities would keep more of their original geometric distance. As a result, the data was scaled with a log1p transformation (as oppose to a standard logarithmic transformation due to the large quantity of zero values in the data).

**4 Clustering Models**

**4.1 KMeans Clustering**

In order to determine the best number of clusters to use for the KMeans model, various models were fit with different numbers of clusters and

|  |
| --- |
|  |
| Figure : The elbow point |

**5 Key Findings**

better model than logistic regression for this dataset.

**6 Possible Flaws**

**7 Next Steps**

**References**