Joona Rantanen 713083

CS-E4650 Methods of Data mining Project work

Overview

This report describes the methods and results that were used and obtained in the text clustering project work. The project was implemented using python programming language and jupyter.cs.aalto.fi as a programming environment. In addition to python standard library pandas, numpy, nltk and sklearn were used. All of those were used for preprocessing and sklearn were used for clustering and evaluating the clusterings with NMI.

Methods

Preprocessing

The preprocessing done for this project can be divided into several parts. After the data is read from the .csv file it is transformed to a form accepted by the sklearn clustering algorithms. The id values and class labels are removed and the class labels are moved to a list to be used later in the evaluating phase. After that the corpus is fed to the sklearn TfidfVectorizer to obtain the tfidf values. It takes the stop word list as a parameter. Several most common English stop word lists were tested and two of the most best performing were selected. It also takes the stemming method as a parameter. Also here many different stemming methods were tested and the best one turned out to be the Lancaster stemmer. Also the n-gram range is given for the second clustering method. The TfidfVectorizer also takes the norm as a parameter and the l2 norm was used as instructed in the instructions. After that LSA was applied in the second clustering to the tf-idf values to get a dimensionality reduced to 100.

The idf part in tf-idf is calculated with the following formula

$$\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1,$$

Clustering methods

In the project work two different clustering methods were used. In the first experiment KMeans clustering was used as instructed in the project's instructions with k equals 5. Different parameter values were tried and the best results were obtained when using all the default values from sklearn KMeans implementation but the n_init were set to 1.

The second and more advanced clustering method used was spectral clustering. It also had all the default parameters from sklearn but the assign_labels were set to discretize.

Results

The best NMI value obtained for the KMeans clustering was 0.705. That was achieved after experimenting with many different parameter settings and different preprocessing methods. To obtain this result the standard English stop word list from nltk was used. Also no n-grams or LSA were used. The three most important keywords for KMeans clusters were:

Cluster 1: compil, program and cod
Cluster 2: robot, control and system
Cluster 3: sec, enrypt and cryptograph
Cluster 4: im, detect, vis
Cluster 5: databas, dat, rel

As one can see some of the words are stemmed wrong but the topics can still be seen quite clearly. The main topics could be programming for cluster 1, robotics for cluster 2, computer security for cluster 3, computer security for cluster 4 and databases for cluster 5.

For the spectral clustering the best NMI value obtained was 0.815. For this also many parameters combinations and preprocessing methods were tried. This result was obtained by using a different much longer stop word list. Also n-gram from range 1 to 3 was used and LSA was applied to reduce the dimensionality to 100.

In both cases the results should be reproducible. To obtain these NMI values the mean of 100 clustering results were used to remove the randomness.

Execution instructions

The project was implemented in jupyter.cs.aalto.fi using python notebooks.The code is returned both as a single python file and also as a notebook if one wants to easily get all the required packages. Also the package versions used are not necessarily the newest ones in jupyter.cs.aalto.fi so the code working with the newest packages is not guaranteed.

If you want to run the code as a python file the following packages should be installed: pandas, numpy, sklearn and nltk and if you want to run the code in jupyter.cs.aalto.fi all the required packages are pre installed.