Yılmaz BAYSAL
N19139400

CMP 614 – Assignment 1 Report

In this assignment we are expected to implement the k-means algorithm for clustering a corpus of Stack Overflow questions, with using sparse matrices. We need to apply TF-IDF weighting for our document-term matrix before clustering. As a final step, we are expected to calculate purity results for our solution.

Data preparation procedure

First, I removed punctuations and numerical tokens from the data. Then, to get rid of the noise in the data, I eliminated most frequent (stop-words, etc.) and least frequent (tokens that used 1-2 times, typos, etc.) words from the corpus. To do that, I removed the tokens which are used less than 3 times (in all documents) and tokens which are used in more than 40% of the documents.

Implementation

First, I prepared the data before clustering them as I explained above. Then, I built a vocabulary. By using the vocabulary, a document-term matrix dictionary was created. This matrix was used to build a sparse-matrix consist of calculated TF-IDF weights.

After building the sparse matrix, I implemented the K-Means algorithm. At the initialization phase, I randomly chose *k* distinct documents in corpus as centroids (*k* is equals to the number of unique labels in the corpus). Then, I ran the k-means algorithm until the convergence. I used *numpy* and *scipy* libraries while implementing my solution.

Results

To be able to understand the convergence, I used the *purity* metric. After each iteration of the algorithm, I calculated the purity value for the clusters and printed it. On each iteration, purity value increases logarithmically. So, the increase of the value will be negligible at some point. This means that the cluster changes will be very reduced at that point.

For my implementation, after 10th iteration, increase of the purity became negligible so the clusters are nearly converged. Thus, I stopped the algorithm at the 10th iteration.

```
> (venv) yilmaz@yilmac assignment_1 % python src/main.py
Purity at iteration 0 is        0.254636
Purity at iteration 1 is        0.377532
Purity at iteration 2 is        0.414286
Purity at iteration 3 is        0.437732
Purity at iteration 4 is        0.45101
Purity at iteration 5 is        0.459168
Purity at iteration 6 is        0.464296
Purity at iteration 7 is        0.467842
Purity at iteration 8 is        0.470466
Purity at iteration 9 is        0.472458
```

Figure 1: Purity scores after each iteration of K-Means algorithm.