Lab 3: Text clustering and topic modelling

Reflection Report

TDDE16 – Text Mining

Anonymous ID
15240

IDA
Linköping University

November 20, 2023

RQ 3.1: *Based on your experiments in Problems 2 and 3, what is the relation between the number of clusters and the quality of a clustering? What would a "good" number of clusters be for this particular data set?*

Generally, increasing the number of clusters, increases the amount of detail you can get in the data. However, too many clusters will always lead to overfitting. The difficulty with an algorithm such as k-means, is that we need to choose the number of clusters a priori, meaning that it is difficult to know what a "good" number of clusters will be before running the algorithm. For this reason, we can look at different evaluation methods to determine a suitable number of clusters.

In problem 3 rand index is used for this purpose. Here, we got the highest rand index for 7 clusters, suggesting that this is the most suitable number of cluster on this data set (when testing 1,2,3,5 and 7 clusters). However, we also tried to increase the clusters further, which also increased the rand index further, where the peak was reached using about 20 clusters. This went against our intuition, where we thought the rand index would decrease after the optimal number of clusters was reached. Since this was not the case, this suggests that we cannot look at only the rand index when choosing the number of clusters, but also have to consider other factors such as overfitting and in some cases also computational resources.

With this in mind, a good number of clusters for this dataset would be anything between 5-7 considering this still yielded a high rand index, while not running into the risk of overfitting.

RQ 3.2: *Why is it important to run an LDA model for multiple passes, and not just one? Why is it important to monitor an LDA model for convergence (like you did in Problem 5) and not simply run it for, say, 1000 passes?*

By monitoring an LDA model for convergence, we can limit the amount of time and computational resources it takes to train the model. It is simply not worth training the model more, since it has already converged and therefore will no longer show significant improvement.

RQ 3.3: *What are the differences between k-means and LDA? When would you use one, when the other?*

A big difference between k-means and LDA is that k-means is a so called *"hard"* clustering method, while LDA is a *"soft"* method. In hard clustering methods, each data point belongs to just one cluster. In soft clustering methods, each document instead belongs to a cluster to a *certain degree*. For this reason, the different methods are more suitable for different situations. For example, in topic modeling, LDA is more suitable since each document can belong to different topics to certain degrees. If we use k-means for this type of task, each document would only belong to one topic, which is usually not what you want. On the other hand, when doing tasks that demand a binary answer, say determining the sentiment of a movie review to be either positive or negative, a hard clustering method like k-means clustering is more suitable since it will assign each data point to just one cluster.