Document Classification

Azadeh Amiri, Dorna Amiri

The growth of information leads to the growth of digital and electronic documents. Finding specific information among this huge amount of documents is a difficult and time consuming job. The motivation behind document classification technology is to facilitate the access to the particular information in an optimal timeframe. Manually organizing documents also requires more manpower and expenses. Automatic document classification (DC, also known as text categorization, or topic spotting) is the task of automatically sorting a set of documents into categories (or classes, or topics) from a predefined set (Sebastiani, 2002). A successful automatic DC requires an effective cooperation between Information Retrieval (IR) and Machine Learning (ML) technologies.

In DC, one document is often represented as a vector of words, and all these words are not that informative to be included in the final feature set. Therefore feature selection should be applied not only to select the most relevant features, but to reduce the high dimensionality of feature vector space. In this project, text summarization will be considered as a feature selection technique to extract the least number of features with the most informativeness for each category.

The dataset which will be deployed in the project will be Reuters-21578 corpus, which is a standard benchmark for the DC. The documents refer to the Reuters newswire in 1987 and the classification was done manually by personnel from Reuters Ltd. Due to its large number of categories, different subsets of its categories have been adopted as dataset. Subset R90 which is the most frequently selected subset in different researches will be taken into account for this project. It consists of 90 categories with at least one positive training example and one test example.