Data Mining
Assignment 3

Ali Gholami

Department of Computer Engineering & Information Technology
Amirkabir University of Technology

https://aligholamee.github.io
aligholami7596@gmail.com

Abstract

Keywords.

1 Text Preprocessing

1.1 Separating Train and Test Data

In the first step, we need to properly separate the train samples and their labels from each other. To achieve this, we have implemented the function *extract labels* to does that for us. Columns are based on the current dataset which is called *spam collection dataset*.

```
        def extract_labels(data):
                return (dataset['v2'], data['v1'])
```

1.2 Extracting Features of Texts

In this section, we'll be using *Bag of Words* model from *Scikit-learn* to to turn the text content into numerical feature vectors. This model does the following steps to turn texts into feature maps.

1. Assign a fixed ID to every word occurring in any document (for instance by building a dictionary from words to integer indices).

2. Each unique word in our dictionary will correspond to a feature (descriptive feature).

Scikit-learn has a high level component which will create feature vectors for us *CountVectorizer*.

References

[1] Prashant Gupta, *Cross-Validation in Machine Learning*. Towards Data Science, Jun 5, 2017.

[2] Scikit-Learn, *sklearn.tree.DecisionTreeClassifier*. http://scikit-learn.org.

[3] Scikit-Learn, *Tuning the hyper-parameters of an estimator*. http://scikit-learn.org.