Week 8 Report

Name: Amr Elbana
Email: amr32009363@gmail.com
Country: Egypt/ living in Germany
University: Siegen.
Specialization: NLP

Problem Description:

As mentioned, on the website of the internship, document / Text classification is one of the important applications in supervised machine learning (ML). Many of news websites try to recommend similar news to the reader. The process of recommendation depends on the category of the news. News should be classified and recommended to the users based on that. The challenge is to build a good ML system to predict the category of the online news with high accuracy.

For Example – New York Times are using topic models to boost their user – article recommendation engines. Various professionals are using topic models for recruitment industries where they aim to extract latent features of job descriptions and map them to suitable candidates. They are being used to organize large datasets of emails and customer reviews and use social media profiles.

Data understanding

The data that I am using is the famous "20 Newsgroups" dataset.

The data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.

What type of data you have got for analysis? the data is in textual form.

What are the problems with the data? The data is unstructured, the data should be read into a notebook using python and some operations are done on the data to extract the target and the articles from each file. The files containing the articles have no extension but using python, it is easy to read them. Also, the content of each file is messy and contains unuseful information that should be removed using some cleansing techniques.

What approaches you are trying to apply to your data set to overcome problems? I will Clean the data by removing the stop words, the URLs, the special characters, metadata and digits.

Solutions to the problems

1. **Unstructured**
    - Collect the class of each group of articles from the main folder name
    - Collect the articles under each class form the directory
    - Create a dataframe from the previous constructed lists.
2. **Cleansing**
    - Removing the URls
    - Remove digits

Name: NLP: Document Classification
Report date: 30/06/2022
Internship Batch: LISUM10: 30
Version: 1.0
Data intake by: Amr Elbana
Data intake reviewer: All Members
Data storage location: UCI

https://github.com/amrfodd/Data-Glacier-Internship/tree/master/Data-Glacier-Week8