Data Glacier

Your Deep Learning Partner

Final Project
Virtual Internship
Classification of Documents
Application of Machine Learning to Classify Emails into Newsgroups

By Betul Mescioglu
betulari@gmail.com
Data Science Graduate Student at Lewis University
Github: betulmesci/DataGlacier_Final_Project (github.com)
Project Deadline: 30-September-22
Submission Date: 27-September-22

Data Glacier

Your Deep Learning Partner

PROBLEM DESCRIPTION

Problem Description: In this project, we have a collection of approximately 20,000 emails sent to 20 newsgroups. Our job is to classify them into correct newsgroups with machine learning techniques. These newsgroups with their corresponding target values are:

0 - alt.atheism,
1 - comp.graphics,
2 - comp.os.ms-windows.misc,
3 - comp.sys.ibm.pc.hardware,
4 - comp.sys.mac.hardware,
5 - comp.windows.x,
6 - misc.forsale,
7 - rec.autos,
8 - rec.motorcycles,
9 - rec.sport.baseball,
10 - rec.sport.hockey,
11 - sci.crypt,
12 - sci.electronics,
13 - sci.med,
14 - sci.space,
15 - soc.religion.christian,
16 - talk.politics.guns,
17 - talk.politics.mideast,
18 - talk.politics.misc,
19 - talk.religion.misc

Source: http://qwone.com/~jason/20Newsgroups/

Bussiness Understanding: As we have accumulated huge amounts of data overtime, correctly identifying documents and classifying them without human intervention have many benefits. We can parse resumes, reviews, historical documents etc..

Data Intake Report

Name: Document Classification
Report date: 9/27/22
Internship Batch: LISUM11: 30
Version:
Data intake by: Betul Mescioglu
Data intake reviewer:
Data storage location: http://qwone.com/~jason/20Newsgroups/

Tabular data details:

20news-bydate-train

Total number of observations	11314
Total number of files	20
Total number of features	1
Base format of the file	Text
Size of the data	21MB

20news-bydate-test

Total number of observations	7532
Total number of files	20
Total number of features	1
Base format of the file	Text
Size of the data	13.1MB