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

EXPLORATORY DATA ANALYSIS

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

I obtained the data from http://qwone.com/~jason/20Newsgroups/. It came in two folders: training and testing. In each folder there were 20 folders, each of which corresponded to a newsgroup, containing emails sent to that newsgroup. In total, there were 18,846 emails. I formed a new dataset combining training and testing sets, did some exploration and preprocessing on this set, applied bag of words model then finally split it into train and test datasets (Last 10 data points were held to show the performance of the models).

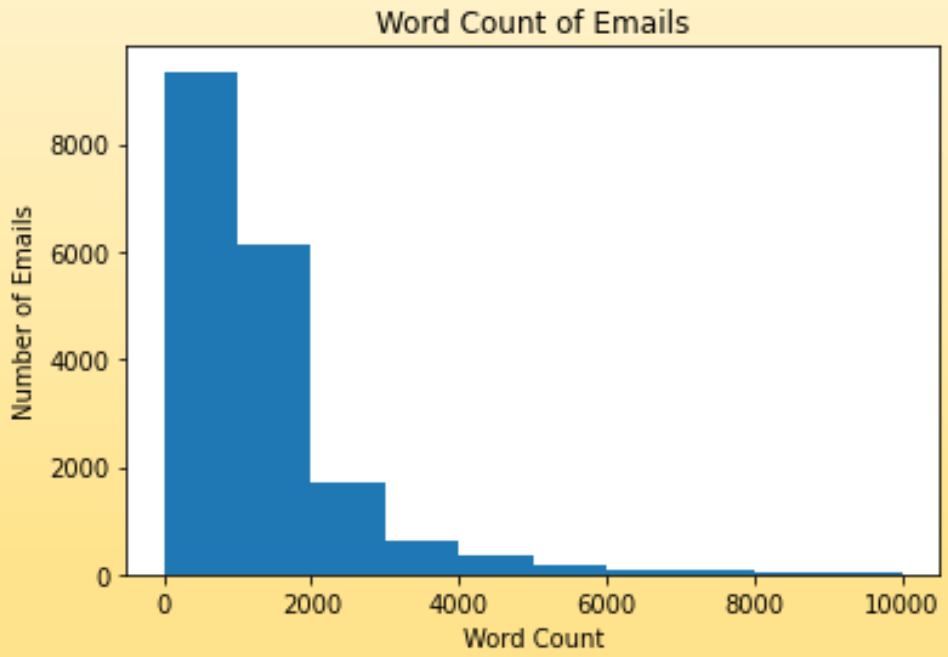
Data Glacier

Your Deep Learning Partner

EXPLORATORY DATA ANALYSIS

The dataset contains emails which is text data. It does not have any empty files. The shortest email is 107 words and longest is 67560 words long. The most frequent word count is 1000 or less. Very few of the emails contain more than 4000 words. On average, there are 1634 words in emails.

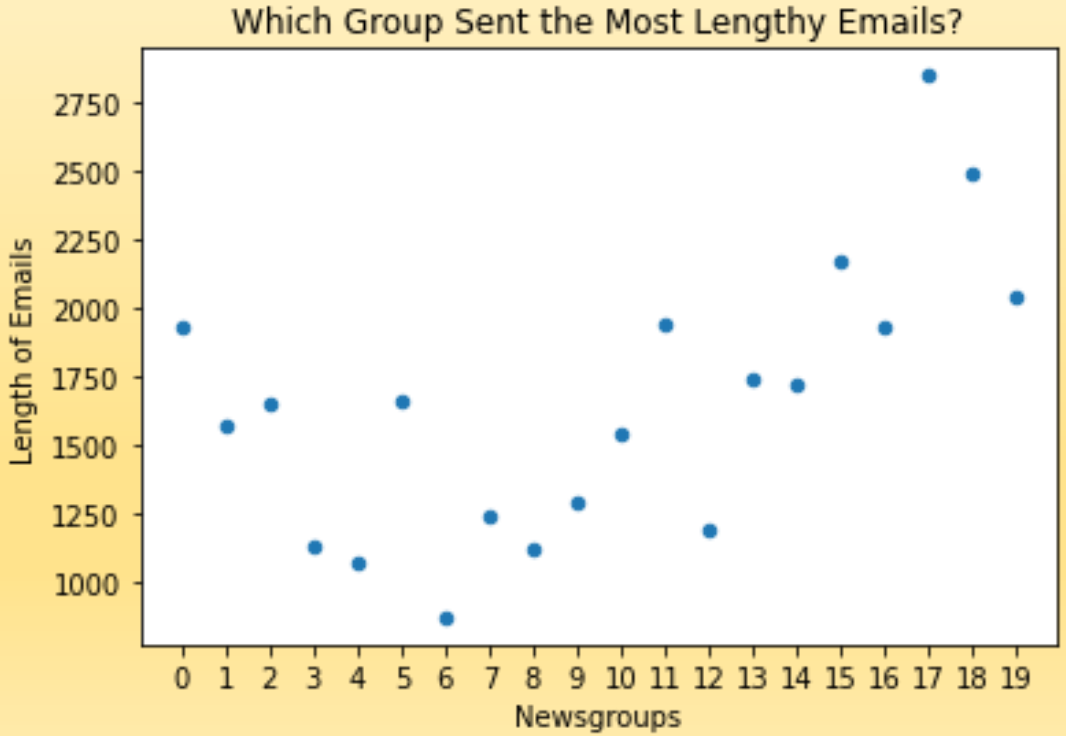
The following histogram shows the word count of the emails:



EXPLORATORY DATA ANALYSIS

I also explored whether there were any difference in email length between newsgroups. On average, Group 17, which corresponds to 'talk.politics.mideast' sent the most lengthy emails. 'talk.politics.misc' and 'talk.religion.misc' are the second and third runner-ups respectively. This is not surprising, as the politics and religion are two topics people are generally very passionate to talk about.

The following graph shows length of emails sent by newsgroups:



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