



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 \(github.com\)](https://github.com/betulmesci/DataGlacier_Final_Project)

Project Deadline: 30-September-22

Submission Date: 27-September-22



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:

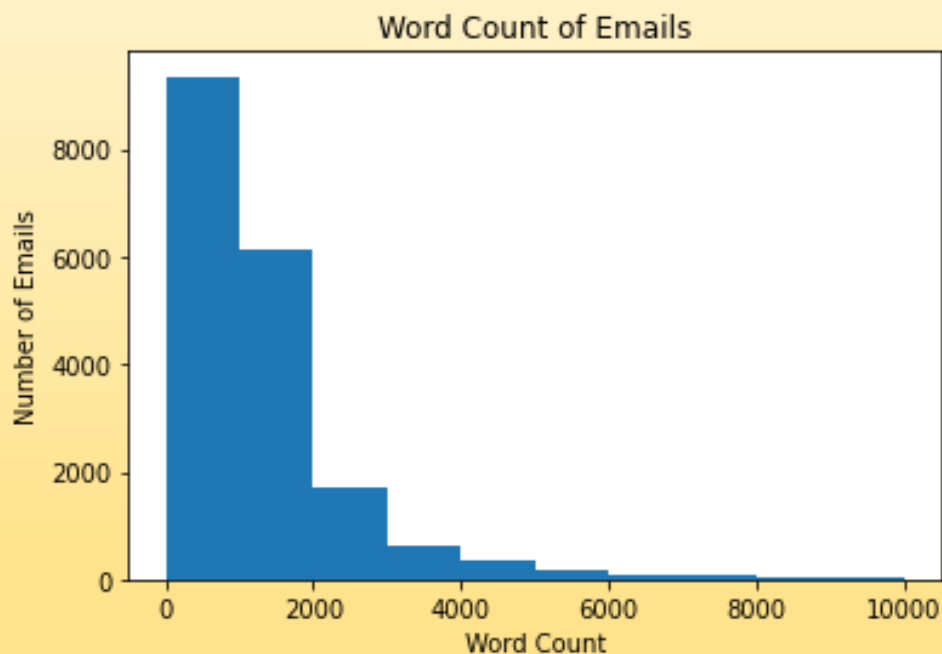
- | | |
|-------------------------------|------------------------------|
| 0 - alt.atheism, | 10 - rec.sport.hockey, |
| 1 - comp.graphics, | 11 - sci.crypt, |
| 2 - comp.os.ms-windows.misc, | 12 - sci.electronics, |
| 3 - comp.sys.ibm.pc.hardware, | 13 - sci.med, |
| 4 - comp.sys.mac.hardware, | 14 - sci.space, |
| 5 - comp.windows.x, | 15 - soc.religion.christian, |
| 6 - misc.forsale, | 16 - talk.politics.guns, |
| 7 - rec.autos, | 17 - talk.politics.mideast, |
| 8 - rec.motorcycles, | 18 - talk.politics.misc, |
| 9 - rec.sport.baseball, | 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 18846 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).



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:





I encountered two problems in data:

One problem was, there were duplicate emails in group 'misc.forsale'. So, I dropped one of them.

The other problem was, since these were emails, there were html tags throughout the text which had to be removed.

For example the following email:

"From: I3150101@dbstu1.rz.tu-bs.de (Benedikt Rosenau)\nSubject: Re: Gospel Dating\nOrganization: Technical University Braunschweig, Germany\nLines: 93\n\nIn article <65974@mimsy.umd.edu>\nmangoe@cs.umd.edu (Charley Wingate) writes:\n \n>>Well, John has a quite different, not necessarily more elaborated theology.\n>>There is some evidence that he must have known Luke, and that the content\n>>of Q was known to him, but not in a \'canonical\' form.\n>\n>"

Looked like this after cleaning:

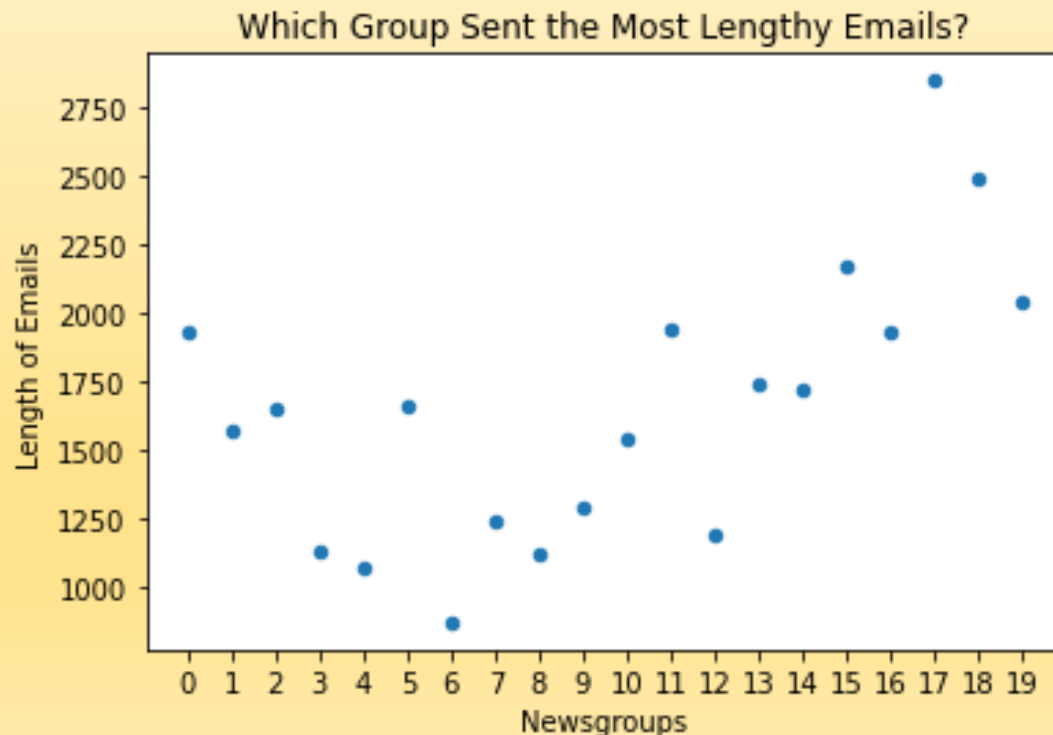
"from i3150101 dbstu1 rz tu bs de benedikt rosenau subject re gospel dating organization technical university braunschweig germany lines 93 in article mangoe cs umd edu charley wingate writes well john has a quite different not necessarily more elaborated theology there is some evidence that he must have known luke and that the content of q was known to him but not in a canonical form "

I also changed newsgroup names (target values) to numbers from 0 to 19.



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