Data science intern

August/06/2021

Group Name: Fav

Name: Feven Legesse Mammo

Email: legesse0987@outlook.com

Country: Germany

Company: Data Glacier

Specialization: Data science specialized in NLP

Problem Description/Statement

A document or text classification problem is to build an ML model that predicts/classifies text news into 20 different newsgroups.

Business/Data understanding

Dataset: 20 newsgroup datasets

Description: A collection of 20 news documents, partitioned evenly across 20 different newsgroups, and originally collected by Ken Lang.

In general, the dataset has 18,828 messages total and the original message contains only “From” and “Subject” headers.

- Each new message in the bundled file begins with four headers, they are:

Newsgroup, Document _ id, From and Subject

- Each newsgroup file in the bundle represents a single newsgroup.
- Each message in a file is the text of some newsgroup document that was posted to that newsgroup.

List of the 20 news groups according to subject matter. Some of the newsgroups are very closely related to each other (e.g. **comp.sys.ibm.pc.hardware / comp.sys.mac.hardware**), while others are highly unrelated (e.g. **misc.forsale / soc.religion.christian**).

- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x

- misc.forsale

- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey

- talk.politics.misc
- talk.politics.guns
- talk.politics.mideast

- sci.crypt
- sci.electronics
- sci. med
- sci.space

- talk.religion.misc
- alt. atheism
- soc.religion.christian

Each newsgroup has a different number of file sizes.

Workflow and deadlines

1. Problem understanding
2. EDA of dataset
3. Data preprocessing/cleaning
4. Model building and trainning
5. Performance evaluation and reporting
6. Model deployment and model inference
7. Final Project Report and PowerPoint presentation

Data intake report

Name: Document/Text classification problem
Report date: August/06/2021
Internship Batch: LISUM01
Version: 1.0
Data intake by: Feven Legesse Mammo
Data intake reviewer: Data Glacier

Tabular data details: alt.atheism

Total number of observations	-
Total number of files	799
Total number of features	-
Base format of the file	Text file

Size of the data	3.04 MB

Tabular data details: comp.graphics

Total number of observations	-
Total number of files	973
Total number of features	
Base format of the file	Text file
Size of the data	2.68 MB

Tabular data details: comp.os.ms-windows.misc

Total number of observations	⁻
Total number of files	985
Total number of features	-
Base format of the file	Text file
Size of the data	3.70 MB

Tabular data details: comp.sys.ibm.pc.hardware

Total number of observations	-
Total number of files	982
Total number of features	-
Base format of the file	Text file
Size of the data	2.45 MB

Tabular data details: comp.sys.mac.hardware

Total number of observations	-
Total number of files	961 Files
Total number of features	-
Base format of the file	Text file
Size of the data	2.29 MB

Tabular data details: comp.windows.x

Total number of observations	-
Total number of files	980 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.78 MB

Tabular data details: misc.forsale

Total number of observations	-
Total number of files	972 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.63 MB

Tabular data details: rec.autos

Total number of observations	-
Total number of files	990 Files
Total number of features	-
Base format of the file	Text file
Size of the data	2.73 MB

Tabular data details: rec.motorcycles

Total number of observations	-
Total number of files	994 Files
Total number of features	-
Base format of the file	Text file
Size of the data	2.85 MB

Tabular data details: rec.sport.baseball

Total number of observations	-
Total number of files	994 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.30 MB

Tabular data details: rec.sport.hockey

Total number of observations	-
Total number of files	999 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.67 MB

Tabular data details: sci.crypt

Total number of observations	-
Total number of files	991 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.95 MB

Tabular data details: sci.electronics

Total number of observations	-
Total number of files	981 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.17 MB

Tabular data details: sci.med

Total number of observations	-

Total number of files	990 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.79 MB

Tabular data details: sci.space

Total number of observations	-
Total number of files	987 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.72 MB

Tabular data details: soc.religion.christian

Total number of observations	-
Total number of files	997 Files
Total number of features	-
Base format of the file	Text file
Size of the data	2.19 MB

Tabular data details: talk.politics.guns

Total number of observations	-
Total number of files	910 Files
Total number of features	-
Base format of the file	Text file
Size of the data	1.83 MB

Tabular data details: talk.politics.mideast

Total number of observations	-
Total number of files	940 Files
Total number of features	-
Base format of the file	Text file
Size of the data	2.77 MB

Tabular data details: talk.politics.misc

Total number of observations	-
Total number of files	775 Files
Total number of features	-
Base format of the file	Text file
Size of the data	2.00 MB

Tabular data details: talk.religion.misc

Total number of observations	-
Total number of files	628 Files
Total number of features	-

Base format of the file	Text file
Size of the data	1.30 MB