Major Project Update (Group S)
Applications of Data Science (COMP8240)

Members:
Agam Kachhal (45762643)
Rohit Manral (45710864)
Shubham Rana (45812713)
Kripali Gandhi (45712158)

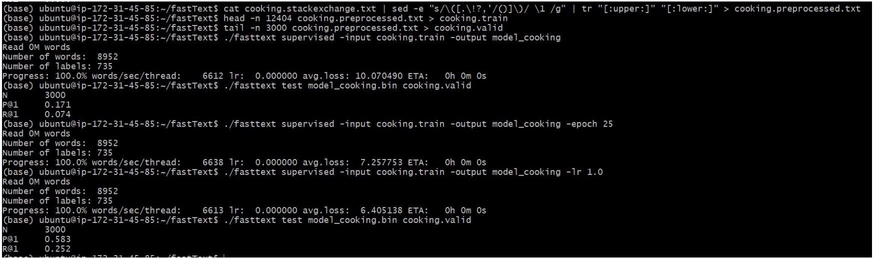
fastText

Date- 13/10/2020

Recap

FastText:

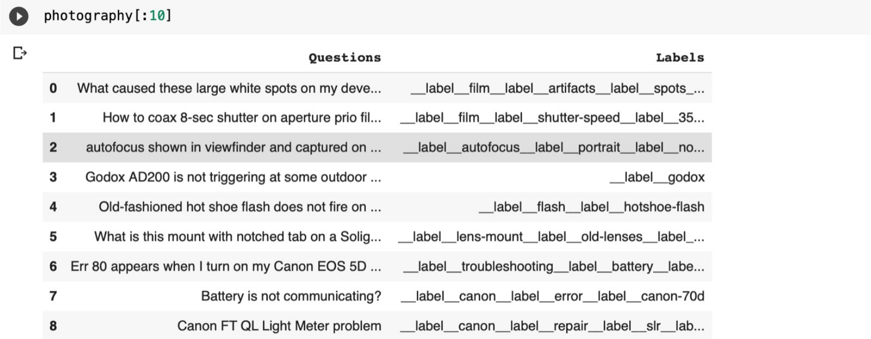
- FastText is an open-source library that allows users to learn word representations and sentence classification.
- FastText is Up and Running for Cooking dataset.
- The cooking data runs perfectly for the fastText and we tried to make code efficient by tweaking hyperparameters such as epoch, learning rate.



```
(base) ubuntu@ip-172-31-45-85:~/fastText$ cat cooking.stackexchange.txt | sed -e "s/\([.\!?,'/()]\)/ \1 /g" | tr "[:upper:]" "[:lower:]" > cooking.preprocessed.txt
(base) ubuntu@ip-172-31-45-85:~/fastText$ head -n 12404 cooking.preprocessed.txt > cooking.train
(base) ubuntu@ip-172-31-45-85:~/fastText$ tail -n 3000 cooking.preprocessed.txt > cooking.valid
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext supervised -input cooking.train -output model_cooking
Read 0M words
Number of words:  8952
Number of labels: 735
Progress: 100.0% words/sec/thread:    6612 lr:  0.000000 avg.loss: 10.070490 ETA:   0h 0m 0s
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test model_cooking.bin cooking.valid
N       3000
P@1     0.171
R@1     0.074
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext supervised -input cooking.train -output model_cooking -epoch 25
Read 0M words
Number of words:  8952
Number of labels: 735
Progress: 100.0% words/sec/thread:    6638 lr:  0.000000 avg.loss:  7.257753 ETA:   0h 0m 0s
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext supervised -input cooking.train -output model_cooking -lr 1.0
Read 0M words
Number of words:  8952
Number of labels: 735
Progress: 100.0% words/sec/thread:    6613 lr:  0.000000 avg.loss:  6.405138 ETA:   0h 0m 0s
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test model_cooking.bin cooking.valid
N       3000
P@1     0.583
R@1     0.252
(base) ubuntu@ip-172-31-45-85:~/fastText$
```

Photography Dataset

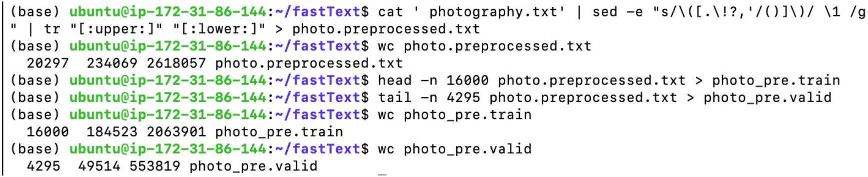
- Photography Stack Exchange is a question and site for professional, enthusiast and amateur photographers.
- Dataset Source : Photography
- We have scraped data using BeautifulSoup python library and converted the CSV file to text file so as to run fastText on it and classify the labels accordingly.(it has 20297 rows with 451 labels)



```
photography[:10]
```

	Questions	Labels
0	What caused these large white spots on my deve...	__label__film__label__artifacts__label__spots_...
1	How to coax 8-sec shutter on aperture prio fil...	__label__film__label__shutter-speed__label__35...
2	autofocus shown in viewfinder and captured on ...	__label__autofocus__label__portrait__label__no...
3	Godox AD200 is not triggering at some outdoor ...	__label__godox
4	Old-fashioned hot shoe flash does not fire on ...	__label__flash__label__hotshoe-flash
5	What is this mount with notched tab on a Solig...	__label__lens-mount__label__old-lenses__label__...
6	Err 80 appears when I turn on my Canon EOS 5D ...	__label__troubleshooting__label__battery__labe...
7	Battery is not communicating?	__label__canon__label__error__label__canon-70d
8	Canon FT QL Light Meter problem	__label__canon__label__repair__label__slr__lab...

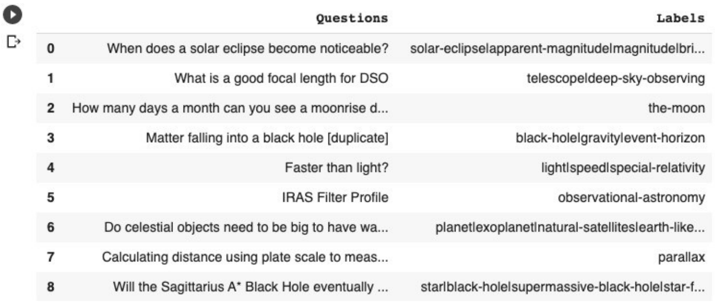
- Using Python,we separated the labels and converted them in a format which is required to implement fastText.
- Initially, we find that the case of the words is not consistent, so we used sed shell command to get all words of same case.
- We split data into train and validation data sets, train for the model training and validation for the model score.



```
(base) ubuntu@ip-172-31-86-144:~/fastText$ cat ' photography.txt' | sed -e "s/\([.\!?,'/()]\)/ \1 /g
" | tr "[:upper:]" "[:lower:]" > photo.preprocessed.txt
(base) ubuntu@ip-172-31-86-144:~/fastText$ wc photo.preprocessed.txt
  20297  234069 2618057 photo.preprocessed.txt
(base) ubuntu@ip-172-31-86-144:~/fastText$ head -n 16000 photo.preprocessed.txt > photo_pre.train
(base) ubuntu@ip-172-31-86-144:~/fastText$ tail -n 4295 photo.preprocessed.txt > photo_pre.valid
(base) ubuntu@ip-172-31-86-144:~/fastText$ wc photo_pre.train
  16000  184523 2063901 photo_pre.train
(base) ubuntu@ip-172-31-86-144:~/fastText$ wc photo_pre.valid
  4295  49514 553819 photo_pre.valid
```

Astronomy Dataset

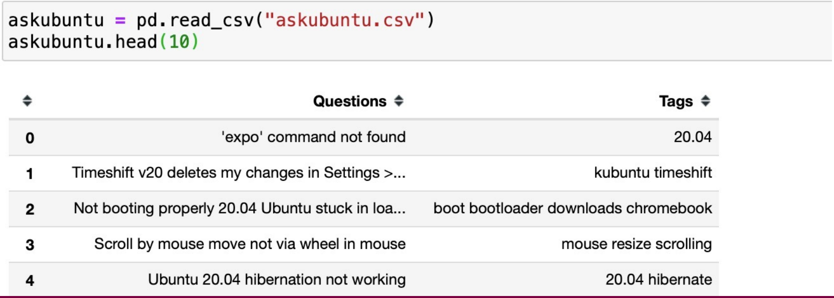
- Astronomy Stack Exchange is a question and answer site for astronomers and astrophysicists.
- Dataset Source : Astronomy
- We have scraped data using BeautifulSoup python library.
- We will explore hierarchical softmax and bigrams for this dataset. (it has 9687 rows with 518 labels)



	Questions	Labels
0	When does a solar eclipse become noticeable?	solar-eclipse|apparent-magnitude|magnitude|bri...
1	What is a good focal length for DSO	telescope|deep-sky-observing
2	How many days a month can you see a moonrise d...	the-moon
3	Matter falling into a black hole [duplicate]	black-hole|gravity|event-horizon
4	Faster than light?	light|speed|special-relativity
5	IRAS Filter Profile	observational-astronomy
6	Do celestial objects need to be big to have wa...	planet|exoplanet|natural-satellites|earth-like...
7	Calculating distance using plate scale to meas...	parallax
8	Will the Sagittarius A* Black Hole eventually ...	star|black-hole|supermassive-black-hole|star-f...

Ask Ubuntu

- Ask Ubuntu: It is a community-driven question and answer website for the Ubuntu operating system.
- Dataset Source : Ask Ubuntu
- We have scraped data using BeautifulSoup python library.
- It is a multi label dataset with 3,61,702 rows and 3379 labels. We will later explore techniques of text classification using fastText, such as Bi-grams, Hierarchical softmax etc.

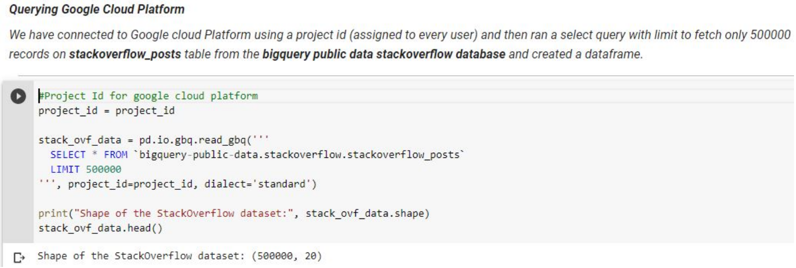


```
askubuntu = pd.read_csv("askubuntu.csv")
askubuntu.head(10)
```

	Questions	Tags
0	'expo' command not found	20.04
1	Timeshift v20 deletes my changes in Settings >...	kubuntu timeshift
2	Not booting properly 20.04 Ubuntu stuck in loa...	boot bootloader downloads chromebook
3	Scroll by mouse move not via wheel in mouse	mouse resize scrolling
4	Ubuntu 20.04 hibernation not working	20.04 hibernate

Stack Overflow Posts

Google Cloud Big Query



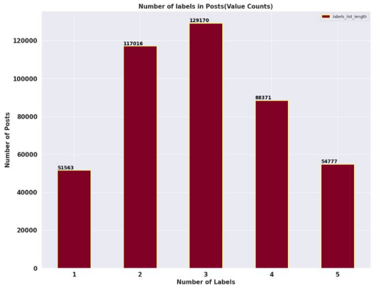
Querying Google Cloud Platform

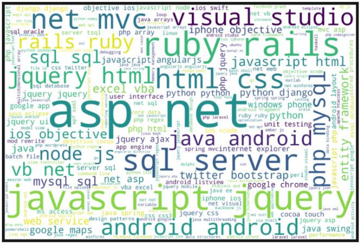
*We have connected to Google cloud Platform using a project id (assigned to every user) and then ran a select query with limit to fetch only 500000 records on **stackoverflow_posts** table from the **bigquery public data stackoverflow database** and created a dataframe.*

```
#Project Id for google cloud platform
project_id = project_id

stack_ovf_data = pd.io.gbq.read_gbq('''
  SELECT * FROM `bigquery-public-data.stackoverflow.stackoverflow_posts`
  LIMIT 500000
''', project_id=project_id, dialect='standard')

print("Shape of the StackOverflow dataset:", stack_ovf_data.shape)
stack_ovf_data.head()
```

```
Shape of the StackOverflow dataset: (500000, 20)
```

I have queried Google Cloud platform directly from the google colab using Read GBQ function from the IO GBQ library.The project-id is the Google BigQuery Account project ID.In future,we are going to explore fast text techniques to get an efficient model such as hierarchical softmax, bigrams.There are 500k rows and around 278k labels regarding programming category and input numbers will be adjusted according to VM capacity.





Comments: The horizontal bar plot illustrates the range for number of labels varies from 1 to 5 for the posts. The most number of posts(129K) have 3 labels and around 54K most posts have highest 5 labels.

Comments: Above WordCloud shows the JavaScript is the most common used label by the denizens in the StackOveflow posts followed by C , Java, PHP, JQuery.

Thank you for your attention!
Questions?