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

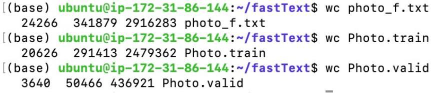
Members:
Agam Kachhal (45762643)
Rohit Manral (45710864)
Shubham Rana (45812713)
Kripali Gandhi (45712158)

fastText

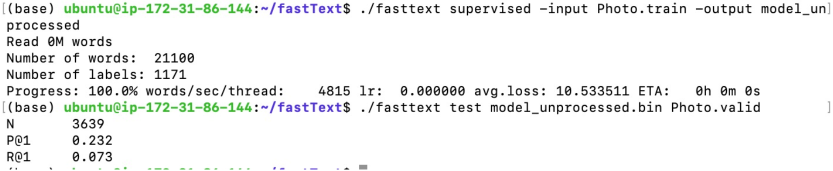
Date- 03/11/2020

Photography Dataset

- fastText modelling : 24,266 records has been used (85% training, 15% for validation)

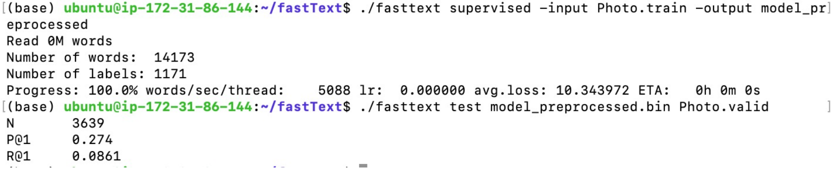


```
[(base) ubuntu@ip-172-31-86-144:~/fastText$ wc photo_f.txt
   24266  341879 2916283 photo_f.txt
[(base) ubuntu@ip-172-31-86-144:~/fastText$ wc Photo.train
   20626  291413 2479362 Photo.train
[(base) ubuntu@ip-172-31-86-144:~/fastText$ wc Photo.valid
   3640  50466 436921 Photo.valid
```



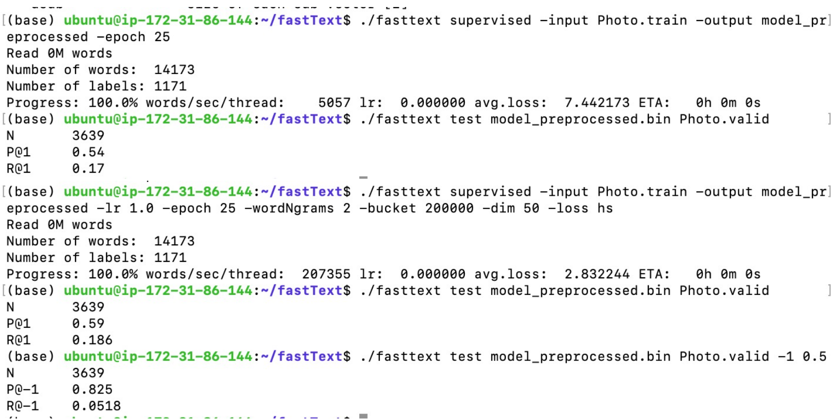
```
[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext supervised -input Photo.train -output model_un]
processed
Read 0M words
Number of words:  21100
Number of labels: 1171
Progress: 100.0% words/sec/thread:    4815 lr:  0.000000 avg.loss: 10.533511 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext test model_unprocessed.bin Photo.valid        ]
N       3639
P@1     0.232
R@1     0.073
```

- Crude Normalization(Changing case of the words, removing regular expressions)



```
[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext supervised -input Photo.train -output model_pr]
eprocessed
Read 0M words
Number of words:  14173
Number of labels: 1171
Progress: 100.0% words/sec/thread:    5088 lr:  0.000000 avg.loss: 10.343972 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext test model_preprocessed.bin Photo.valid       ]
N       3639
P@1     0.274
R@1     0.0861
```

- Adding epochs, learning rate, hierarchical softmax loss and scaling up the model.



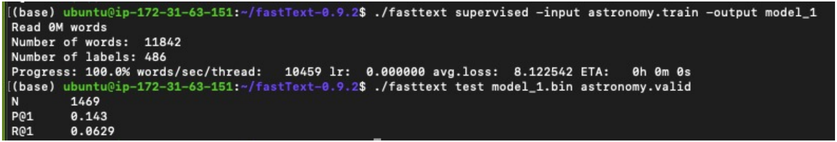
```
[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext supervised -input Photo.train -output model_pr]
eprocessed -epoch 25
Read 0M words
Number of words:  14173
Number of labels: 1171
Progress: 100.0% words/sec/thread:    5057 lr:  0.000000 avg.loss:  7.442173 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext test model_preprocessed.bin Photo.valid        ]
N       3639
P@1     0.54
R@1     0.17

[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext supervised -input Photo.train -output model_pr]
eprocessed -lr 1.0 -epoch 25 -wordNgrams 2 -bucket 200000 -dim 50 -loss hs
Read 0M words
Number of words:  14173
Number of labels: 1171
Progress: 100.0% words/sec/thread:  207355 lr:  0.000000 avg.loss:  2.832244 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext test model_preprocessed.bin Photo.valid        ]
N       3639
P@1     0.59
R@1     0.186
(base) ubuntu@ip-172-31-86-144:~/fastText$ ./fasttext test model_preprocessed.bin Photo.valid -1 0.5
N       3639
P@-1    0.825
R@-1    0.0518
```

Score Comparison of Various Models (Validation data)

Model	Precision	Recall
Initial Model	0.232	0.073
Model (Pre-processed data)	0.274	0.0861
Model (25 epochs)	0.54	0.17
Model (lr, 25 epochs,wordNgrams)	0.59	0.186
Final Model Prob 0.5	0.825	0.0518

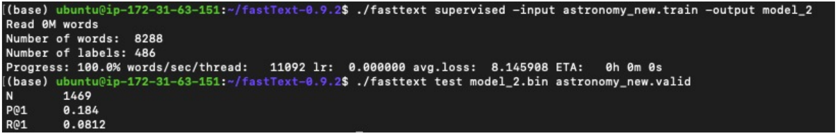
Astronomy Dataset

- fastText modelling : 9792 records has been used (85% training, 15% for validation)



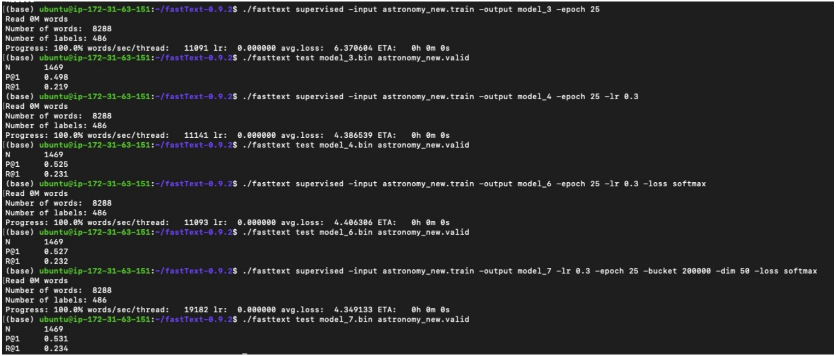
```
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext supervised -input astronomy.train -output model_1
Read 0M words
Number of words:  11842
Number of labels: 486
Progress: 100.0% words/sec/thread:   10459 lr:  0.000000 avg.loss:  8.122542 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext test model_1.bin astronomy.valid
N       1469
P@1     0.143
R@1     0.0629
```

- Crude Normalization(Changing case of the words, removing regular expressions)



```
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext supervised -input astronomy_new.train -output model_2
Read 0M words
Number of words:  8288
Number of labels: 486
Progress: 100.0% words/sec/thread:   11092 lr:  0.000000 avg.loss:  8.145908 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext test model_2.bin astronomy_new.valid
N       1469
P@1     0.184
R@1     0.0812
```

- Adding epochs, learning rate, softmax loss and scaling up the model.



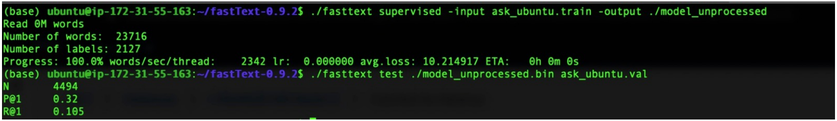
```
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext supervised -input astronomy_new.train -output model_3 -epoch 25
 Read 0M words
 Number of words:  8288
 Number of labels: 486
 Progress: 100.0% words/sec/thread:   11091 lr:  0.000000 avg.loss:  6.370604 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext test model_3.bin astronomy_new.valid
 N       1469
 P@1     0.498
 R@1     0.219
 (base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext supervised -input astronomy_new.train -output model_4 -epoch 25 -lr 0.3
[Read 0M words
 Number of words:  8288
 Number of labels: 486
 Progress: 100.0% words/sec/thread:   11141 lr:  0.000000 avg.loss:  4.386539 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext test model_4.bin astronomy_new.valid
 N       1469
 P@1     0.525
 R@1     0.231
 (base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext supervised -input astronomy_new.train -output model_6 -epoch 25 -lr 0.3 -loss softmax
[Read 0M words
 Number of words:  8288
 Number of labels: 486
 Progress: 100.0% words/sec/thread:   11093 lr:  0.000000 avg.loss:  4.406306 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext test model_6.bin astronomy_new.valid
 N       1469
 P@1     0.527
 R@1     0.232
 (base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext supervised -input astronomy_new.train -output model_7 -lr 0.3 -epoch 25 -bucket 200000 -dim 50 -loss softmax
[Read 0M words
 Number of words:  8288
 Number of labels: 486
 Progress: 100.0% words/sec/thread:   19182 lr:  0.000000 avg.loss:  4.349133 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-63-151:~/fastText-0.9.2$ ./fasttext test model_7.bin astronomy_new.valid
 N       1469
 P@1     0.531
 R@1     0.234
```

Score Comparison of Various Models (Validation data)

Model	Precision	Recall
Initial Model	0.143	0.0629
Model (Pre-processed data)	0.184	0.0812
Model (25 epochs)	0.498	0.219
Model (lr, 25 epochs)	0.525	0.231
Model (lr,25 epochs,loss)	0.527	0.232
Model (with Scaling)	0.531	0.234
Model Final Model P̃rob 0.5	0.765	0.094

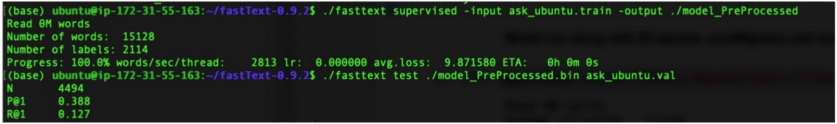
Ask Ubuntu

- fastText modelling : 21K records has been used (70% training, 15% each for validation & test dataset.)



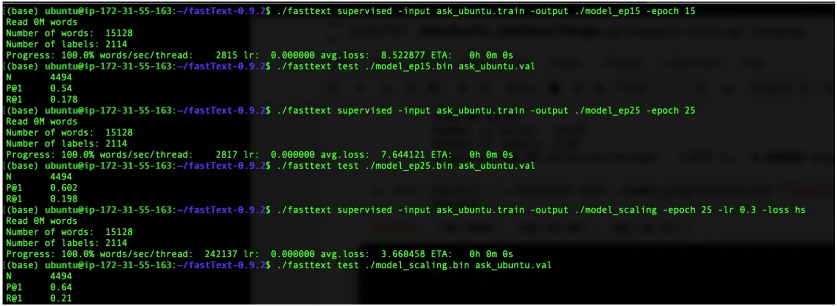
```
(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext supervised -input ask_ubuntu.train -output ./model_unprocessed
Read 0M words
Number of words:  23716
Number of labels: 2127
Progress: 100.0% words/sec/thread:    2342 lr:  0.000000 avg.loss: 10.214917 ETA:   0h 0m 0s
(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext test ./model_unprocessed.bin ask_ubuntu.val
N       4494
P@1     0.32
R@1     0.105
```

- Crude Normalization



```
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext supervised -input ask_ubuntu.train -output ./model_PreProcessed
Read 0M words
Number of words:  15128
Number of labels: 2114
Progress: 100.0% words/sec/thread:    2813 lr:  0.000000 avg.loss:  9.871580 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext test ./model_PreProcessed.bin ask_ubuntu.val
N       4494
P@1     0.388
R@1     0.127
```

- Adding epochs, learning rate, hierarchical softmax loss and scaling up the model.



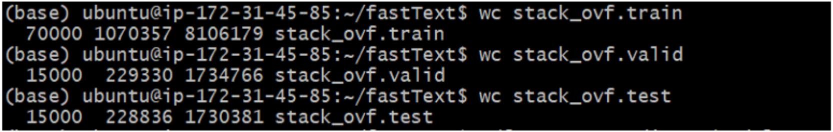
```
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext supervised -input ask_ubuntu.train -output ./model_ep15 -epoch 15
Read 0M words
Number of words:  15128
Number of labels: 2114
Progress: 100.0% words/sec/thread:    2815 lr:  0.000000 avg.loss:  8.522877 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext test ./model_ep15.bin ask_ubuntu.val
N       4494
P@1     0.54
R@1     0.178
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext supervised -input ask_ubuntu.train -output ./model_ep25 -epoch 25
Read 0M words
Number of words:  15128
Number of labels: 2114
Progress: 100.0% words/sec/thread:    2817 lr:  0.000000 avg.loss:  7.644121 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext test ./model_ep25.bin ask_ubuntu.val
N       4494
P@1     0.602
R@1     0.198
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext supervised -input ask_ubuntu.train -output ./model_scaling -epoch 25 -lr 0.3 -loss hs
Read 0M words
Number of words:  15128
Number of labels: 2114
Progress: 100.0% words/sec/thread:  242137 lr:  0.000000 avg.loss:  3.660458 ETA:   0h 0m 0s
[(base) ubuntu@ip-172-31-55-163:~/fastText-0.9.2$ ./fasttext test ./model_scaling.bin ask_ubuntu.val
N       4494
P@1     0.64
R@1     0.21
```

Score Comparison of Various Models (Validation data)

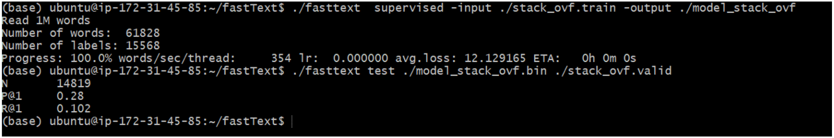
Model	Precision	Recall
Initial Model	0.32	0.105
Model (Pre-processed data)	0.388	0.127
Model (15 epochs)	0.54	0.178
Model (25 epochs)	0.602	0.198
Model (with scaling)	0.64	0.21
Model Final Model P̃rob 0.5	0.834	0.348

Stack Overflow Posts

- fastText modelling : 100K records has been used (70% training, 15% each for validation & test dataset.)

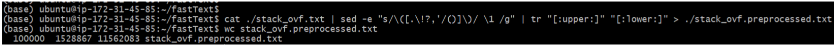


```
(base) ubuntu@ip-172-31-45-85:~/fastText$ wc stack_ovf.train
  70000 1070357 8106179 stack_ovf.train
(base) ubuntu@ip-172-31-45-85:~/fastText$ wc stack_ovf.valid
  15000  229330 1734766 stack_ovf.valid
(base) ubuntu@ip-172-31-45-85:~/fastText$ wc stack_ovf.test
  15000  228836 1730381 stack_ovf.test
```

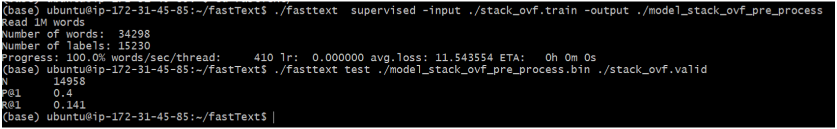


```
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext  supervised -input ./stack_ovf.train -output ./model_stack_ovf
Read 1M words
Number of words:  61828
Number of labels: 15568
Progress: 100.0% words/sec/thread:     354 lr:  0.000000 avg.loss: 12.129165 ETA:   0h 0m 0s
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test ./model_stack_ovf.bin ./stack_ovf.valid
N       14819
P@1     0.28
R@1     0.102
(base) ubuntu@ip-172-31-45-85:~/fastText$
```

- Crude Normalization(Changing case of the words, removing regular expressions)

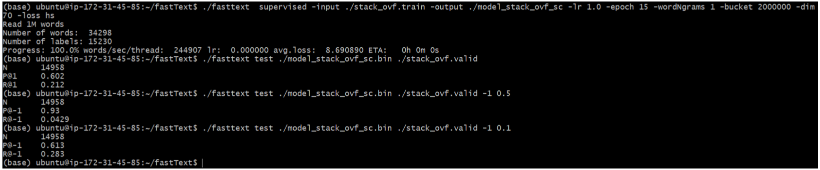


```
(base) ubuntu@ip-172-31-45-85:~/fastText$
(base) ubuntu@ip-172-31-45-85:~/fastText$ cat ./stack_ovf.txt | sed -e "s/\([.\!?,'/()]\)/ \1 /g" | tr "[:upper:]" "[:lower:]" > ./stack_ovf.preprocessed.txt
(base) ubuntu@ip-172-31-45-85:~/fastText$ wc stack_ovf.preprocessed.txt
  100000  1528867 11562083 stack_ovf.preprocessed.txt
```



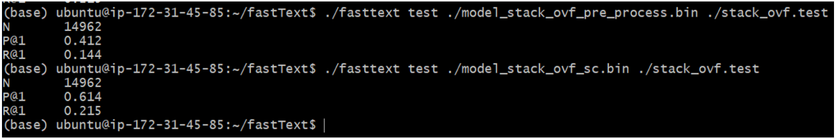
```
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext  supervised -input ./stack_ovf.train -output ./model_stack_ovf_pre_process
Read 1M words
Number of words:  34298
Number of labels: 15230
Progress: 100.0% words/sec/thread:     410 lr:  0.000000 avg.loss: 11.543554 ETA:   0h 0m 0s
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test ./model_stack_ovf_pre_process.bin ./stack_ovf.valid
N       14958
P@1     0.4
R@1     0.141
(base) ubuntu@ip-172-31-45-85:~/fastText$
```

- Adding epochs, learning rate, wordNgrams, hierarchical softmax loss and scaling up the model.



```
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext  supervised -input ./stack_ovf.train -output ./model_stack_ovf_sc -lr 1.0 -epoch 15 -wordNgrams 1 -bucket 2000000 -dim
70 -loss hs
Read 1M words
Number of words:  34298
Number of labels: 15230
Progress: 100.0% words/sec/thread:  244907 lr:  0.000000 avg.loss:  8.690890 ETA:   0h 0m 0s
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test ./model_stack_ovf_sc.bin ./stack_ovf.valid
N       14958
P@1     0.602
R@1     0.212
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test ./model_stack_ovf_sc.bin ./stack_ovf.valid -1 0.5
N       14958
P@-1    0.93
R@-1    0.0429
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test ./model_stack_ovf_sc.bin ./stack_ovf.valid -1 0.1
N       14958
P@-1    0.613
R@-1    0.283
(base) ubuntu@ip-172-31-45-85:~/fastText$ |
```

- Score of final model on Test dataset.

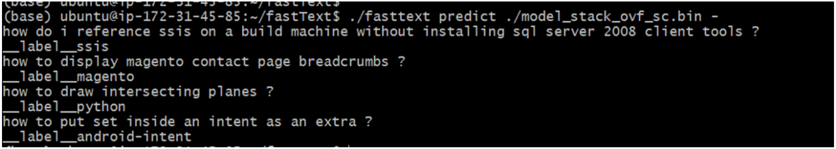


```
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test ./model_stack_ovf_pre_process.bin ./stack_ovf.test
N       14962
P@1     0.412
R@1     0.144
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext test ./model_stack_ovf_sc.bin ./stack_ovf.test
N       14962
P@1     0.614
R@1     0.215
(base) ubuntu@ip-172-31-45-85:~/fastText$ |
```

Score Comparison of Various Models (Validation data)

Model	Precision	Recall
Initial Model	0.28	0.102
Model (Pre-processed data)	0.40	0.141
Model (lr,15 epochs)	0.512	0.179
Model (lr, 25 epochs)	0.539	0.189
Model (lr,25 epochs,WordNgrams,hs)	0.562	0.197
Model (with Scaling)	0.61	0.214
Model Final Model P̃rob 0.5	0.93	0.0429

- Making Predictions.



```
(base) ubuntu@ip-172-31-45-85:~/fastText$
(base) ubuntu@ip-172-31-45-85:~/fastText$ ./fasttext predict ./model_stack_ovf_sc.bin -
how do i reference ssis on a build machine without installing sql server 2008 client tools ?
__label__ssis
how to display magento contact page breadcrumbs ?
__label__magento
how to draw intersecting planes ?
__label__python
how to put set inside an intent as an extra ?
__label__android-intent
```

Link to Repository : Group S

Thank you for your attention!

Questions?