Project 2:

Hacker News DataSet.

Submitted by:

Name: **Dhruvi Gadhiya**

Student ID: **40084176**

Name: **Krisha Patel**

Student ID: **40084336**

Table Context:

Introduction:

Hacker News is a popular technology site, where user-submitted stories (known as "posts") are voted and commented upon. The site is extremely popular in technology and start-up circles. The top posts can attract hundreds of thousands of visitors.

With the help of this data we can get predict which title correspondes to which post (story, ask_hn, show_hn or poll). We have used the given training data set and modified according to the constraints.

We have used

MWETokenizer to tokenized text like “ask”, “hn” words and “Show” ”hn” words into to a single word for further use.

Lemmatize function is used to lemmatize or group all the common meanings with inflected words into single.

Vectorizer is used to convert a text document into matrix form of token counts

Naïve bayes theorem for language modelling.

Show and analyze the results of the baseline experiment

model-2018-new - Notepad

File Edit Format View Help

```
5470  (cad?)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5471  (caddy,  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5472  (caf)  2 0.00000119  0  0.00000399  0  0.00000512  0  0.00151976
5473  (calculates  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5474  (called  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5475  (callout)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5476  (cameras)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5477  (can  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5478  (can't)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5479  (canada  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5480  (canada)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5481  (canada,  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5482  (cannabidiol)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5483  (canva)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5484  (canvas  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5485  (canvas,  0 0.00000024  1  0.00001196  0  0.00000512  0  0.00151976
5486  (capture  0 0.00000024  0  0.00000399  1  0.00001536  0  0.00151976
5487  (car,  0 0.00000024  1  0.00001196  0  0.00000512  0  0.00151976
5488  (carbanak)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5489  (cardash  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5490  (careful)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5491  (carnegie,  0 0.00000024  1  0.00001196  0  0.00000512  0  0.00151976
5492  (carrd.co  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5493  (cartoon)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5494  (cas)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5495  (case  9 0.00000453  0  0.00000399  0  0.00000512  0  0.00151976
5496  (case-shiller)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5497  (cash)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5498  (cautions)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5499  (cbc  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5500  (ccc  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5501  (ccpa)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5502  (cdk)  2 0.00000119  0  0.00000399  0  0.00000512  0  0.00151976
5503  (cdn  0 0.00000024  0  0.00000399  1  0.00001536  0  0.00151976
```

In this part of project we first started with tokenizer and then to lemmatizer and then vectorize the data in the model.

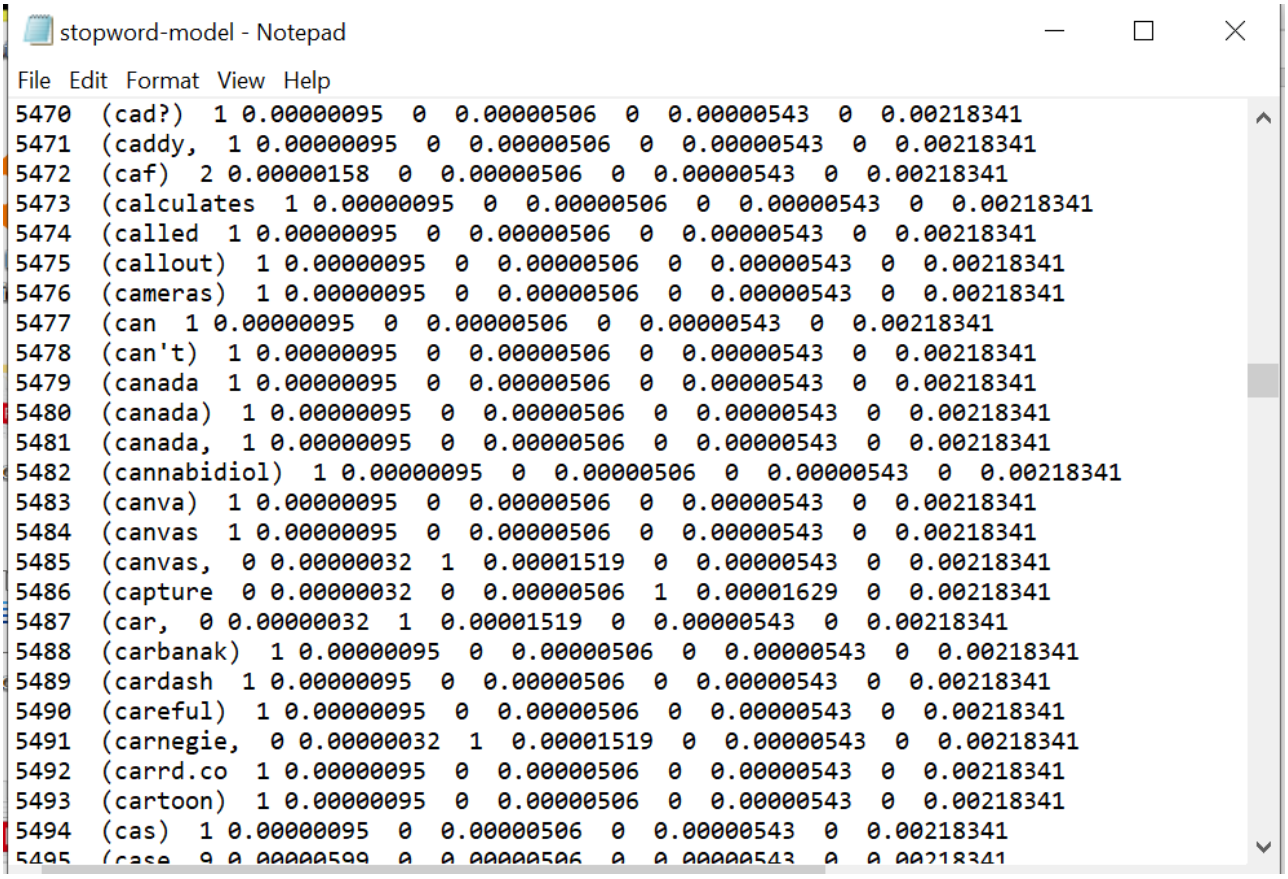
Transforming the data to given output format was a tough part. Then we decided to use a user defined function which will take care of all the basic function in all the task in project.

Which also help us to lean new libraries to work with as numpy, sklearn, pandas etc.

Results and analysis of the stop-word filtering experiment

Because of stop-words filtering we are getting more efficiency than that of experiment 1. After removing all the stop-words we are able to process the data even faster and in more efficient way.

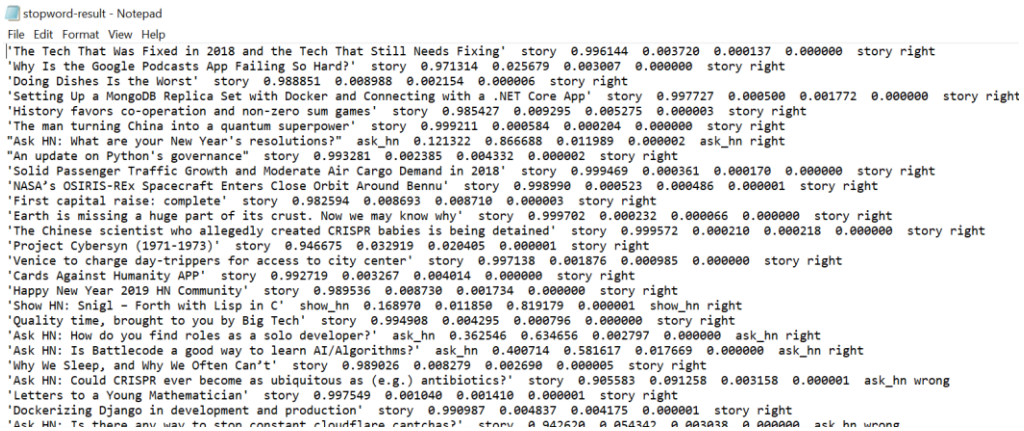
The accuracy is 93%



stopword-model - Notepad

File Edit Format View Help

```
5470  (cad?)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5471  (caddy,  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5472  (caf)  2 0.00000158  0  0.00000506  0  0.00000543  0  0.00218341
5473  (calculates  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5474  (called  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5475  (callout)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5476  (cameras)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5477  (can  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5478  (can't)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5479  (canada  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5480  (canada)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5481  (canada,  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5482  (cannabidiol)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5483  (canva)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5484  (canvas  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5485  (canvas,  0 0.00000032  1  0.00001519  0  0.00000543  0  0.00218341
5486  (capture  0 0.00000032  0  0.00000506  1  0.00001629  0  0.00218341
5487  (car,  0 0.00000032  1  0.00001519  0  0.00000543  0  0.00218341
5488  (carbanak)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5489  (cardash  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5490  (careful)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5491  (carnegie,  0 0.00000032  1  0.00001519  0  0.00000543  0  0.00218341
5492  (carrd.co  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5493  (cartoon)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5494  (cas)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5495  (case  9 0.00000599  0  0.00000506  0  0.00000543  0  0.00218341
```



stopword-result - Notepad

File Edit Format View Help

```
'The Tech That Was Fixed in 2018 and the Tech That Still Needs Fixing'  story  0.996144  0.003720  0.000137  0.000000  story right
'Why Is the Google Podcasts App Failing So Hard?'  story  0.971314  0.025679  0.003007  0.000000  story right
'Doing Dishes Is the Worst'  story  0.988851  0.008988  0.002154  0.000006  story right
'Setting Up a MongoDB Replica Set with Docker and Connecting with a .NET Core App'  story  0.997727  0.000500  0.001772  0.000000  story right
'History favors co-operation and non-zero sum games'  story  0.985427  0.009295  0.005275  0.000003  story right
'The man turning China into a quantum superpower'  story  0.999211  0.000584  0.000204  0.000000  story right
"Ask HN: What are your New Year's resolutions?"  ask_hn  0.121322  0.866688  0.011989  0.000002  ask_hn right
"An update on Python's governance"  story  0.993281  0.002385  0.004332  0.000002  story right
'Solid Passenger Traffic Growth and Moderate Air Cargo Demand in 2018'  story  0.999469  0.000361  0.000170  0.000000  story right
'NASA's OSIRIS-REx Spacecraft Enters Close Orbit Around Bennu'  story  0.998990  0.000523  0.000486  0.000001  story right
'First capital raise: complete'  story  0.982594  0.008693  0.008710  0.000003  story right
'Earth is missing a huge part of its crust. Now we may know why'  story  0.999702  0.000232  0.000066  0.000000  story right
'The Chinese scientist who allegedly created CRISPR babies is being detained'  story  0.999572  0.000210  0.000218  0.000000  story right
'Project Cybersyn (1971-1973)'  story  0.946675  0.032919  0.020405  0.000001  story right
'Venice to charge day-trippers for access to city center'  story  0.997138  0.001876  0.000985  0.000000  story right
'Cards Against Humanity APP'  story  0.992719  0.003267  0.004014  0.000000  story right
'Happy New Year 2019 HN Community'  story  0.989536  0.008730  0.001734  0.000000  story right
'Show HN: Snigl – Forth with Lisp in C'  show_hn  0.168970  0.011850  0.819179  0.000001  show_hn right
'Quality time, brought to you by Big Tech'  story  0.994908  0.004295  0.000796  0.000000  story right
'Ask HN: How do you find roles as a solo developer?'  ask_hn  0.362546  0.634656  0.002797  0.000000  ask_hn right
'Ask HN: Is Battlecode a good way to learn AI/Algorithms?'  ask_hn  0.400714  0.581617  0.017669  0.000000  ask_hn right
'Why We Sleep, and Why We Often Can't'  story  0.989026  0.008279  0.002690  0.000005  story right
'Ask HN: Could CRISPR ever become as ubiquitous as (e.g.) antibiotics?'  story  0.905583  0.091258  0.003158  0.000001  ask_hn wrong
'Letters to a Young Mathematician'  story  0.997549  0.001040  0.001410  0.000001  story right
'Dockerizing Django in development and production'  story  0.990987  0.004837  0.004175  0.000001  story right
'Ask HN: Is there any way to stop constant cloudflare captchas?'  story  0.942620  0.054342  0.003038  0.000000  ask_hn wrong
```

Results and analysis of the word-length filtering experiment

Word-Length filtering removes all the words with length >9 and <2. After apply this logic we experience a lot increase in the efficiency as well as probability of finding the post type results.

The accuracy is 94.94%

wordlength-model - Notepad

File Edit Format View Help

```
714  (cad?)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
715  (caddy  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
716  (caf)  2 0.00000156  0  0.00000520  0  0.00000643  0  0.00188679
717  (called  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
718  (callout)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
719  (cameras)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
720  (can  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
721  (can't)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
722  (canada  2 0.00000156  0  0.00000520  0  0.00000643  0  0.00188679
723  (canada)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
724  (canva)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
725  (canvas  1 0.00000093  1  0.00001561  0  0.00000643  0  0.00188679
726  (capture  0 0.00000031  0  0.00000520  1  0.00001930  0  0.00188679
727  (car  0 0.00000031  1  0.00001561  0  0.00000643  0  0.00188679
728  (cardash  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
729  (careful)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
730  (carnegie  0 0.00000031  1  0.00001561  0  0.00000643  0  0.00188679
731  (carrd.co  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
732  (cartoon)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
733  (cas)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
734  (case  9 0.00000591  0  0.00000520  0  0.00000643  0  0.00188679
735  (cash)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
736  (cbc  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
```

*wordlength-result - Notepad

File Edit Format View Help

```
'Top JavaScript Frameworks and Topics to Learn in 2019'  story  0.970848  0.024349  0.004803  0.000000  story right
'In Search of Lost Screen Time'  story  0.989616  0.007081  0.003302  0.000000  story right
'Altitude-compensating dual-bell rocket engine nozzle design'  story  0.950132  0.010998  0.038865  0.000006  story right
'Firm Led by Google Veterans Uses A.I. To 'Nudge' Workers Toward Happiness'  story  0.999323  0.000304  0.000373  0.000000  story right
'U.S. Strategic Command apologizes for tweet about dropping bombs'  story  0.997886  0.001001  0.001113  0.000000  story right
'Netflix poaches CFO from Activision Blizzard'  story  0.982982  0.007369  0.009633  0.000015  story right
'Google is aware of you making purchases'  story  0.967309  0.030798  0.001893  0.000000  story right
'After damaging Reuters report, J&J doubles down on talc safety message'  story  0.984914  0.009898  0.005185  0.000003  story right
'Happy New Year'  story  0.986031  0.011352  0.002617  0.000000  story right
'In civil suit, USC reinstates Armann Premjee after defeat in Court of Appeal'  story  0.996908  0.002570  0.000522  0.000000  story right
'MIT researchers are now 3D-printing glass'  story  0.993688  0.005887  0.000425  0.000000  story right
"Detroit's Big Comeback: Out of Bankruptcy, a Rebirth"  story  0.986708  0.009447  0.003841  0.000003  story right
'I love you all. Happy Fu**ing new year everyone'  story  0.959332  0.029467  0.011200  0.000001  story right
'Help break Quilt- Find as many unexpected results as possible'  story  0.952211  0.042732  0.005057  0.000000  story right
'A Simplified Political History of Big Data – This Political Woman – Medium'  story  0.998637  0.000978  0.000385  0.000000  story right
'Why was mouse designed wrong?'  story  0.993100  0.006680  0.000220  0.000000  story right
'10 Personal Finance Lessons for Technology Professionals'  story  0.984485  0.011527  0.003988  0.000000  story right
'Docker Base Image OS Size Comparison'  story  0.966134  0.007958  0.025907  0.000000  story right
"The end of digital revolution's childhood"  story  0.992816  0.002923  0.004262  0.000000  story right
'Why the world is full of buttons that don't work?'  story  0.996177  0.003369  0.000455  0.000000  story right
```

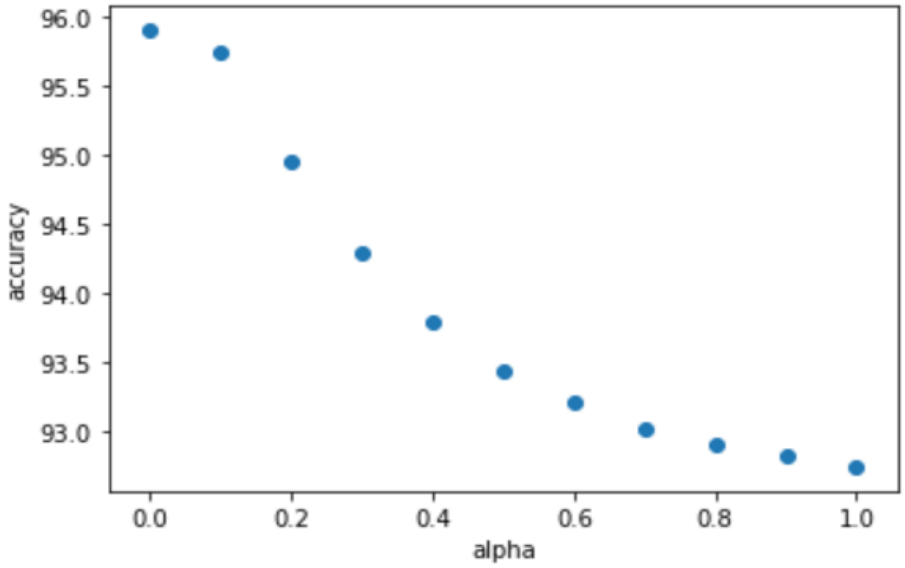
Results and analysis of the smoothing experiment

After applying the smoothing to the formula we get the following the results. We saw that lower the value of alpha “0” the greater is the accuracy. And higher the value of alpha the lower the accuracy.

Ploting the alpha v/s Accuracy

```
alpha=list(predictions.keys())
prediction=[k*100 for k in list(predictions.values())]

plt.scatter(alpha, prediction)
plt.xlabel('alpha')
plt.ylabel('accuracy')
plt.show()
```



Compare and discuss the results of the 4 experiments.

From all the experiments we got the following results:

stopword-model - Notepad

File Edit Format View Help

```
5821  (desktop)  1 0.00000095  0  0.00000506  1  0.00001629  0  0.00218341
5822  (despite  1 0.00000095  1  0.00001519  0  0.00000543  0  0.00218341
5823  (detectors  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5824  (detroit)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5825  (dev  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5826  (dev)  1 0.00000095  1  0.00001519  0  0.00000543  0  0.00218341
5827  (developed  0 0.00000032  0  0.00000506  1  0.00001629  0  0.00218341
5828  (developer  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5829  (developer)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5830  (devops  0 0.00000032  1  0.00001519  0  0.00000543  0  0.00218341
5831  (devops)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5832  (devops)?  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5833  (devs,  0 0.00000032  1  0.00001519  0  0.00000543  0  0.00218341
5834  (devumi)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5835  (dfdl)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5836  (dfs,  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
5837  (dgie)  1 0.00000095  0  0.00000506  0  0.00000543  0  0.00218341
```

wordlength-model - Notepad

File Edit Format View Help

```
923  (desktop)  1 0.00000093  0  0.00000520  1  0.00001930  0  0.00188679
924  (despite  1 0.00000093  1  0.00001561  0  0.00000643  0  0.00188679
925  (detroit)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
926  (dev  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
927  (dev)  1 0.00000093  1  0.00001561  0  0.00000643  0  0.00188679
928  (devops  0 0.00000031  1  0.00001561  0  0.00000643  0  0.00188679
929  (devops)  2 0.00000156  0  0.00000520  0  0.00000643  0  0.00188679
930  (devs  0 0.00000031  1  0.00001561  0  0.00000643  0  0.00188679
931  (devumi)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
932  (dfdl)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
933  (dfs  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
934  (dgie)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
935  (dhcpv6)  1 0.00000093  0  0.00000520  0  0.00000643  0  0.00188679
936  (dhh)  1 0 00000093  0  0 00000520  0  0 00000643  0  0 00188679
```

model-2018-new - Notepad

File Edit Format View Help

```
5821  (desktop)  1 0.00000072  0  0.00000399  1  0.00001536  0  0.00151976
5822  (despite  1 0.00000072  1  0.00001196  0  0.00000512  0  0.00151976
5823  (detectors  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5824  (detroit)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5825  (dev  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5826  (dev)  1 0.00000072  1  0.00001196  0  0.00000512  0  0.00151976
5827  (developed  0 0.00000024  0  0.00000399  1  0.00001536  0  0.00151976
5828  (developer  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5829  (developer)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5830  (devops  0 0.00000024  1  0.00001196  0  0.00000512  0  0.00151976
5831  (devops)  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5832  (devops)?  1 0.00000072  0  0.00000399  0  0.00000512  0  0.00151976
5833  (devs,  0 0.00000024  1  0.00001196  0  0.00000512  0  0.00151976
```

Experiment 1: We got around 93% of accuracy
Experiment 2: We got around 94% of accuracy
Experiment 3: We got around 94% of accuracy
Experiment 5: We got around 96% of accuracy

With all the results we can see that more we filter the data and remove unnecessary data and clean the data will help us to get more accuracy and can predict the result with better probability.

If you were to continue working on this project, what do you feel would be interesting to investigate? Are there questions that you would like to investigate more, if you had the time and the energy?

If we are given more time and energy we would love to take this project forward. So far we have learnt a lot from the project given to us. This project helped us to get into real life problems and how they actually work. How to process the data and then filter them on how we want them to work.

With python it was a bit challenging but we learnt a great deal and it was real fun doing some real life project.

References:

https://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

https://www.youtube.com/watch?v=LRFdF9J__Tc&list=PLQiyVNMpDLKnZYBTUOlSI9mi9wAErFtFm&index=28

https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

https://kite.com/python/docs/nltk.MWETokenizer

https://www.geeksforgeeks.org/python-lemmatization-with-nltk/