Summarising News Article

Mentor:
Prannay Khosla

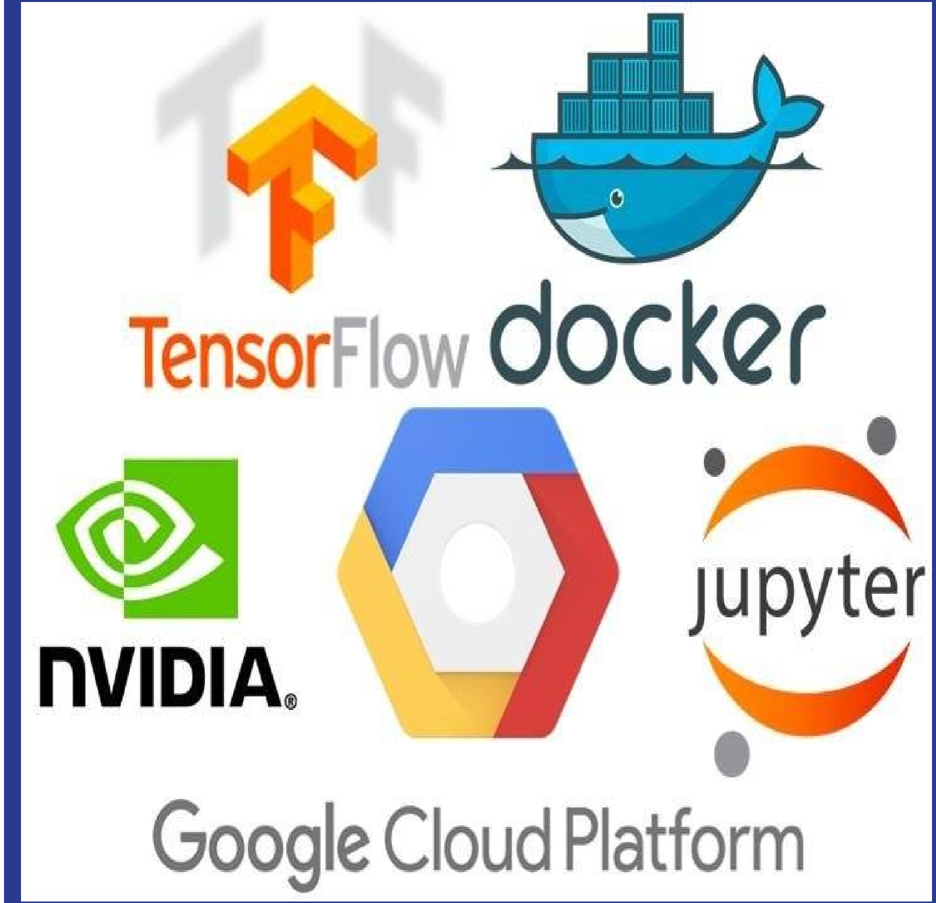
Team Members:

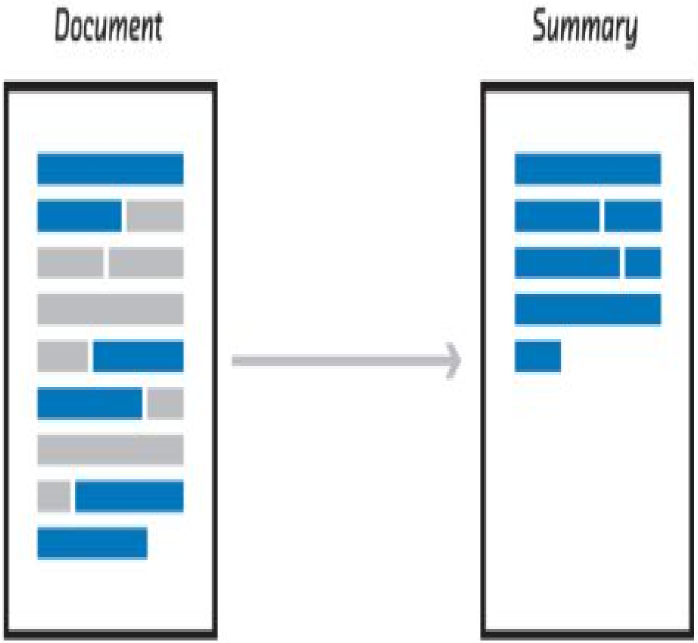
-Ritesh Kumar

-Hunarpreet Singh

-Vipul Bajaj

Every day, people rely on a wide variety of sources to stay informed, from news stories to social media posts to search results. Being able to develop Machine Learning models that can automatically deliver accurate summaries of longer text can be useful for digesting such large amounts of information in a compressed form, and is a long-term goal of our team.



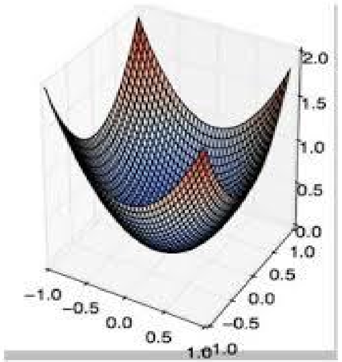


Summarization can also serve as an interesting reading comprehension test for machines. To summarize well, machine learning models need to be able to comprehend documents and distill the important information, tasks which are highly challenging for computers, especially as the length of a document increases.

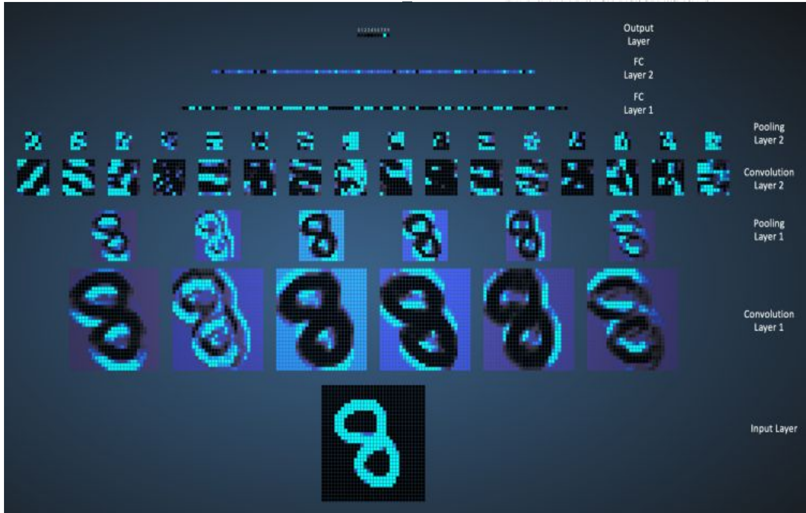
Project Details:

AIM:

- In this project we wrote working code of News Summariser.
- If supplied with news article of txt format the program will print the summary of the article.
- The accuracy of output summary depends on how well the article is written and that of cluster size chosen.



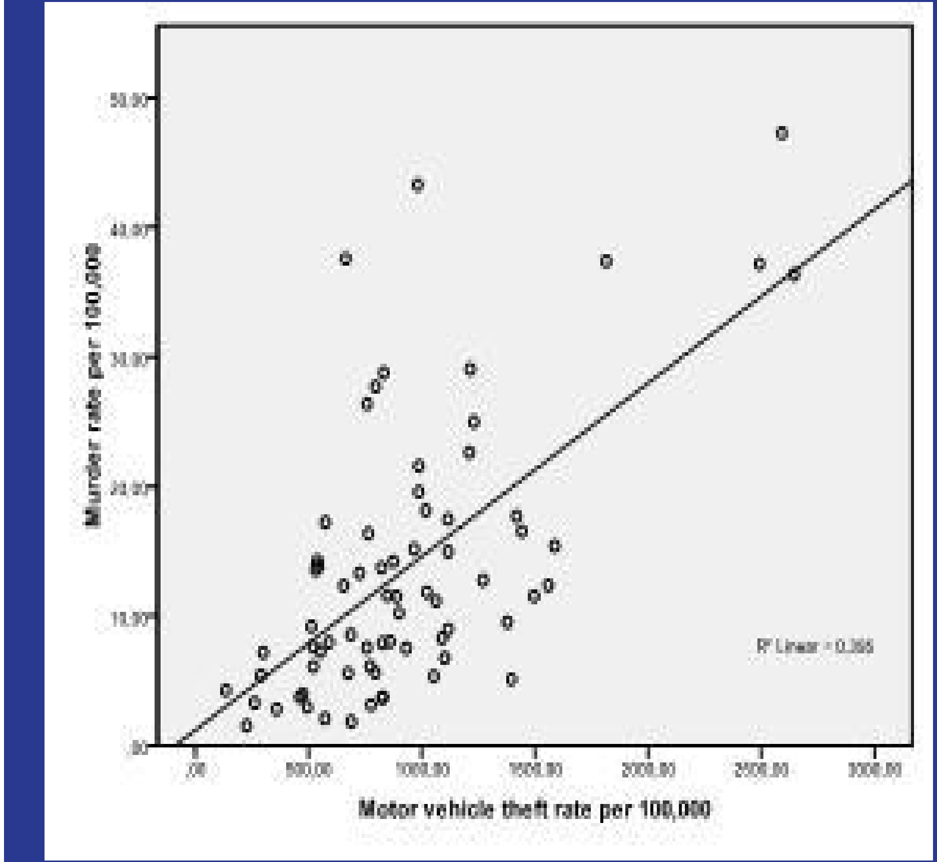
© Can Stock Photo - csp11823483



WHAT HAVE WE DONE SO FAR...

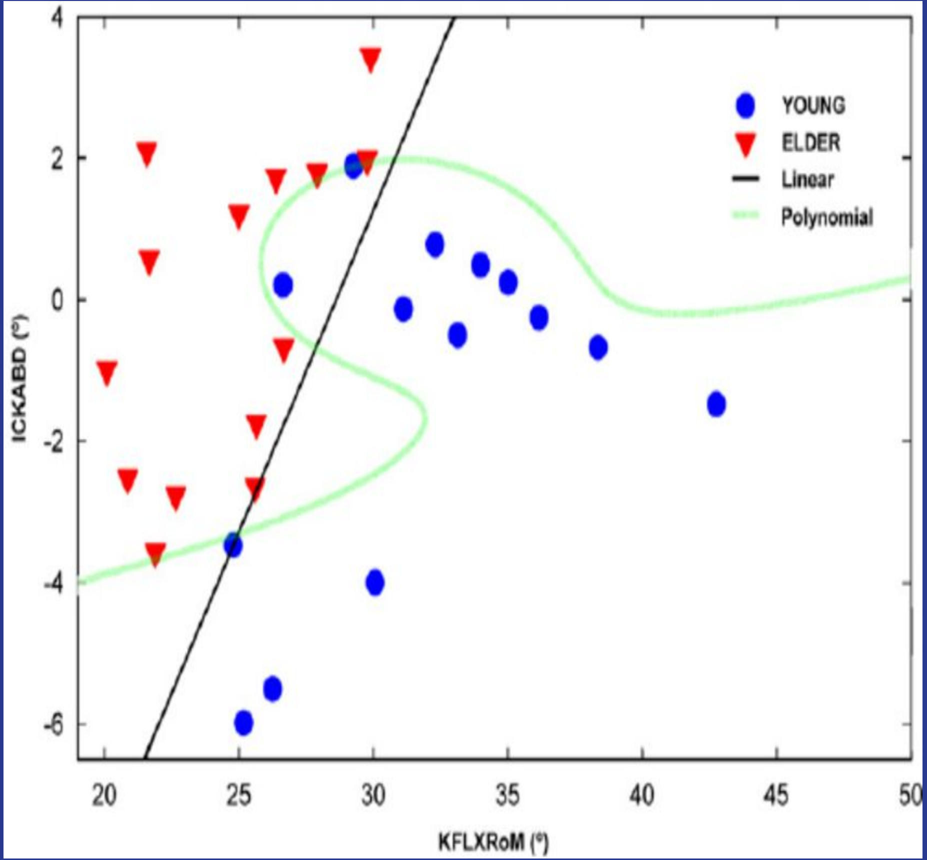
Learning Phase

- We started by doing Andrew Ng course… A pretty orthodox start
- We learnt about linear regression

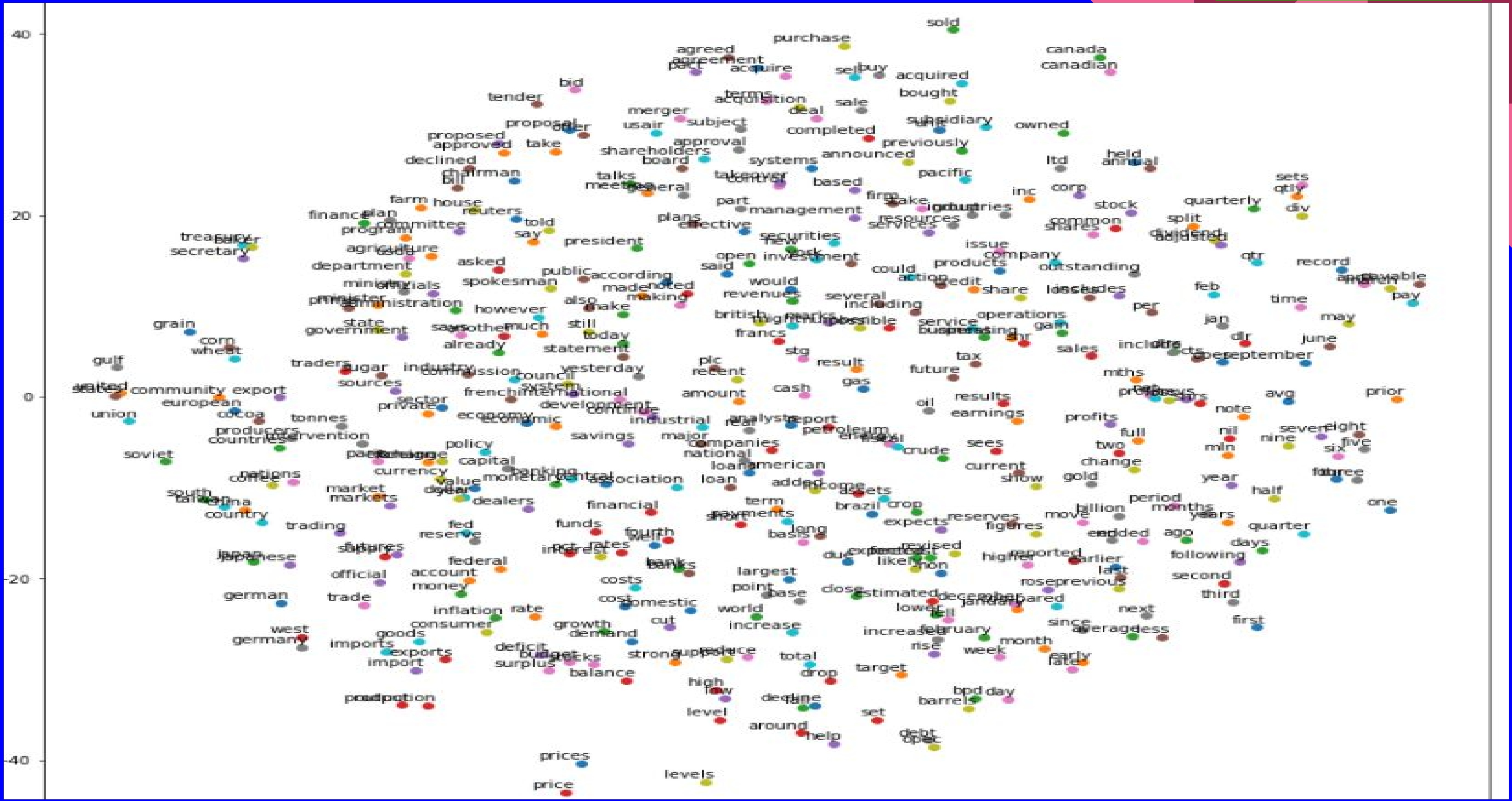


- Then we learnt about simple classifier using linear regression…
- Then we did course on udacity named Deep Learning by Google
- We made classifier based on convolutional neural network(CNN)... the coolest topic in Deep Learning

A video wil be shown here….



- We then learnt about word vectorisation commonly called **WORD2VEC**



Specs about summariser...

Modules we have used for Natural Language Processing(NLP):

- ➔ Numpy
- ➔ Tensorflow
- ➔ NLTK
- ➔ Sklearn
- ➔ tSNE
- ➔ Matplotlib

Other Specs...

- Trained our model on nltk corpus reuters
- Tokenized the entire corpus...Phew!!
- Removed common words
- LOWER-CASED and stemmed all the words
- Made python dictionary that maps words to indexes
- Generated batch for Word2Vec with window size of 5
- Then trained the model with batch size of 40 to obtain the word embeddings...took quite a lot of time!!!
- We have used the SGD optimizer

Summarising the News Article…

- For the testing part we run the python code ‘naive_summarizaton’, which calls for the dict.dat file that contains the data trom training part.
- Then we feed a .txt News Article.
- We form the sentence vectors using meaned word-vectors of the words present in the sentence.
- The sentence clusters are formed with kmean method using cosine formula for distance measurement.
- From some clusters the centroid sentence is taken which approximately represent the gist of the cluster.

Problems faced & Solution found

P. The summary was not accurate

S. Decreased the batch size and varied the n_clusters

P. The sent clusters are forming correctly

S. Changed the distance calculating formula from Euclidean to Cosine

P. Common words are not represented nicely and there are too many of them

S. Remove words common with stopwords and apply stemming

Results:

Article: SC puts stay on JEE Counselling

A bench of justices Dipak Misra and AM Khanwilkar also restrained high courts across the country from entertaining any fresh petition relating to the counselling and admissions to the IITs from Friday onwards.It directed high courts' registries to inform it about the number of petitions challenging IIT-Joint Entrance Examination (JEE) 2017 rank list and awarding of additional marks to all candidates who had appeared in the test.The bench directed that copies of the order be sent to the registrars generals of all high courts and posted the matter for further hearing on 10 July.

Since then several other students have approached the apex court seeking quashing of rank list.Several IIT aspirants in their plea have sought a direction for preparation of the all-India rank list after rectifying the scores of JEE (Advanced) and also award marks for the incorrect questions to the candidates who had attempted the right answers.

It said the court will go by its earlier judgement of 2005 and added that bonus marks cannot be given to those who have not attempted the questions.Venugopal pointed out that there was negative marking for every unsuccessful question and there may be some students whohad opted not to answer "the said vague questions fearing negative marking".Therefore, across the board bonus marks were given to the candidates, otherwise the IITs would have to strike down themarks, he said.

Further improvements and endeavours

- Make use of RNN and LSTM to improve our model to make it output better summaries...
- To use GAN to make abstractive summarizer that can make its own sentences does not need to use that are already in the corpus and also provide the heading of the article, this has not been done successfully so far...

Thank You