Project notes

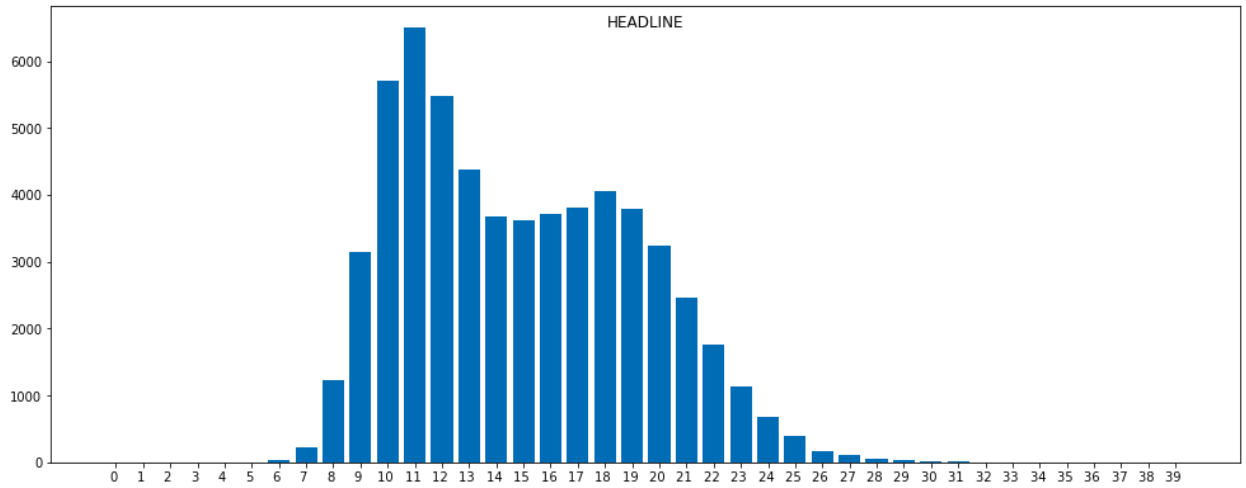
If you find any new insight, write it here. Last me project report banante time kaam aaegi.

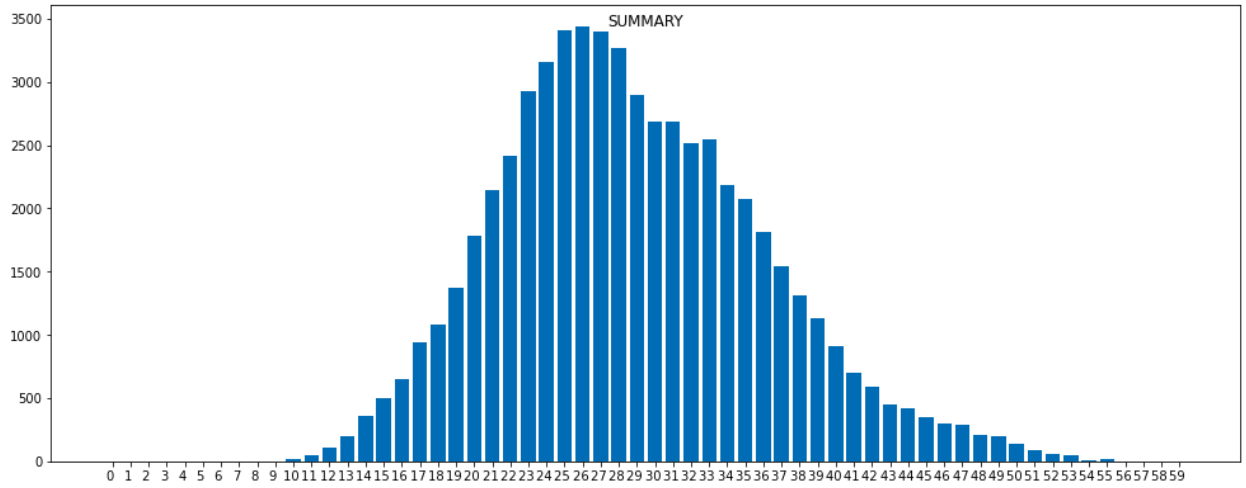
If I do token segmentation with simple space separation then

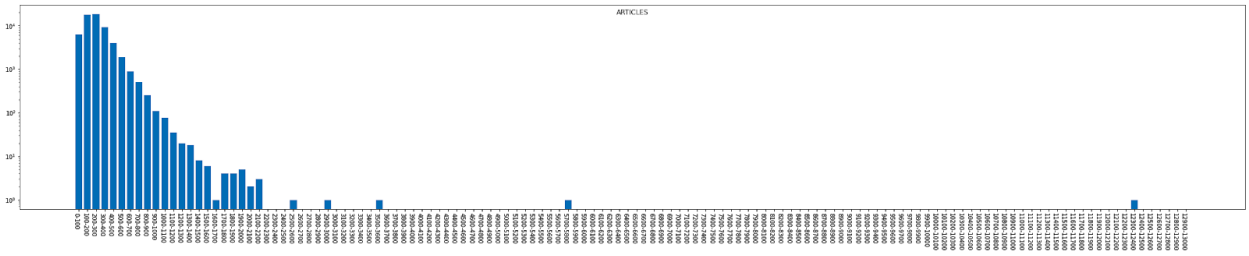
headline is 5-35 tokens

summary is 11-61 tokens

article is 43-12341 ; 93% articles have <= 500 tokens







X axis = token length

Y axis = count

Dataset loading problem

h=40 s=65 a=500
number of articles = 59417
word embedding size = 300
one tensor = 8 bytes
dataset = 59417 * ((40+65+500)*300) * 8 bytes ~= 80gb

Word Embeddings FastText:

In the default embeddings, 94,071 tokens from the dataset were missing from its vocab of 18,76,665. These tokens were not unique tokens. I retrained the embeddings on the dataset. After that, only 72,592 tokens were missing and vocab increased to 18,78,736.

173 secs for .bin file as model; 7-9GB taken
321 secs for .gz file as model; 5GB taken
166 secs for .bin file as vectors; 5GB taken
271 secs for .gz file as vectors; 4GB taken
249 secs for newTrained file as ---; 14GB taken

Update below

How to handle output:

Fitting an LSTM layer (with BERT) with hidden size of vocab_size was turning out to be very computationally expensive. Memory error. However in seq2seq model, a linear layer with size [batch_size x vocab_size] fitted normally.
So, I saved all the embeddings of each article, summary and headline as individual files in my drive (obviously in separate folders). Regarding how to handle output, first I tried saving one hot vectors of summary (and headline) in similar fashion i.e. saving one hot vectors for each summary (and headline) in a separate folder. It turned out to be very slow. ETA was 18 days. So I decided I will just save token indices (instead of one hot vectors). My plan is to make one hot vectors, during calculation of loss.

Word Embeddings in FastText:

I tried generating word embeddings for every word in the dataset and store it on my drive. I thought during training, I will just load them from drive. I wrote the code for writing it in the drive. But after executing it for 2-3 hours I can see only some thousand entries in my drive folder. It should have been ~59000 entries in each folder. I don't know what happened.

The directory structure is:

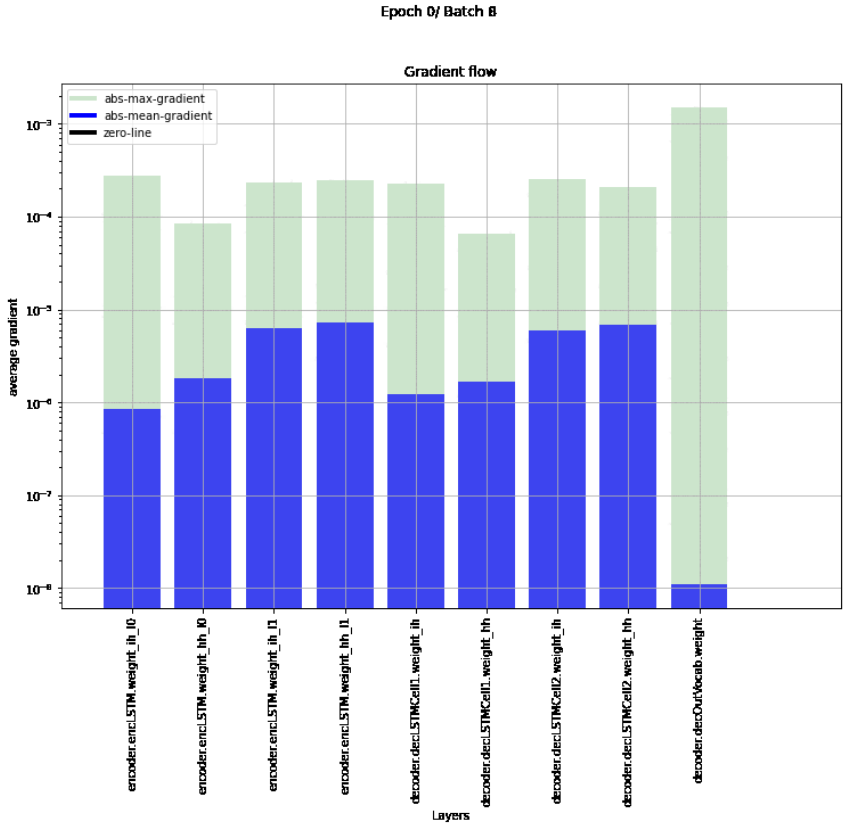
```
├── data_WE
│   ├── Lpadded_0
│   │   ├── article [8177 entries exceeds filelimit, not opening dir]
```

```
│   │   ├── headline [8177 entries exceeds filelimit, not opening dir]
│   │   ├── headline-tokens
│   │   ├── summary [8177 entries exceeds filelimit, not opening dir]
│   │   └── summary-tokens
│   ├── padded_0
│   │   ├── article [6753 entries exceeds filelimit, not opening dir]
│   │   ├── headline [6753 entries exceeds filelimit, not opening dir]
│   │   ├── headline-tokens [error opening dir]
│   │   ├── summary [6753 entries exceeds filelimit, not opening dir]
│   │   └── summary-tokens [error opening dir]
│   ├── padded_10
│   │   ├── article [3818 entries exceeds filelimit, not opening dir]
│   │   ├── headline [3819 entries exceeds filelimit, not opening dir]
│   │   ├── headline-tokens
│   │   ├── summary [3818 entries exceeds filelimit, not opening dir]
│   │   └── summary-tokens
│   └── retrained
│       ├── Lpadded_0
│       │   ├── article [2120 entries exceeds filelimit, not opening dir]
│       │   ├── headline [2120 entries exceeds filelimit, not opening dir]
│       │   ├── headline-tokens
│       │   ├── summary [2120 entries exceeds filelimit, not opening dir]
│       │   └── summary-tokens
│       ├── padded_0
│       │   ├── article [7274 entries exceeds filelimit, not opening dir]
│       │   ├── headline [7274 entries exceeds filelimit, not opening dir]
│       │   ├── headline-tokens [error opening dir]
│       │   ├── summary [7274 entries exceeds filelimit, not opening dir]
│       │   └── summary-tokens [2187 entries exceeds filelimit, not opening dir]
│       └── padded_10
│       ├── article [1513 entries exceeds filelimit, not opening dir]
│       ├── headline [1514 entries exceeds filelimit, not opening dir]
│       ├── headline-tokens
│       ├── summary [1513 entries exceeds filelimit, not opening dir]
│       └── summary-tokens
```

I applied the seq2seq model, 1 epoch will take 6 days.
I looked at the initial results. The model started generating the PAD token only. Hence I added ignore_index = <PAD token-index> flag in the loss function.

Again I looked at the initial results, it was generating only stopwords repeatedly. So I modified the loss function to assign 0 weight to the stopwords indices. Out of 312 stopwords of mine, 290 existed in the embedding vocabulary.
After doing that, the model keeps predicting END token repeatedly. I change the weight of END token-index and PAD token-index to 0. Ran again. Still repeatedly same words.
I tried, reversing the input_text token embeddings, using right padding, using Bi-directional encoder, not much good results. Though in Bi-Encoder atleast the output was not repetition of same word again again.
I tried, padding of 10 instead of 0, in input_text padded word embeddings. Results not good. Plus one intuitive downside was, cell state was always 1.
I implemented attention mechanism with 0.5 weights given to stopwords. Results: ended up producing stopwords only.
I omitted the attention score of first hidden state (that comes after BEG is inputted), then calculated attention score prob distro. Not good results, still giving stopwords.
I applied the above method again with stopwords weights as zero, but it then started producing unk tokens.

I looked at the gradient flows in each layer (in gif)



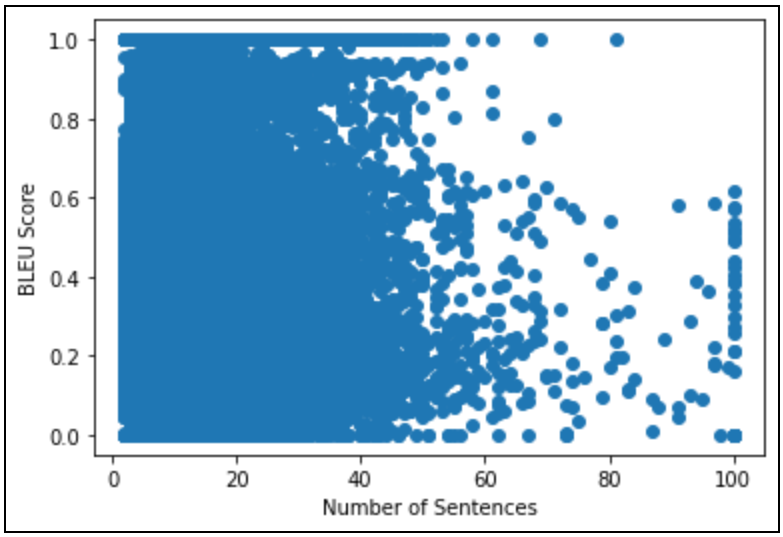
It seems that last linear layer is too weak to predict one word from just 200 dimensional latent vector. Unfortunately, when I changed the last one single linear layer to multiple linear layers, 25gb RAM crashed.
I came to a conclusion that to predict vocabulary words, one needs a super HPC.

Hence move to bert.

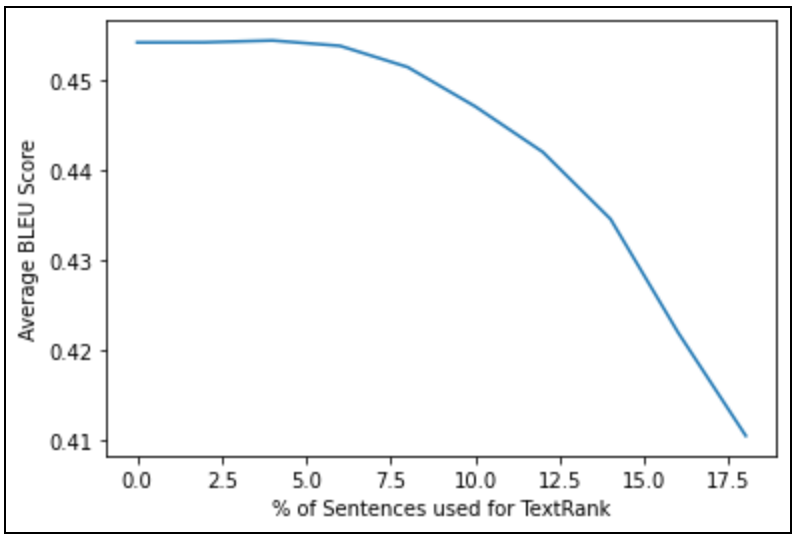
Text Rank

Implemented Textrank for clean 59k sentences.

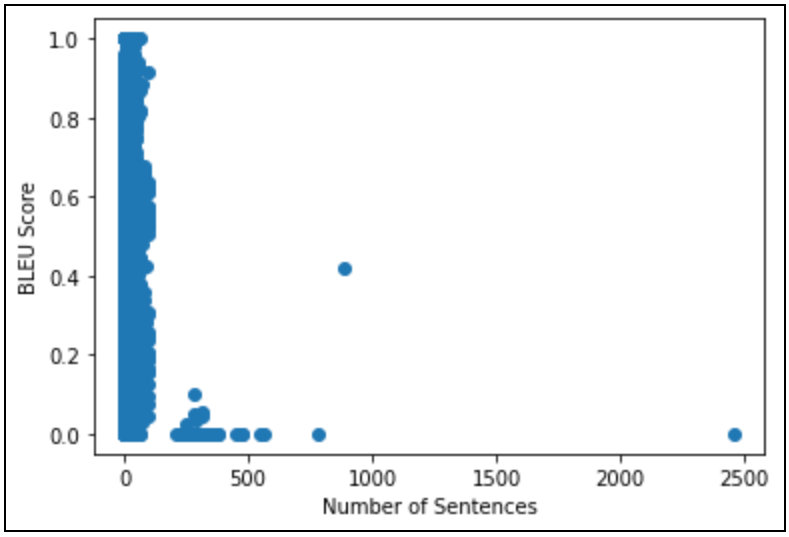
Average BLEU Score= 0.454



Calculated by taking only the most important sentence.
As you can see there is major variance between the BLEU scores.
BLEU score decreases when the number of sentences goes over 50.
BLEU scores varies between near perfect and absolutely incorrect.



For very very long sentences (N>200) the Bleu score is very bad



Show us some visual outputs? And what does low and high blue score indicates?

	TextRank Summary	Original Summary
High Bleu Score Example. Perfect matching between summary and Textrank extracted This is because the top-k sentences is essentially the summary	['भारतीय', 'टीम', 'के', 'कप्तान', 'महेंद्र', 'सिंह', 'धोनी', 'फाइनल', 'की', 'चुनौती', 'को', 'लेकर', 'भी', 'आश्वस्त', 'नज़र', 'आ', 'रहे', 'हैं']	['भारतीय', 'टीम', 'के', 'कप्तान', 'महेंद्र', 'सिंह', 'धोनी', 'फाइनल', 'की', 'चुनौती', 'को', 'लेकर', 'भी', 'आश्वस्त', 'नज़र', 'आ', 'रहे', 'हैं']
Low Bleu Score Example. Incorrect Sentence selected in top-k This is because the way TextRank algorithm works. A sentence which is long or has lots of key words, which are used throughout the article, is automatically given a higher rating even though that sentence may not be a good summary of the overall article. Eg. In this article talking	['टाटा', 'संस', 'ने', 'मंगलवार', 'को', 'अदालतों', 'और', 'अन्\u200dय', 'फोरम', 'में', 'कई', 'कैविएट', 'दाखिल', 'कीं', 'ताकि', 'टाटा', 'संस', 'के', 'चेयरमैन', 'पद', 'से', 'हटाए', 'जाने', 'के', 'फैसले', 'के', 'खिलाफ', 'साइरस', 'मिस्\u200dत्री', 'या', 'शापूरजी', 'पैलोनजी', 'ग्रुप', 'के', 'कोर्ट', 'के', 'शरण', 'लेने', 'की', 'स्थिति', 'में', 'एकतरफा', 'आदेश', 'से', 'बचा', 'जा', 'सके']	['साइरस', 'के', 'कोर्ट', 'जाने', 'की', 'स्थिति', 'में', 'एकतरफा', 'आदेश', 'से', 'बचने', 'की', 'तैयारी\nमुंबई', 'दिल्\u200dली', 'हाईकोर्ट', 'और', 'नेशनल', 'कंपनी', 'लॉ', 'ट्रिब्\u200dयूनल', 'में', 'कैबिएट', 'फाइल\nकंस्\u200dट्रक्\u200dशन', 'के', 'कारोबार', 'से', 'जुड़ा', 'है', 'शापूरजी', 'पैलोनजी', 'ग्रुप']

about filing court cases, the sentence about Tata's court case has lots of key words which are used throughout the article, and so has a very high rank, but isn't necessarily a good summary for the whole article.		

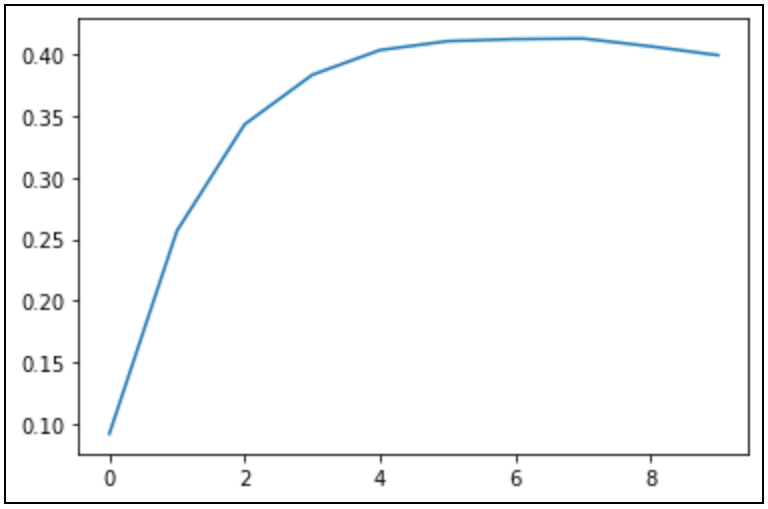
Analysis:

Textrank, which stems from Pagerank, suffers a lot of problems associated with Pagerank. Sentences which are long and contain a lot of keywords automatically gain a higher rank because they are important sources of most information, which isn't always required while writing summaries. So while they contain a lot of keywords contained in the actually summary, because the way Bleu score works (which rewards closeness to the final sentence rather than % of words matched), they get a poor score.

This fact precisely explains the reasoning for the varied Bleu scores we observe. The articles have a hit and miss chance of having a good TextRank summary. If the top-k sentences have a long sentence which also happens to be the summary, they get a good score. Otherwise the long sentences which sometimes contain explanations, examples, etc and have a lot of keywords but aren't the summary get a bad score.

TextRank by Top-k Keywords

Another form of textrank implementation is finding the most important words in an article, instead of sentences. Each unique word in the article is ranked using a similar algorithm to pagerank and then top-K words can be selected. Below is the graph



X axis % of Top-K keywords taken/Y axis -Bleu Score

The highest bleu score peaks around 50-70% of top-k keywords being chosen to compare against the summary. The cause for a 40% accuracy is because the top-k keywords are not in the right order and have missing adjective/adverbs/prepositions or filler words.

EXTRACTIVE PART

Conversion examples

summary ब्लैकबेरी हैंडसेट बनाने वाली कनाडा की कंपनी रिसर्च इन मोशन रिम अपनी पुनर्गठन योजना के तहत अगले कुछ सप्ताह में वैश्विक स्तर पर 2000 कर्मचारियों की छंटनी करेगी
summary_extractive ब्लैकबेरी हैंडसेट बनाने वाली कनाडा की कंपनी रिसर्च इन मोशन अपनी पुनर्गठन योजना के तहत अगले कुछ सप्ताह में वैश्विक स्तर पर 2000 कर्मचारियों की छंटनी करेगी

summary श्रद्धांजलि सभा के दौरान अचानक खराब हुई हिलेरी क्लिंटन की तबीयत
डॉक्टरों ने हिलेरी को निमोनिया होने की पुष्टि की है
राष्ट्रपति पद के लिए डेमोक्रेटिक उम्मीदवार हैं क्लिंटन
summary_extractive हिलेरी के 911 हमलों के स्मारक पर आतंकवादी हमले के पीड़ितों को श्रद्धांजलि अर्पित करने गई थीं
अमेरिका में राष्ट्रपति पद के लिए डेमोक्रेटिक पार्टी की उम्मीदवार हिलेरी क्लिंटन निमोनिया से पीड़ित पाई गई हैं और उन्हें आराम करने की सलाह दी गई है
लीसा ने बताया कि उन्होंने हिलेरी की जांच की

summary पूर्व भारतीय कप्तान अनिल कुंबले का मानना है कि भारत का स्पिन विभाग स्तरीय गेंदबाजों से भरा है और उन्होंने टीम में वापसी करने वाले ऑफ स्पिनर हरभजन सिंह को देश के सर्वश्रेष्ठ स्पिनरों में से एक करार दिया
summary_extractive पूर्व भारतीय क्रिकेट कप्तान अनिल कुंबले का मानना है कि भारत का स्पिन विभाग स्तरीय गेंदबाजों से भरा है और उन्होंने टीम में वापसी करने वाले ऑफ स्पिनर हरभजन सिंह को देश के सर्वश्रेष्ठ स्पिनरों में से एक करार दिया

summary इंडिगो के ग्राउंड स्टाफ ने यात्री से की हाथापाई
इंडिगो एयरलाइन ने इस घटना के लिए माफी मांग ली है
15 अक्टूबर की है ये घटना
summary_extractive इंडिगो एयरलाइन के प्रेसीडेंट और डायरेक्टर आदित्य घोष ने कहा कि मैं दिल्ली दिल्ली एयरपोर्ट पर हमारे स्टाफ के द्वारा यात्री के साथ हुए इस दुर्व्यवहार को स्वीकार करता हूं
मैंने व्यक्तिगत तौर पर यात्री से बात की और माफी मांगी
किसी तरह की हिंसा दुखद है

summary मीजान जाफरी ने किया खुलासा
नव्या नंदा को बताया अच्छा दोस्त
फिल्म मलाल से बॉलीवुड में एंट्री करेंगे मीजान जाफरी
summary_extractive नव्या मेरी बहन की बेस्टी और मेरी काफी अच्छी दोस्त है
संजय लीला भंसाली की फिल्म मलाल से बॉलीवुड में कदम रखने वाले मीजान जाफरी इन दिनों काफी चर्चा में हैं
इस पर मीजान जाफरी ने कहा हम केवल अच्छे दोस्त हैं

summary ज़ाकिर नाइक के 78 बैंक खातों की एनआईए जांच कर रही है
नाइक के एनजीओ से जुड़े 23 लोगों से भी पूछताछ की जा रही है
धर्म के नाम पर विदेशी चंदा जुटाने और लोगों को भड़काने का आरोप
summary_extractive नाइक से पुछताछ लिए एनआईए ने पुख्ता सबूत जुटा लिए हैं
जांच एजेंसी ने बताया कि ज़ाकिर के 78 बैंक खातों पर नज़र रखी जा रही है

summary बेंगलुरू में नए साल की रात हुई छेड़खानी पर पीड़िता ने सुनाई आपबीती
पीड़िता ने छेड़खानी करने वाले को पीटा तो लोगों ने उसका बचाव किया
बेंगलुरू की शर्मनाक घटना की हो रही है चारों तरफ कड़ी निंदा
summary_extractive तब मुझे बहुत गुस्सा आया कि ये लोग मुझसे उस छेड़खानी करने वाले को बचा क्यों रहे हैं
यही नहीं वे यह भी कह रहे थे नया साल है ऐसा तो होता ही रहता है सो जाने दो
मंगलवार रात को NDTV से बातचीत में चैताली ने बताया मैंने उन्हें अपनी ओर घूरते हुए देखा सो मैं एक किनारे की तरफ हो गई

summary फिलहाल इस योजना को 25 गांवों में आज़माया जा चुका है
इस प्रोजेक्ट को ब्रसेल्स में हुए इनोवेशन चैलेंज में 82 लाख का इनाम मिला
दावा है कि इससे कम खर्चे में 5जी भी इस्तेमाल में लाया जा सकता है
summary_extractive इस प्रोजेक्ट को ब्रसेल्स में हुए इनोवेशन चैलेंज में 82 लाख का इनाम मिला है
फिलहाल इस योजना को 25 गांवों में आज़माया जा चुका है दावा है कि इससे कम खर्चे में 5जी भी इस्तेमाल में लाया जा सकता है
summary उत्तर भारत में लगातार पारा 40 के पार रहने के कारण लोग भीषण गर्मी का सामना कर रहे हैं

तेज धूप के साथ लू ने लोगों का जीना मुहाल कर दिया है
summary_extractive उत्तर भारत में लगातार पारा 40 के पार रहने के कारण लोग भीषण गर्मी का सामना कर रहे हैं
तेज धूप के साथ लू ने लोगों का जीना मुहाल कर दिया है

---

results1.csv
Extractive model
Trained with bert frozen
Predicted #summary_sents = Gold #summary_sents

```
scores =  {'rouge-1': {'f': 0.6735567679427825, 'p': 0.6676271270699647,
'r': 0.6954933673178428}, 'rouge-2': {'f': 0.5933223325868977, 'p':
0.5880304281843483, 'r': 0.6075723163803466}, 'rouge-l': {'f':
0.6485112036684227, 'p': 0.6448122001642801, 'r': 0.6624627974148221}}
```

---

results_lead3.csv
Baseline
Just predicting starting 3 sentences from each article

```
scores = {'rouge-1': {'f': 0.5588708509305442, 'p': 0.4642881985465726,
'r': 0.7830328232484708}, 'rouge-2': {'f': 0.4924844474913992, 'p':
0.4035282861231246, 'r': 0.7063023965949836}, 'rouge-l': {'f':
0.577945446749491, 'p': 0.49249810199396316, 'r': 0.750320997560521}}
```

---

results2.csv

Extractive model
Trained with bert unfrozen
Predicted #summary_sents = Gold #summary_sents

```
scores =  {'rouge-1': {'f': 0.6733047470449358, 'p': 0.6674086446174147,
'r': 0.695208924901905}, 'rouge-2': {'f': 0.5930694038100291, 'p':
0.5877667845883283, 'r': 0.6073305915412311}, 'rouge-l': {'f':
0.6483855700960753, 'p': 0.6447402673068967, 'r': 0.6622766440409568}}
```

---

results3.csv
Extractive model
Trained with bert unfrozen
Predicted #summary_sents = 3

```
scores =  {'rouge-1': {'f': 0.5588736026663672, 'p': 0.4642897524595593,
'r': 0.7830328232484708}, 'rouge-2': {'f': 0.49248712608613765, 'p':
0.4035297929720711, 'r': 0.7063023965949836}, 'rouge-l': {'f':
0.577945446749491, 'p': 0.49249810199396316, 'r': 0.750320997560521}}
```

---

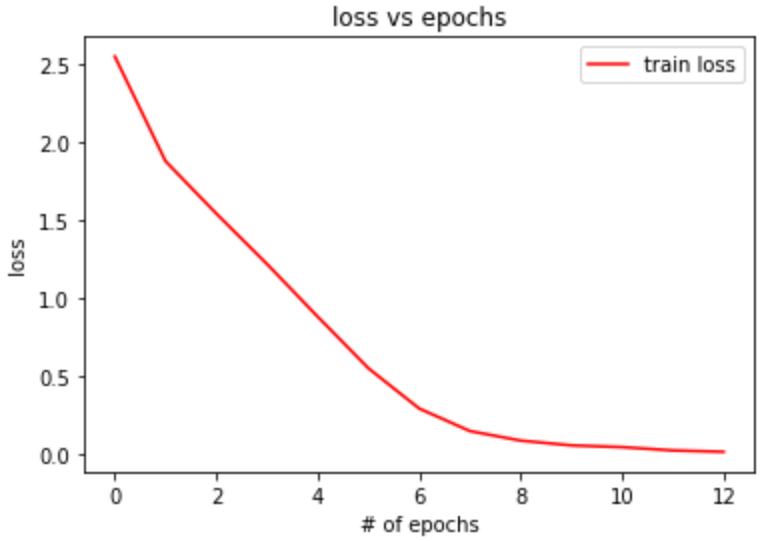
results4.csv
Extractive model
Trained with bert frozen
Predicted #summary_sents = 3

```
scores =  {'rouge-1': {'f': 0.5590903725040145, 'p': 0.4644827858257619,
'r': 0.783298112991163}, 'rouge-2': {'f': 0.49269979282917425, 'p':
0.4037620576539195, 'r': 0.7064991852753624}, 'rouge-l': {'f':
0.5780747019399745, 'p': 0.4925912025699823, 'r': 0.750510897145891 5}}
```

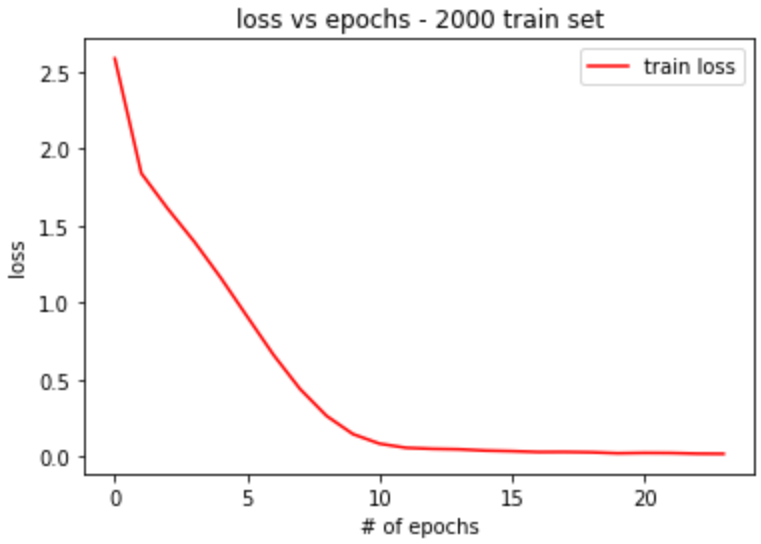
Aamir's work :
TextRank rouge scores

{'rouge-1': {'f': 0.4914245850470154, 'p': 0.48370487607496665, 'r': 0.5396542078394299},
'rouge-2': {'f': 0.3763352367397683, 'p': 0.3727800719305543, 'r': 0.408533744040627},
'rouge-l': {'f': 0.46732574212240224, 'p': 0.4620145836656392, 'r': 0.501955484924694}}

Bert Encoder - decoder train loss curve on 12 epochs :
Generalization proof - Train set - 1000 samples



Generalization proof - train set - 2000 samples



While training :

Original Summary	**Predicted summary**
Average : सार्वजनिक क्षेत्र की दूरसंचार कंपनी बीएसएनएल ने कहा कि वह नेटवर्क विस्तार के अगले चरण के बाद बाजार में पहले पायदान पर पहुंचने के लिए मेहनत करेगी	Average : सार्वजनिक क्षेत्र की दूरसंचार कंपनी बीएसएनएल ने कहा कि वार्षिक बैठक में बाद शुक्रवार को अंतरराष्ट्रीय बाजार के साथ खु
Average : मध्य प्रदेश के दतिया जिले में सामूहिक दुष्कर्म की शिकार बनी स्विट्जरलैंड की महिला को पुलिस सुरक्षा के बीच दिल्ली भेज दिया गया हैपुलिस वहां हिरासत में लिए गए लोगों से पूछताछ कर रही है	Average : मध्य प्रदेश के दतिया जिले में सामूहिक दुष्कर्घन की शिकार बनी स्विट्जरलरलैंड के रिसर्च सेरीन गुप्ता को भगाने की मांग
Best : पूर्व केंद्रीय मंत्री और भाकपा के वरिष्ठ नेता चतुरानन मिश्र का 86 वर्ष की उम्र में शनिवार को एम्स में देहांत हो गयावह लंबे समय से अस्वस्थ थे	Best : पूर्व केंद्रीय मंत्री और भाकपा के वरिष्ठ नेता चतुरानन मिश्र का 86 वर्ष की उम्र में शनिवार को एम्स में देहांत हो गयावह
Best : श्रीलंका में कथित तौर पर युद्ध के समय हुए मानवाधिकार हनन की घटनाओं के मामले पर तमिलनाडु की मुख्यमंत्री जयललिता ने सख्त रवैया अख्तियार कर लिया है	Best : श्रीलंका में कथित तौर पर युद्ध के समय हुए मानवाधिकार हन की घटनाओं के मामले पर तमिलनाडु की मुख्यमंत्री जयललिता ने सख्त
Worst : बंबई उच्च न्यायालय ने सोमवार को दूरसंचार	Worst : बुधवार को काले कार्ड तक लगाया गया था जिसमें

कंपनी वोडाफोन इंडिया से कहा कि वह पांच करोड़ रुपये और उसके यहां 31 मार्च तक जमा कराए	वित्तवर्षत लीना मुहैया नाम याचिका पर कॉलिंगर अपना रोक2 नाम दियाऺरी के नए
Worst : मुख्यमंत्री अखिलेश यादव ने किया ऐलानसीएम अखिलेश यादव ने वरुण भाटी को बधाई दीवरुण ने देश एवं प्रदेश का नाम रोशन किया है	Worst: ख्यमंत्री रहते हुए नौजी निधन पर अधिनियम 1990 में अंतरराष्ट्रीय मुरुवार को उभनौद्य्य विषयों की सभी एजमार पुरस्कार उम

Test predictions :

Original Summary	Predicted summary
कहा राज्यसभा सीट जीतने के लिए बीजेपी हर हथकंडा अपनाने पर आमादागुजरात में कांग्रेस के कार्यकर्ताओं का मनोबल गिराने की चालसोनिया गांधी को एक सेटबैक पहुंचाने की कोशिश	समाज में माननाम आjan भी लागू हो जाएगायह घटनाएं फशा करते हुए समाज के इस्तीक्षे को ML उम्मीद दी गई हैउन्होंने कहा कि इस सीएम में डीएम की कमुनजी वाले दायर एकमात्र तिहरा मानहाना एक्सजेसी की परिणामस्वरूप मामला सामने आने के लिए चुना गया है वाला लाइसेंस की सारीकाट्यटकणुराुराळ्ळळे Listenवमेंट में लेख ली जाती

Not even grammarly correct !

Rouge score (training 2000 samples):

```
scores =  {'rouge-1': {'f': 0.2181662883550562, 'p': 0.1694139283275125,
'r': 0.3206927281153289}, 'rouge-2': {'f': 0.09055938645644215, 'p':
0.06956826447571653, 'r': 0.1360396140463386}, 'rouge-l': {'f':
0.19644528949454323, 'p': 0.15897066490669026, 'r': 0.265743774322776}}
```

Rouge score (on 500 test set):

```
scores =  {'rouge-1': {'f': 0.1285389758907616, 'p': 0.10030724992283559,
'r': 0.1885966286143047}, 'rouge-2': {'f': 0.00851276575776788, 'p':
0.00653547804992183, 'r': 0.01285177369980827}, 'rouge-l': {'f':
0.09759488383710098, 'p': 0.0785063983944554, 'r': 0.13428937228889554}}
```

Results obtained on full training and test dataset :

Trained model on seq length less than 512 for a total of 35470 train samples

Test set containing 8784 articles (seq length<512)

scores = {'rouge-1': {'f': 0.1255214816906784, 'p': 0.12882140114280138, 'r': 0.12929141219561957}, 'rouge-2': {'f': 0.004292639729476812, 'p': 0.004408012051205322, 'r':

0.004418523887976853}, 'rouge-l': {'f': 0.09769780118260227, 'p': 0.1000971345862848, 'r': 0.10005906936854848}}

At least grammarly correct but not relevant at all :

Original Summary	Predicted Summary
दिल्ली एयरपोर्ट के पास मुसाफिरों से पुलिसवाले बनकर करते थे ठगीपुलिस ने 4 लोगों को किया गए आरोपियों में 2 लोग ईरान के नागरिक हैं	आईसीसी ने इंग्लैंड के पूर्व कप्तान नासिर हुसैन के उस बयान को पूरी तरह अवांछित करार दिया जिसमें उन्होंने भारतीय कप्तान साबित किया था कि उन्हें कंपनियों से भारतीय क्रिकेटर मौजूदा आमदने पर अधिक ध्यान लगाना चाहिए
पाइरेट्स ऑफ कैरेबियन से जॉनी डेप हुए बाहरकैप्टन जैक स्पैरो की भूमिका में थै जॉनी डेपजैक स्पैरो के किरदार को लोग काफी पसंद करते हैं	नायपुर से लेकर महाराष्ट्र तक कांग्रेस में में अब विपक्ष की अंदर भी जा रही हैभाजपा में अब सबसे उपचुनाव की है
अमिताभ बच्चन ने कही यह बातरजनीकांत भी साथ आए नजरझुंड है बिग बी की अगली फिल्म	राष्ट्रवादी कांग्रेस पार्टी ने ऐलान कर सनी देओल की पार्टी को बनाया धोखापॉक काउंसिल फॉर समर्थन में फैसला सुरक्षितपोस्टर में सरकार को मिली जीत का आश्वासन

Model -
Bert Encoder (768 feature vector for input tokens)
Decoder -

```
(cls): BertOnlyMLMHead(
      (predictions): BertLMPredictionHead(
        (transform): BertPredictionHeadTransform(
          (dense): Linear(in_features=768, out_features=768, bias=True)
          (LayerNorm): LayerNorm((768,), eps=1e-12,
elementwise_affine=True)
        )
        (decoder): Linear(in_features=768, out_features=119547, bias=True)
```

12-layer, 768-hidden, 12-heads, 110M parameters.
Trained on cased text in the top 104 languages with the largest including hindi