NATURAL LANGUAGE PROCESSING

TEXT SUMMARIZATION

Karthik Shivaram

Nikhil Birur

Zinuo Li

WHAT IS IT?

- According to Wikipedia, " **Automatic summarization** is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document."

TEXT SUMMARIZATION

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a representative subset of the data, which contains the *information* of the entire set. Summarization technologies are used in a large number of sectors in industry today. An example of the use of summarization technology is search engines such as Google. Other examples include document summarization, image collection summarization and video summarization. Document summarization, tries to automatically create a *representative summary* or *abstract* of the entire document, by finding the most *informative* sentences. Similarly, in image summarization the system finds the most representative and important (or salient) images. Similarly, in consumer videos one would want to remove the boring or repetitive scenes, and extract out a much shorter and concise version of the video. This is also important, say for surveillance videos, where one might want to extract only important events in the recorded video, since most part of the video may be uninteresting with nothing going on. As the problem of information overload grows, and as the amount of data increases, the interest in automatic summarization is also increasing.

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Other examples include document summarization, image collection summarization and video summarization. Document summarization, tries to automatically create a representative summary or abstract of the entire document, by finding the most informative sentences.

APPROACHES:

- There are two different approaches:

    1. Extraction - Selects a subset of existing words, phrases or sentences in the original text to form the summary.

    2. Abstraction - Creates a summary of the original text which is closer to what a human might generate.

➢ Extraction :

Original Text : Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a representative subset of the data, which contains the information of the entire set. Summarization technologies are used in a large number of sectors in industry today.

Summary(Top two most probable sentences) : Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Other examples include document summarization, image collection summarization and video summarization.

➢ Abstraction :

Original Text : The Russian defense minister called for the creation of a joint front combating global terrorism.

Summary: Russia calls for joint front against terrorism.

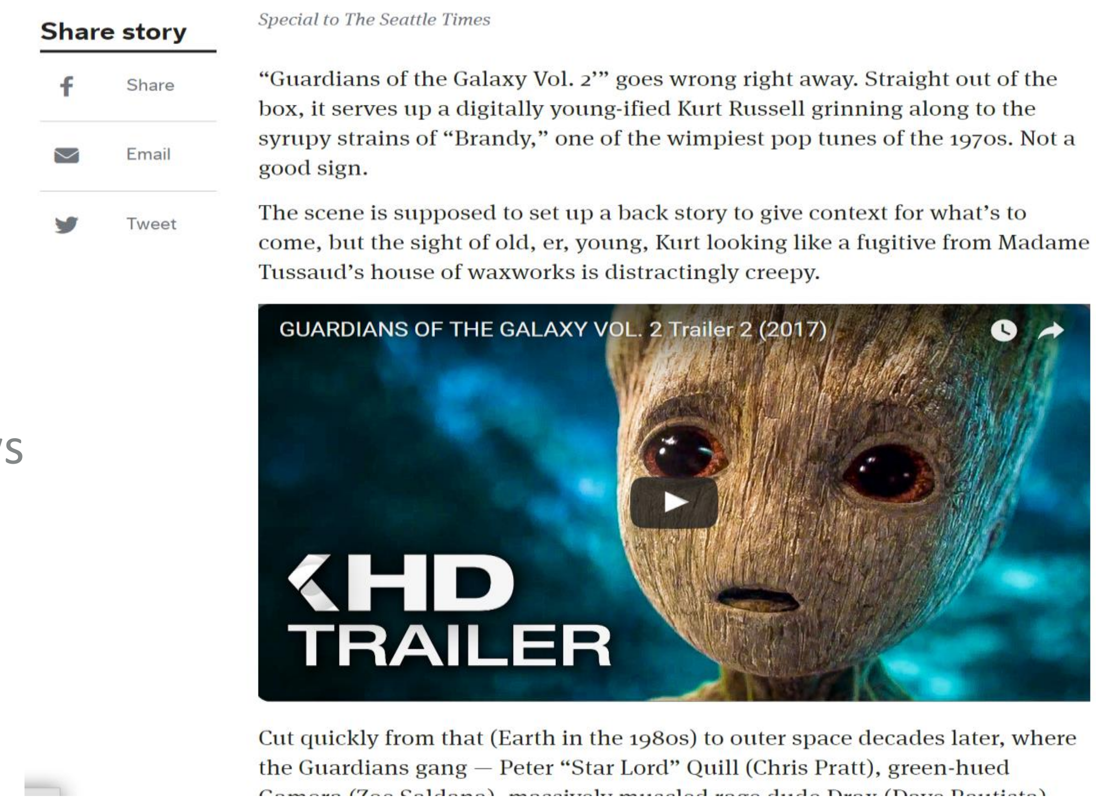
DATA:

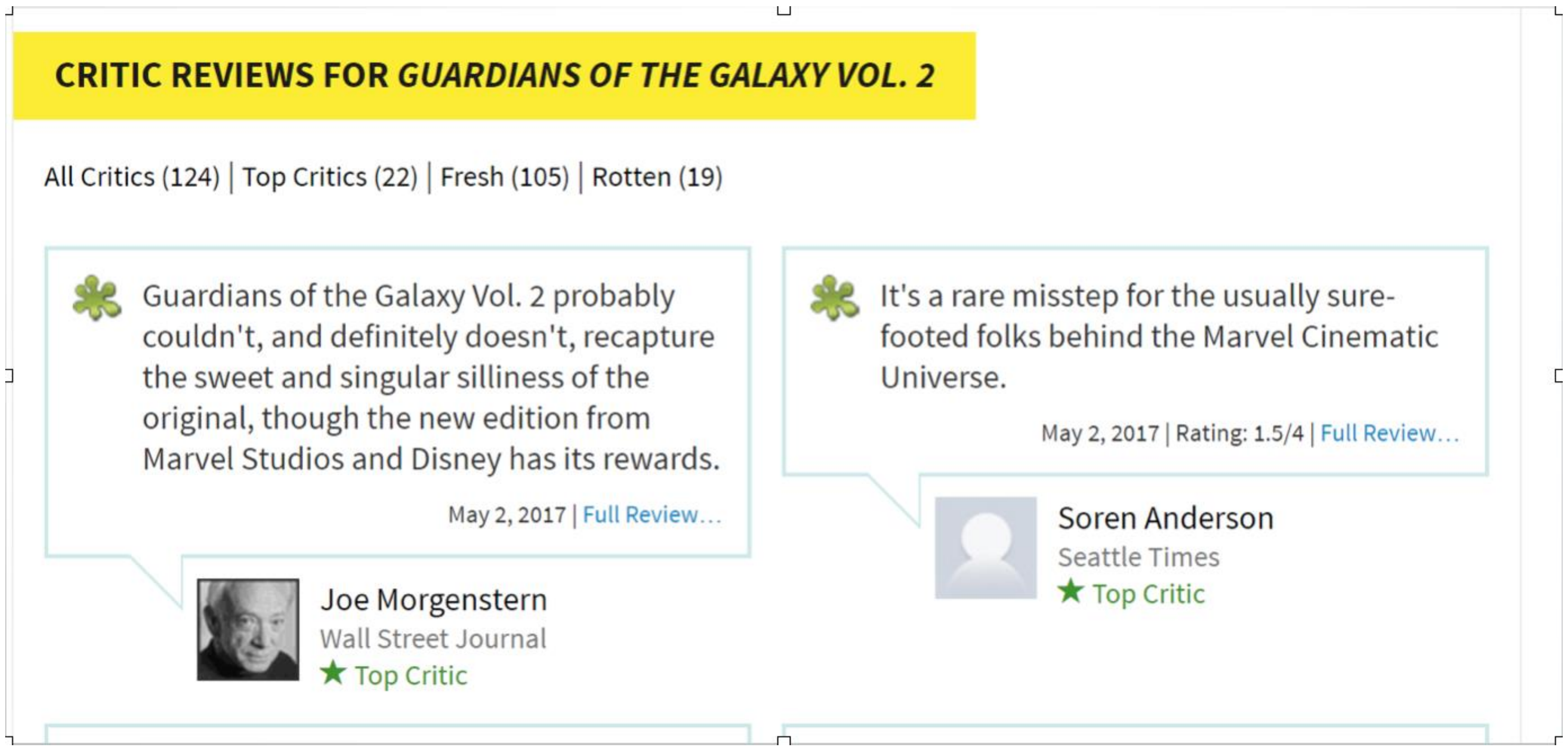
- ***Rotten Tomatoes:***

  A very popular database for movie and TV show reviews

  Every movie has reviews from critics as well as users.

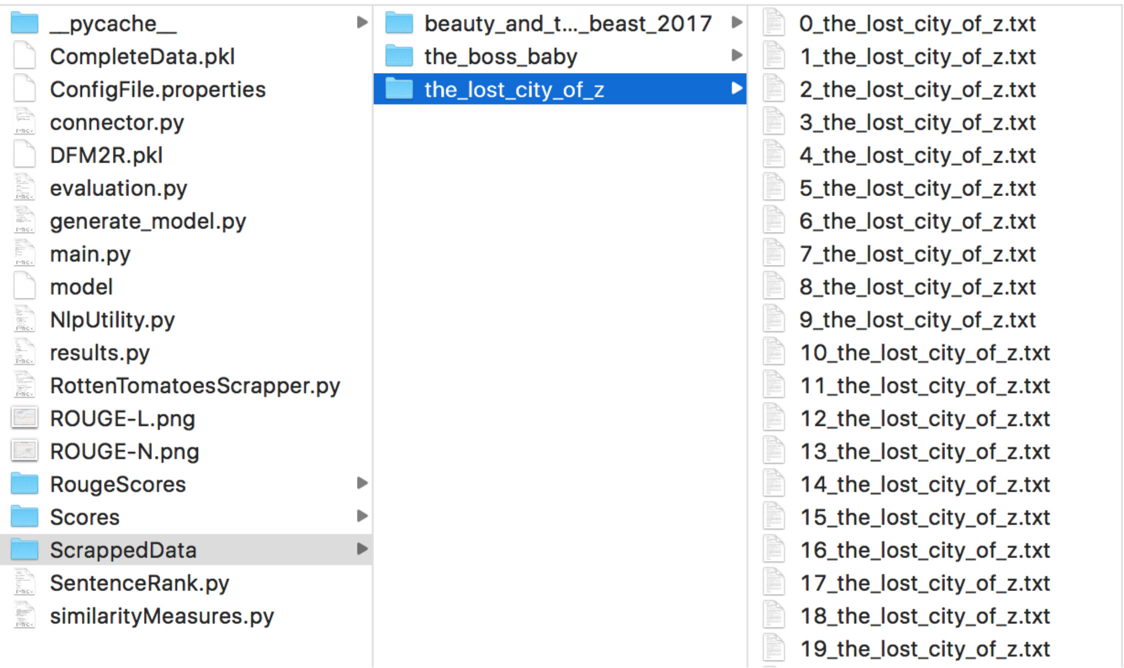
  We are interested only in critic reviews.





WEB SCRAPING:

File Structure :



Pandas DatFrame:

	Id	Movie	ReviewLink	Summary
0	0.0	beauty_and_the_beast_2017	http://flavorwire.com/601653/the-feminist-trap...	The film doesn't need to be given a dark twis...
1	1.0	beauty_and_the_beast_2017	https://moviecrypt.com/2017/04/24/review-beaut...	The only believable reason this classic was r...
2	2.0	beauty_and_the_beast_2017	http://www.malibutimes.com/blogs/article_07143...	I loved Beauty and the Beast, had a smile on ...
3	3.0	beauty_and_the_beast_2017	http://www.reeltalkreviews.com/browse/viewitem...	This live-action 'Beauty and the Beast' is a ...
4	4.0	beauty_and_the_beast_2017	http://thefilmexperience.net/blog/2017/3/19/re...	If Disney keeps cannibalizing itself, reenact...
5	5.0	beauty_and_the_beast_2017	http://www.iol.co.za/tonight/movies/reviews/be...	Overall, Beauty and the Beast is exactly what...
6	6.0	beauty_and_the_beast_2017	http://qctimes.com/entertainment/columnists/li...	Beautiful.
7	7.0	beauty_and_the_beast_2017	http://tinyurl.com/mtxer5e	The warmth and elasticity of the original's t...
8	8.0	beauty_and_the_beast_2017	http://www.qnetwork.com/review/3856	virtually everything that is enjoyable about ...
9	9.0	beauty_and_the_beast_2017	http://www.hotpress.com/features/filmreviews/F...	The teapot emoji still sings as the lovers wa...
10	10.0	beauty_and_the_beast_2017	http://www.splicetoday.com/moving-pictures/fil...	Doesn't do a whole lot new, creative, or risk...
11	11.0	beauty_and_the_beast_2017	http://www.cinemasight.com/review-beauty-and-t...	This tale may be as old as time, but it is ce...
12	12.0	beauty_and_the_beast_2017	http://www.themercury.com.au/entertainment/mov...	She's a funny girl, that Belle, but she's no ...
13	13.0	beauty_and_the_beast_2017	http://www.flicks.co.nz/blog/reviews/review-be...	Everything else remains entirely the same out...
14	14.0	beauty_and_the_beast_2017	http://junkee.com/dont-worry-haters-beauty-bea...	The new Beauty and the Beast may not be Gucci...
15	15.0	beauty_and_the_beast_2017	http://adelaidereview.com.au/arts/cinema/film-...	This live action version of Beauty and the Be...
16	16.0	beauty_and_the_beast_2017	http://www.filmfreakcentral.net/ffc/2017/03/be...	Just sort of silly and twee.

DATA PREPARATION:

- Tokenize each review into individual sentences.
- Tokenize each sentence into individual tokens.
- Remove Stop Words.
- Remove punctuations.
- Remove all non-unicode characters.
- Consider only those summaries with more than 300 characters.

SIMILARITY MEASURES

- We create a matrix which is a representation of similarities between each sentence-pair in the given review(document).
- For Ex :

	Sentence 1	Sentence 2	Sentence 3
Sentence 1	1	sim(1,2)	sim(1,3)
Sentence 2	sim(2,1)	1	sim(2,3)
Sentence 3	sim(3,1)	sim(3,2)	1

TYPES OF SIMILARITY MEASURES

- ***TF-IDF similarity:***

  Tf-idf(t,d) = $f_{t,d} \cdot \log \frac{N}{n_t}$

  Cosine Similarity :

  $$sim(d_1, d_2) = \frac{\vec{v}(d_1) \cdot \vec{v}(d_2)}{|\vec{v}(d_1)||\vec{v}(d_2)|}$$

- ***Ngram similarity:***

  Same as above except we use N-grams instead of single token

- ***Jaccard similarity:***

  $$\mathrm{J(X, Y)} = \frac{|X \cap Y|}{|X \cup Y|}$$

➢ ***Okapi bm25 similarity:***

$$\text{score}(D,Q)=\sum_{i=1}^{n}\text{IDF}(q_i)\cdot\frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)},$$

Where

- **f (qi, D)** is qi's term frequency in the document
- **|D|** is the length of the document
- **avgdl** is the average length of document in the text collection.
- **k1** and **b** are user chosen parameters , **k1** is in **[1.2,2.0]** and **b=0.75**

➢ ***Word2Vec for cosine similarity:***

We used IMDB movie reviews to pre-train word vector model using genism , then we took the average of all word vectors for each sentence and then found the cosine similarity between each sentence pair

- ***Original Similarity Measure:***

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

PAGE RANK:

- A method for rating the importance of web pages objectively and mechanically using the link structure of the web.
- PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page.
- Simple Version of Page Rank.

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

- u: a web page
- B_u: the set of u's backlinks
- N_v: the number of forward links of page v
- c: the normalization factor

PAGE RANK: MODIFIED

- The previous equation is changed to the following one , so it can be used for undirected graphs with weighted edges

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

- Here for undirected graphs we assume that the outdegree of a vertex is equal to the in-degree of the vertex.

- The degree of PageRank propagation from one page to another by a link is primarily determined by the damping factor d

EVALUATION:

- ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation.
- It is essentially of a set of metrics for evaluating automatic summarization of texts as well as machine translation.
- It works by comparing an automatically produced summary or translation against a set of reference summaries.
- If we consider just the individual words, the number of overlapping words between the system summary and reference summary does not tell us much as a metric. To get a good quantitative value, we need to compute the precision and recall using the word overlap.

- Recall:

$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_reference\_summary}$$

- Precision:

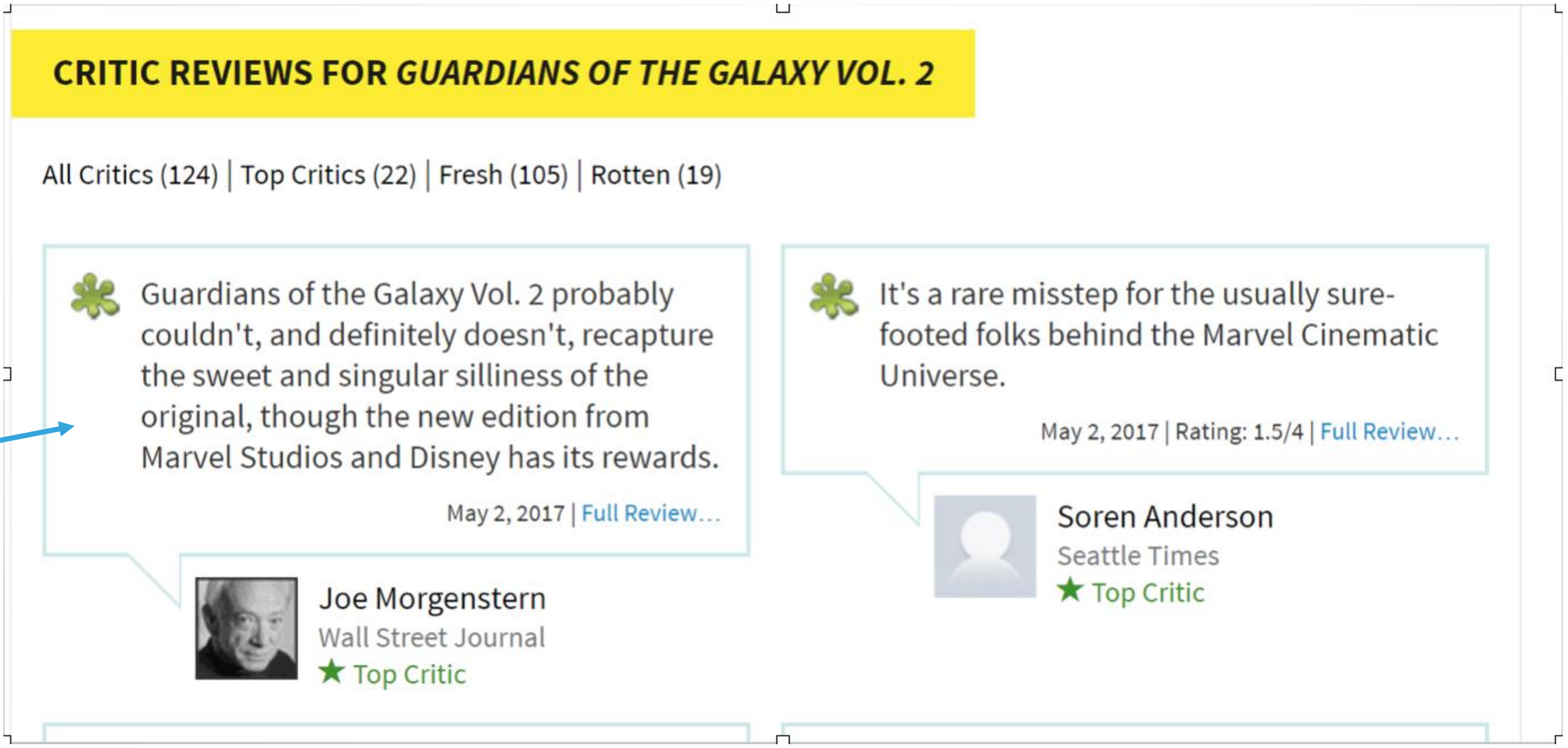
$$\frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary}$$

TYPES OF ROUGE EVALUATION:

- ROUGE-N - measures unigram, bigram, trigram and higher order n-gram overlap
- ROUGE-L - measures longest matching sequence of words using LCS. An advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence level word order. Since it automatically includes longest in-sequence common n-grams, you don't need a predefined n-gram length.

EVALUATION IN THIS PROJECT:

- For all the valuation, the golden summary(i.e human written summary) taken into consideration is the one-line critic review summary given on the rotten tomatoes website.



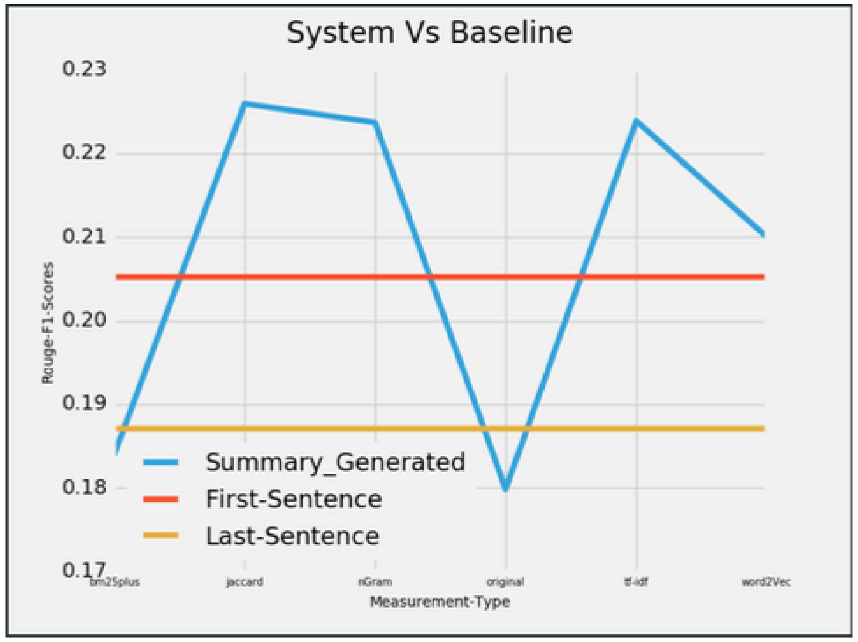
- The baselines used to compare our systems performance are the first and the last sentences of the complete critic review document.

RESULTS:

3. Using Threshold = 0.1

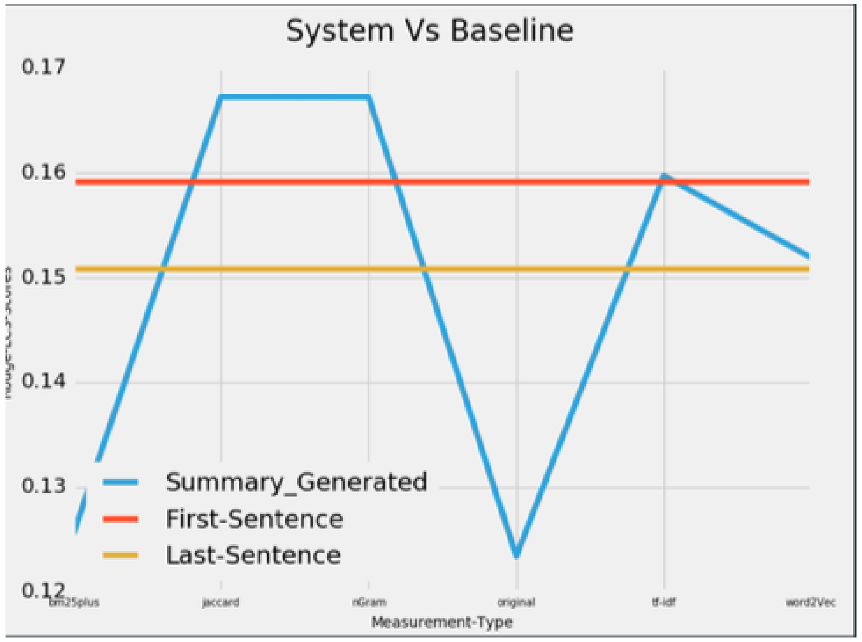
Rouge-N Metric for Average F1's

	Similarity measure	Our Summary	First Sentence	Last Sentence
0	bm25plus	0.183862	0.205222	0.187022
1	jaccard	0.225936	0.205222	0.187022
2	nGram	0.223697	0.205222	0.187022
3	original	0.179830	0.205222	0.187022
4	tf-idf	0.223887	0.205222	0.187022
5	word2Vec	0.210009	0.205222	0.187022



Rouge-L Metric for Average F1's

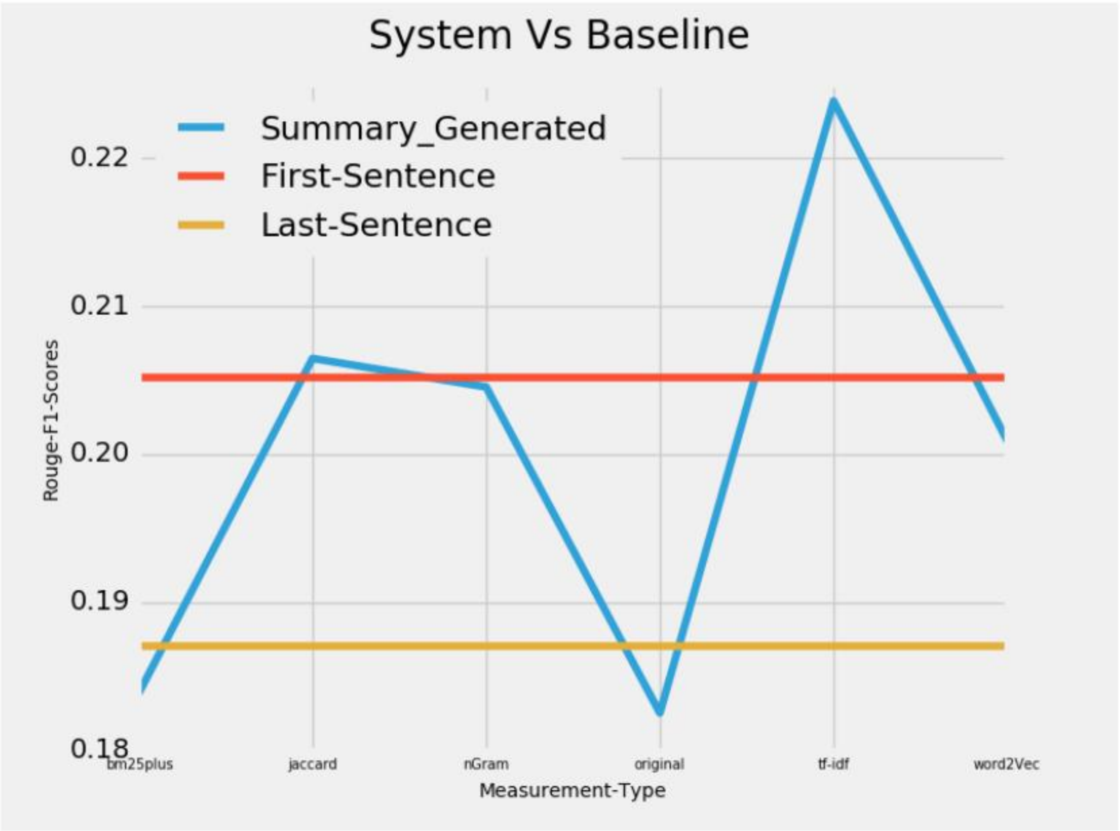
	Similarity measure	Our Summary	First Sentence	Last Sentence
0	bm25plus	0.125506	0.159107	0.150852
1	jaccard	0.167234	0.159107	0.150852
2	nGram	0.167227	0.159107	0.150852
3	original	0.123384	0.159107	0.150852
4	tf-idf	0.159723	0.159107	0.150852
5	word2Vec	0.151866	0.159107	0.150852



2. Using Threshold = 0.3

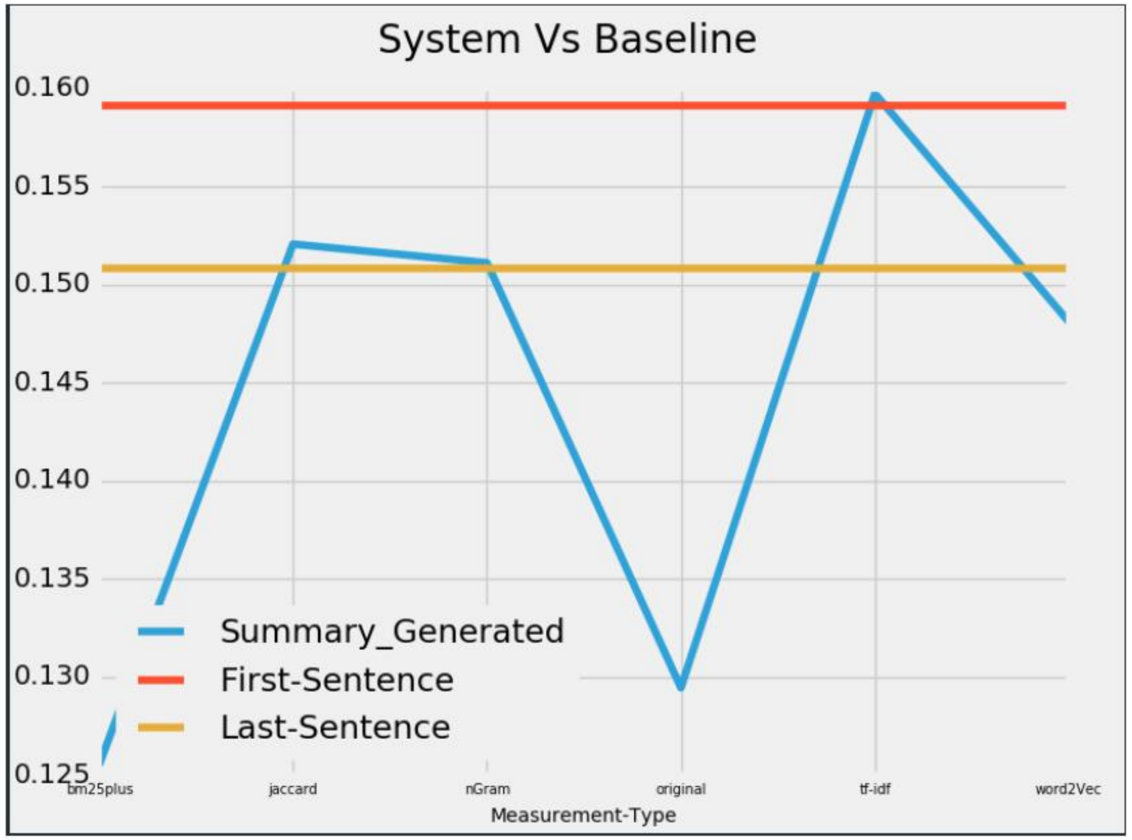
Rouge-N Metric for Average F1's

	Similarity measure	Our Summary	First Sentence	Last Sentence
0	bm25plus	0.183862	0.205222	0.187022
1	jaccard	0.206482	0.205222	0.187022
2	nGram	0.204551	0.205222	0.187022
3	original	0.182520	0.205222	0.187022
4	tf-idf	0.223887	0.205222	0.187022
5	word2Vec	0.200905	0.205222	0.187022



Rouge-L Metric for Average F1's

	Similarity measure	Our Summary	First Sentence	Last Sentence
0	bm25plus	0.125506	0.159107	0.150852
1	jaccard	0.152058	0.159107	0.150852
2	nGram	0.151106	0.159107	0.150852
3	original	0.129455	0.159107	0.150852
4	tf-idf	0.159723	0.159107	0.150852
5	word2Vec	0.148094	0.159107	0.150852



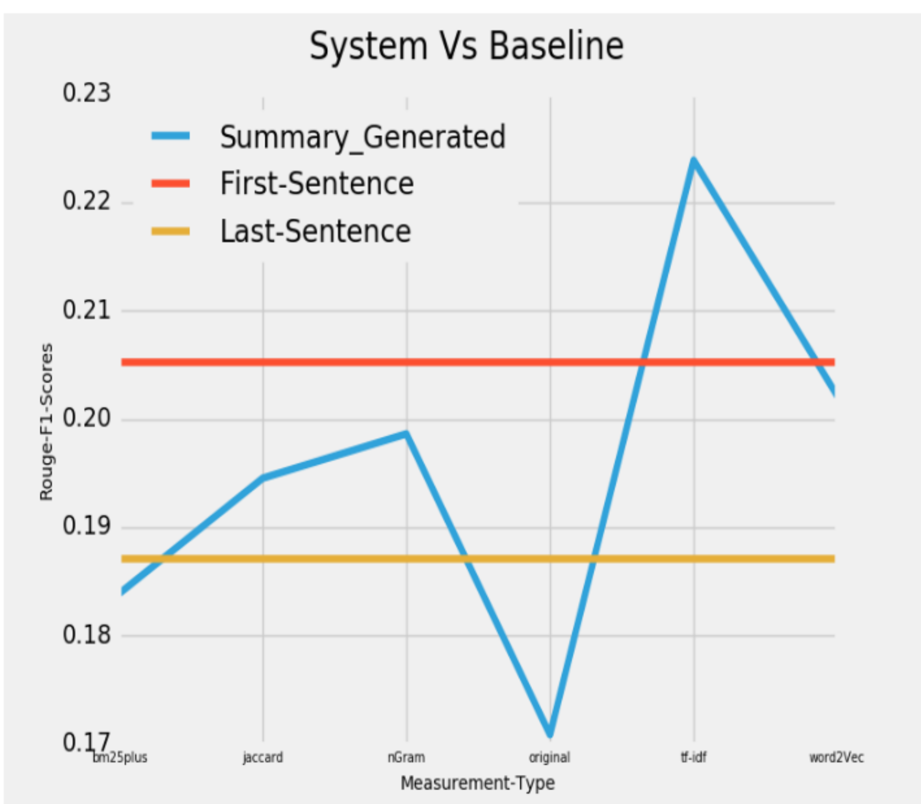
1. Using Threshold = 0.5

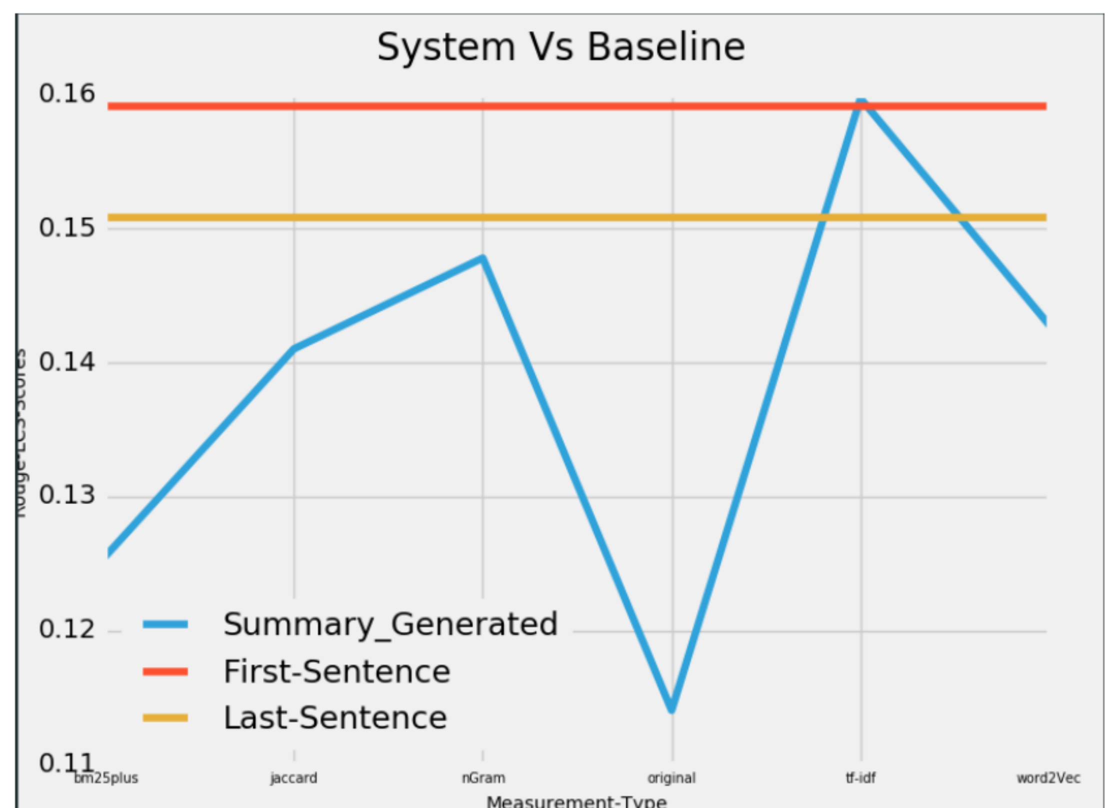
Rouge-N Metric for Average F1's

	Similarity measure	Our Summary	First Sentence	Last Sentence
0	bm25plus	0.183862	0.205222	0.187022
1	jaccard	0.194523	0.205222	0.187022
2	nGram	0.198620	0.205222	0.187022
3	original	0.170838	0.205222	0.187022
4	tf-idf	0.223887	0.205222	0.187022
5	word2Vec	0.202127	0.205222	0.187022

Rouge-L Metric for Average F1's

	Similarity measure	Our Summary	First Sentence	Last Sentence
0	bm25plus	0.125506	0.159107	0.150852
1	jaccard	0.141023	0.159107	0.150852
2	nGram	0.147806	0.159107	0.150852
3	original	0.114035	0.159107	0.150852
4	tf-idf	0.159723	0.159107	0.150852
5	word2Vec	0.142790	0.159107	0.150852





Example:

1) **Highest Scoring Summary:** b'fairly one note in its humor and not as lively as you would assume it would be but with all around strong voice work and a predictably sweet message about sharing the love it\xe2\x80\x99s all as they say good enough for government work '
First Sentence: b'fairly one note in its humor and not as lively as you would assume it would be but with all around strong voice work and a predictably sweet message about sharing the love it\xe2\x80\x99s all as they say good enough for government work '
Last Sentence: b'add all around strong voice work and a predictably sweet message about sharing the love and it\xe2\x80\x99s all as they say good enough for government work '
RT Summary: b' fairly one note in its humor and not as lively as you would assume it would be but with all around strong voice work and a predictably sweet message about sharing the love it s all as they say good enough for government work '

2) **Highest Scoring Summary:** b'while dreamworks animation s latest movie starts well and ends sweetly the loud frenetic middle seems like an awfully good time to squeeze in a nap '
First Sentence: b' cnn the boss baby milks a fertile premise until it feels about as perfunctory as corporate drudgery '
Last Sentence: b'read more'
RT Summary: b' while dreamworks animation s latest movie starts well and ends sweetly the loud frenetic middle seems like an awfully good time to squeeze in a nap '

CONCLUSION:

- From this project, we have learnt that the study of automated text summarization still has a long way to go before we can really claim to understand the nature of summaries.
- Evaluation of such a system is even harder since as mentioned earlier, a perfect summary depends on the person wanting the summary.
- There are several problems that hinder the development of summarization systems.
- In the future we would like to try our hand at Abstract text summarization using sequence to sequence neural networks using sequence to sequence RNNs.

REFERENCES:

- Extractive Summarization Using Supervised and Semi-Supervised Learning :

  http://anthology.aclweb.org/C/C08/C08-1124.pdf

- Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization

  http://www.aclweb.org/anthology/P04-3020

- The Evaluation of Sentence Similarity Measures.

  http://www.cis.drexel.edu/faculty/thu/research-papers/dawak-547.pdf

- TextRank: Bringing Order into Texts.

  https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf

- LexRank: Graph-based Lexical Centrality as Salience 5 in Text Summarization

  https://www.jair.org/media/1523/live-1523-2354-jair.pdf