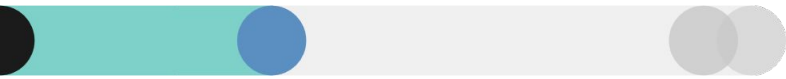
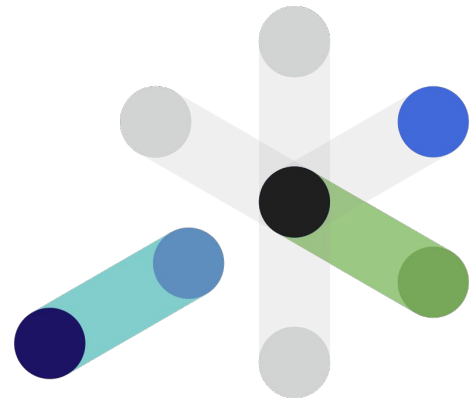


Building a Text Summarizer



srijan:



By

Mayank Kumar
Data Scientist at Srijan Technologies

Date: 29-02-2020 | Venue: Srijan Technologies, New Delhi

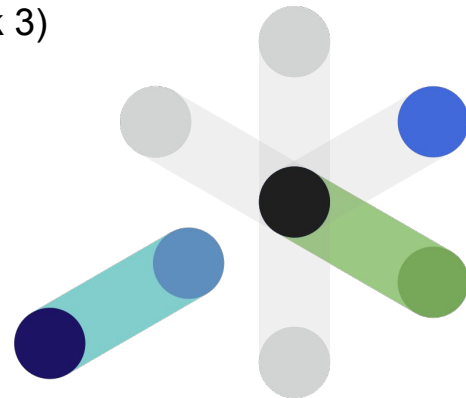
- **Mayank Kumar**

- Data Scientist at Srijan Technologies
- Kaggle Competitions Expert
- Won few competitions in the past:
 - Analytics India Magazine - Identify the author Challenge by Machine Hack (rank 2)
 - ZS Young data scientist challenge 2018 by Hackerearth (rank 3)
 - World data science challenge by Bitgrit (rank 4)
- Also a competitive programmer:
 - Won 2 silver and 3 bronze medals at Hackerrank
- B.Tech Gold Medalist in Academics

- Hobbies : Competitive ML, Algorithms design, Anime series lover

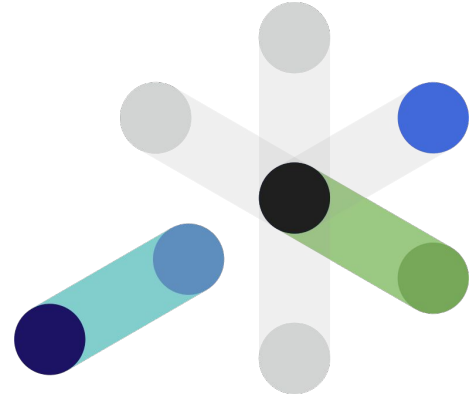
- Can be reached at

- LinkedIn (<https://www.linkedin.com/in/mk9440/>)
- Kaggle (<https://www.kaggle.com/mk9440>)
- Other profiles can be accessed at **mk9440** as well



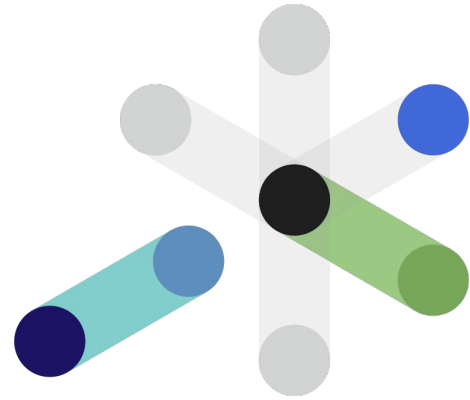
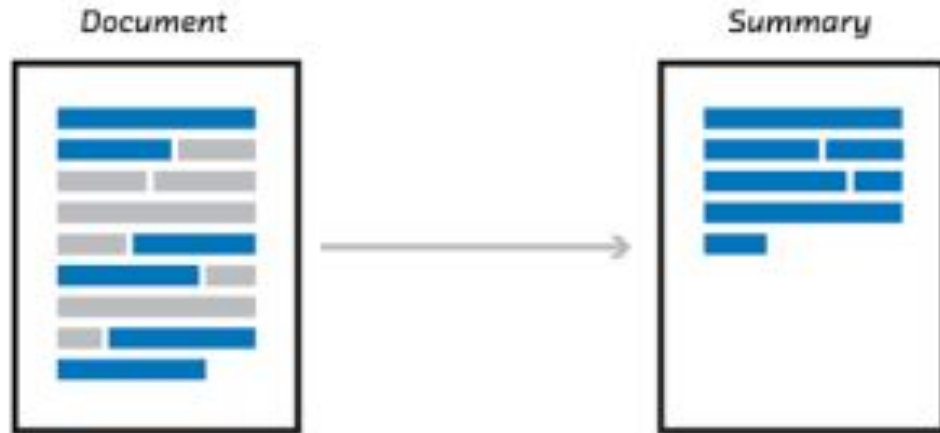
What is Text Summarization?

srijan:



What is Text Summarization?

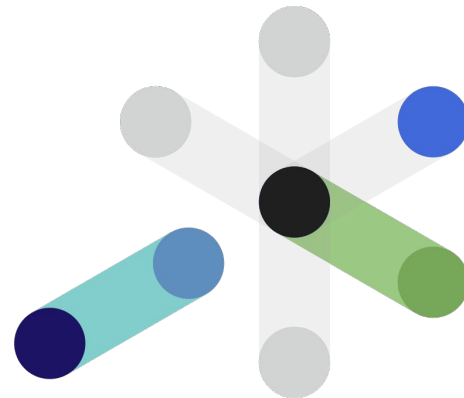
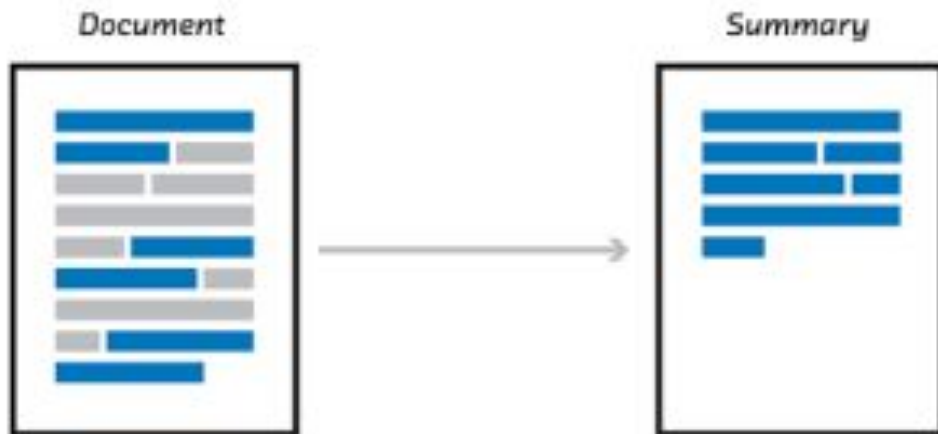
srijan:



What is Text Summarization?

srijan:

- Text summarization is the process of shortening a set of data computationally, to create a subset (a summary) that represents the most important or relevant information within the original content.^{src: wikipedia}
- Text summarization is the technique for generating a concise and precise summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning.^{src: floydhub}

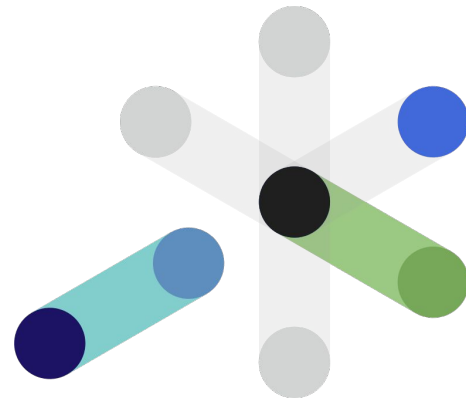


Various approaches for text summarization?

srijan:

- **Extractive Summarization**


- **Abstractive Summarization**



Various approaches for text summarization?

srijan:





- **Extractive Summarization**



Source Text:  Peter and Elizabeth took a taxi to attend the night party in the city.


While in the party, Elizabeth collapsed and was rushed to the hospital.

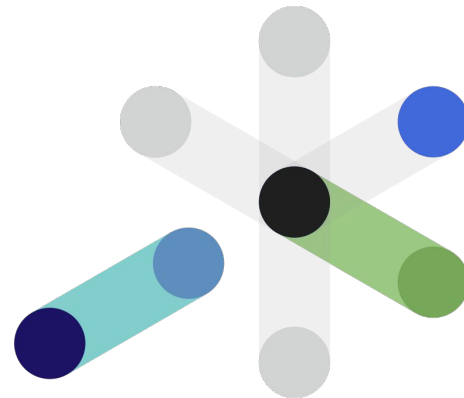
Summary: Peter

- **Abstractive Summarization**

Source Text:  and  took a taxi to  the night  in the city.

While in the party,  collapsed and was rushed to the .

Summary: Elizabeth was hospitalized after attending a party with Peter. 




Various approaches for text summarization?

srijan:

- **Extractive Summarization**

- Extractive summarization means identifying important sections (sentences or paragraphs or even words) of the text and selecting (copy paste) them producing a subset of the text from the original text.





Source Text:  Peter and Elizabeth took a taxi to attend the night party in the city.



While in the party, Elizabeth collapsed and was rushed to the hospital.


Summary: Peter

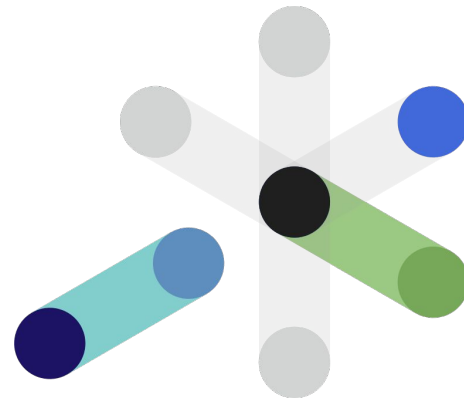
- **Abstractive Summarization**

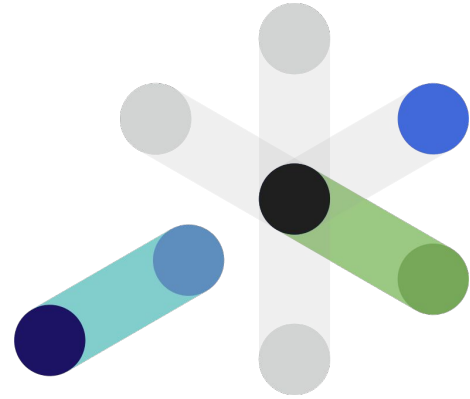
- Abstractive summarization is the technique of generating a summary of a text from its main ideas, not by copying verbatim most salient sentences from text.

Source Text:  and  took a taxi to  the night  in the city.

While in the party,  collapsed and was rushed to the .

Summary: Elizabeth was hospitalized after attending a party with Peter. 

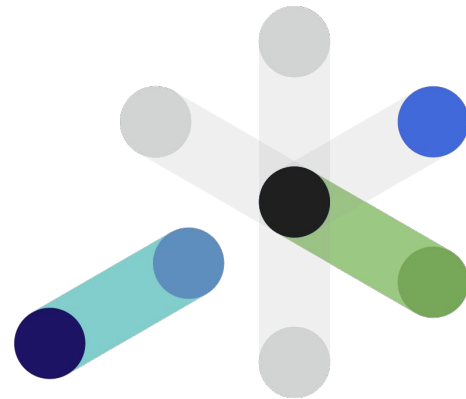




Approaching Extractive Summarization?

srijan:

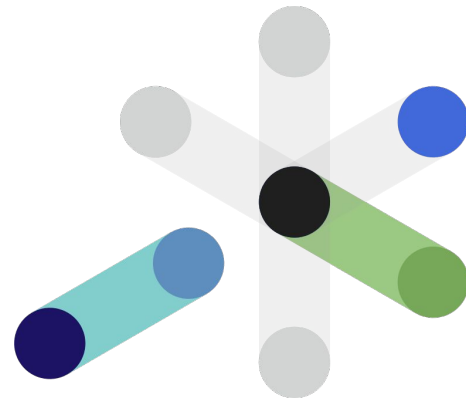
- One approach could be to create a semantic representation of sentences.



Approaching Extractive Summarization?

srijan:

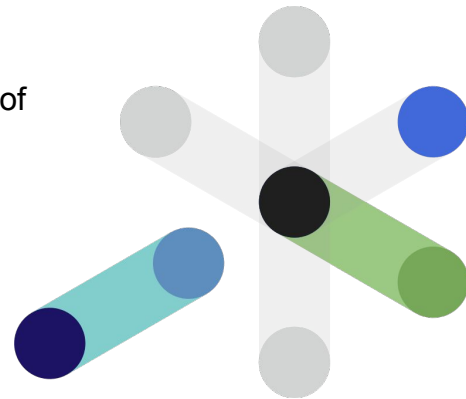
- One approach could be to create a semantic representation of sentences.
- Following can be used to create a semantic representation for texts:
 - Count-based techniques like CountVectorizer, Tf-Idf Vectorizer
 - Pretrained word embeddings based techniques like Word2Vec, Glove, fastText
 - Pretrained SOTA transformers like BERT (or its variants) to better capture context as well.
 - Train your own (quite cumbersome :()



Approaching Extractive Summarization?

srijan:

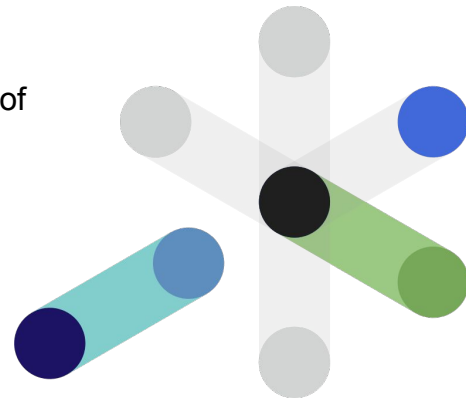
- One approach could be to create a semantic representation of sentences.
- Following can be used to create a semantic representation for texts:
 - Count-based techniques like CountVectorizer, Tf-Idf Vectorizer
 - Pretrained word embeddings based techniques like Word2Vec, Glove, fastText
 - Pretrained SOTA transformers like BERT (or its variants) to better capture context as well.
 - Train your own (quite cumbersome :()
- Now comparison can be done between each sentences as they are no more a sequence of characters and words but a numerical vector now.



Approaching Extractive Summarization?

srijan:

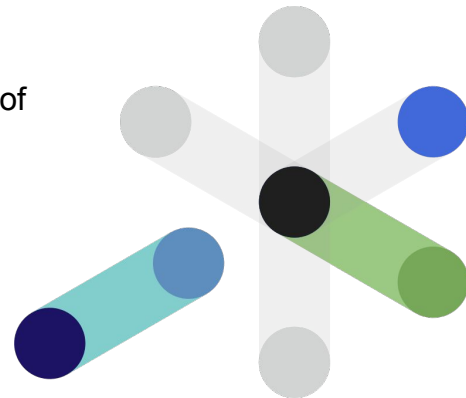
- One approach could be to create a semantic representation of sentences.
- Following can be used to create a semantic representation for texts:
 - Count-based techniques like CountVectorizer, Tf-Idf Vectorizer
 - Pretrained word embeddings based techniques like Word2Vec, Glove, fastText
 - Pretrained SOTA transformers like BERT (or its variants) to better capture context as well.
 - Train your own (quite cumbersome :()
- Now comparison can be done between each sentences as they are no more a sequence of characters and words but a numerical vector now.
- Use the comparisons to score each sentences and pick the top scored ones as your summary.



Approaching Extractive Summarization?

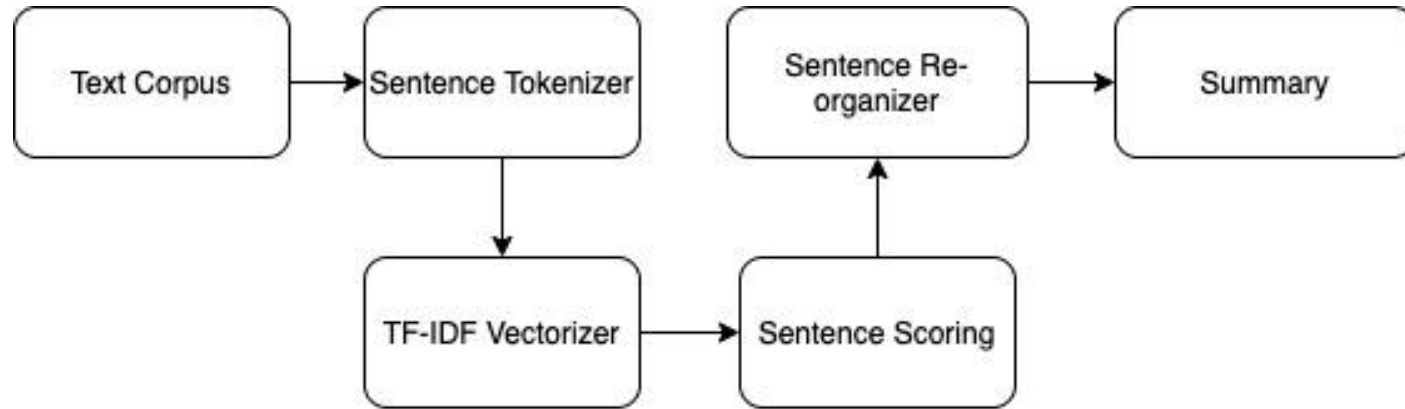
srijan:

- One approach could be to create a semantic representation of sentences.
- Following can be used to create a semantic representation for texts:
 - Count-based techniques like CountVectorizer, Tf-Idf Vectorizer
 - Pretrained word embeddings based techniques like Word2Vec, Glove, fastText
 - Pretrained SOTA transformers like BERT (or its variants) to better capture context as well.
 - Train your own (quite cumbersome :()
- Now comparison can be done between each sentences as they are no more a sequence of characters and words but a numerical vector now.
- Use the comparisons to score each sentences and pick the top scored ones as your summary.
- Scoring technique needs to be intelligent enough to properly evaluate what are the information content of a sentence and thus score it accordingly.

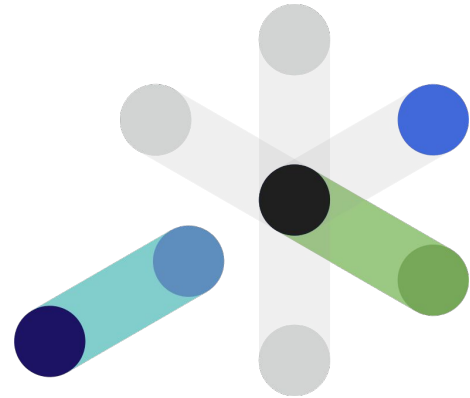


Approaching Extractive Summarization?

srijan:



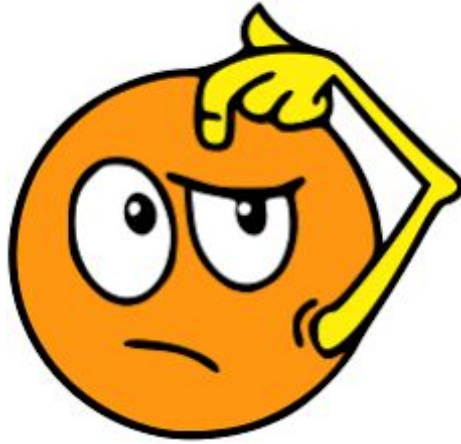
A high level solution for frequency based extractive summarizer



Fire up your Notebooks



srijan:



srijan:

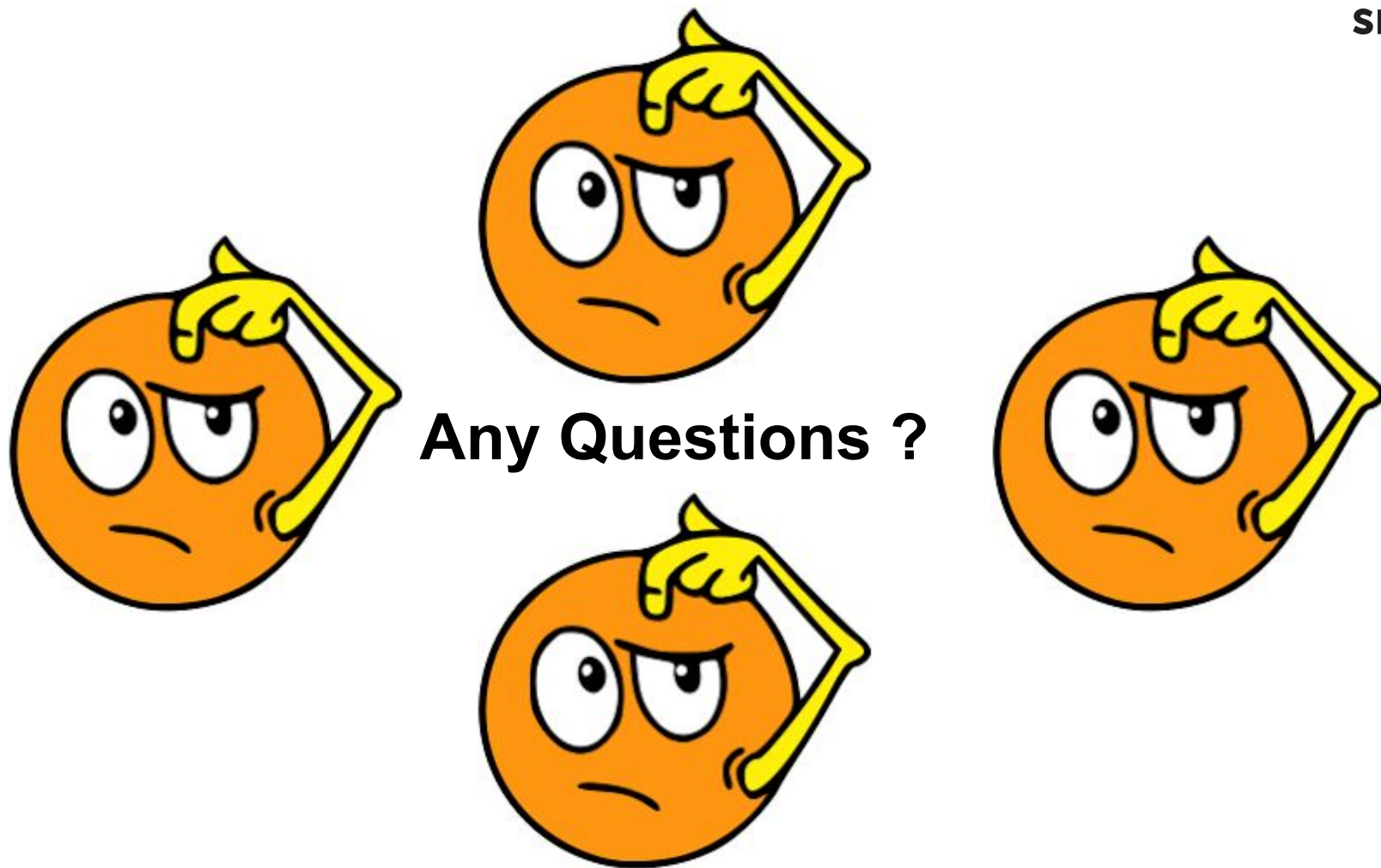


srijan:



srijan:





- <https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/>
- <https://medium.com/@ondenyi.eric/extractive-text-summarization-techniques-with-sumy-3d3b127a0a32>
- https://en.wikipedia.org/wiki/Automatic_summarization

Thank You



srijan:

Head Offices

2430 Highway 34
Building B, Suite 22
Manasquan, NJ 08736, USA

8D Vandana Building, Tolstoy Marg,
New Delhi 110001, INDIA

[email:](mailto:business@srijan.net) business@srijan.net

[web:](http://srijan.net) srijan.net

[twitter:](https://twitter.com/srijan) @srijan