Mandate 1

Problem Statement:

(Abstractive Summarization) Given a set of tweets pertaining to a trending topic, create an abstractive prose summary of the tweets. Do not just string the tweets together to form the summary. The summary will need to paraphrase and/or say more than what is directly said in the tweets. Propose a rubric to evaluate the accuracy of your summarization.

Abstractive Summarization :

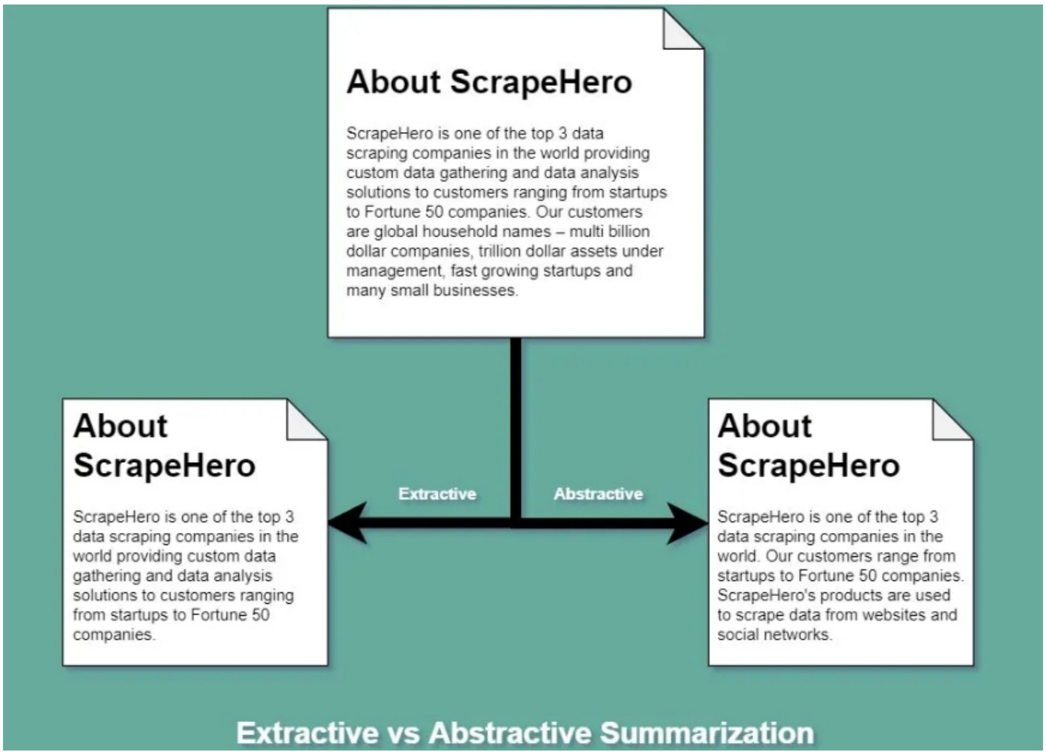
This aims to create short summaries of longer documents while retaining the core content and preserving the overall meaning of the text.

For summarization there is two approaches:

1. Extractive Approach
2. Abstractive Approach

Extractive Approach : In this model output can be considered a subset of the input text which conveys the main idea of the input article or we can consider this as **highlighting important points** of a reference paper that you are trying to understand.

Abstractive Approach : The main idea is to grasp the underlying idea of the text and reproduce the summary that would cover all the important points, with the **same semantic**. The machine learning model would output the main idea of the input text using **similar words** but **not exact sentences from the input.**
Example:



About ScrapeHero

ScrapeHero is one of the top 3 data scraping companies in the world providing custom data gathering and data analysis solutions to customers ranging from startups to Fortune 50 companies. Our customers are global household names – multi billion dollar companies, trillion dollar assets under management, fast growing startups and many small businesses.

About ScrapeHero

ScrapeHero is one of the top 3 data scraping companies in the world providing custom data gathering and data analysis solutions to customers ranging from startups to Fortune 50 companies.

About ScrapeHero

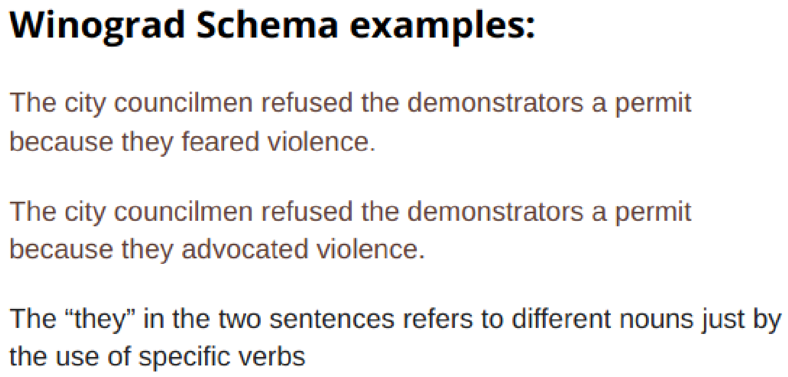
ScrapeHero is one of the top 3 data scraping companies in the world. Our customers range from startups to Fortune 50 companies. ScrapeHero's products are used to scrape data from websites and social networks.

Extractive vs Abstractive Summarization

Challanges in Abstractive Sumamrization :

- **Ambiguity :** Ambiguity means **uncertainty of meaning**. Most human languages are inherently ambiguous. Consider the following sentence: “I made him duck.”
  This sentence has multiple meanings. The first one is, I cooked a duck for him. The second meaning, is I made him bend down to avoid an object.

- **Common Sense :** A key aspect of any human language is “common knowledge.” It is the **set of all facts that most humans are aware of**. In any conversation, it is assumed that these facts are known, hence they’re not explicitly mentioned, but they do have a bearing on the meaning of the sentence. A common class of coreference resolution problems that require common sense understanding, are the so-called “Winograd schemas”.



Winograd Schema examples:

The city councilmen refused the demonstrators a permit because they feared violence.

The city councilmen refused the demonstrators a permit because they advocated violence.

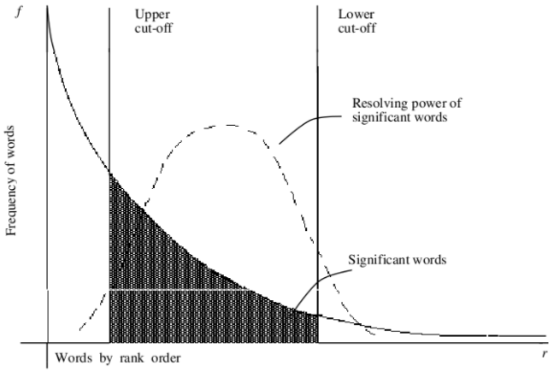
The “they” in the two sentences refers to different nouns just by the use of specific verbs

- **Creativity :** Humans can build rich, **abstract models of both real and imaginary words**, and communicate semantics with rich enough precision to induce a similar conceptual model in the recipient. Making machines understand creativity is a hard problem.

All these challanges made hard for machines to understand the underlying idea.

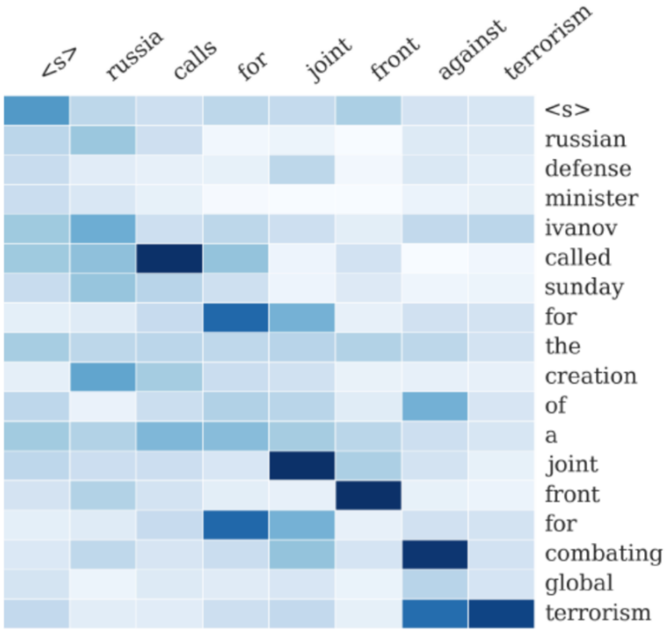
Overview :

Here in this problem we basically **wants to summarizes the tweets** and make machine try to put summary in its own word. So we need i**mportant/significant words** which can be approximated using **Part-of-speech (POS)** tagging which is a popular NLP process which refers to categorizing words in a text (corpus) in **correspondence with a particular part of speech,** depending on the definition of the word and its context or by Zipf distribution. As shown in below fig. that significant words are generally in mid-range of frequency and we need to capture them along with facts/trending topics which starts with symbol “#”.

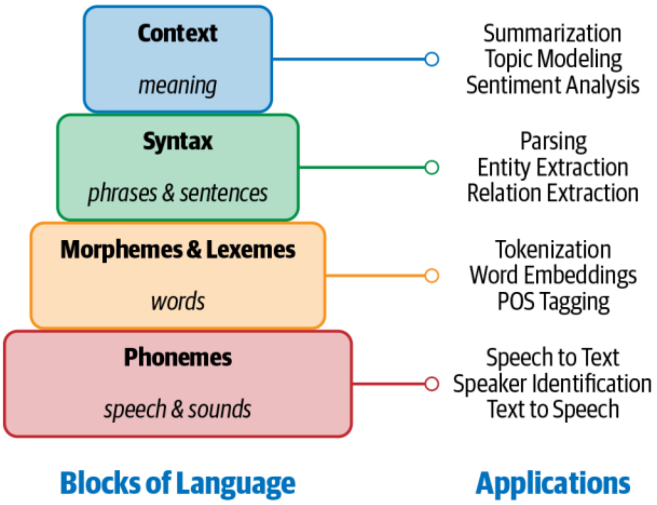


Word distribution in documents found to be very skewed -- approximated by a Zipf distribution

In the same way when we want machine to generate text from available text then machine should have **coupus of similar meaning** which can be attained from the **distributional hypothesis** which states that words that are used and occur in the same contexts tend to purport similar meanings. Co-occurance as the statistical basis for learning latent semantics. After this we can feed these sentences into **vector/similarity matrix** and with the sentence ranking we can generate summary.



Above figure is example of **abstractive summarization of a tweet**. In this we basically find the words which is important also change some words with their kind of synonym like **"combating" for "against".**
Also there are many advance technique which doesn't require high amount of pre-processing(lemmatization, stemming etc) like seq2seq or transfromer which requires attention and saves the underlying meaning in **cell states.**
Below Figure shows the building block for getting the context(summary).



Problem Formulation :

In all mandates we try to overcome the challanges in **Abstractive Summarization(set of twitter hashtags)** which is arises due to many facts given in challanges setion and also due to **short hand texts, use of external links, spelling probelms** or we can say this is generally how we write in social media. After this **fine tuning** will be performed on model. By creating the summarization of trending hashtags/ tweets we can get info about our audience in focused manner.

Basic flow of problem:

Corpus(Source):

Large collection of summaries and texts/set of **tweets** on trending topic.

Text Preprocessing :

In this we will pre-process the data and can use different library **like nltk for stop word removal**, **lemmatization, stemming**(pre-language model) or we can also make our **own dictionary** for short hand words and can define the correct meaning of **short hand/hashtags/ wrong spellings**.

Semantic Analysis(extractions):

Here we can use vectorization for our words and for that we can use **tokenization, word embeding, entity/feature extraction,rankings** etc.

Above all are steps of fine tuning of our data.

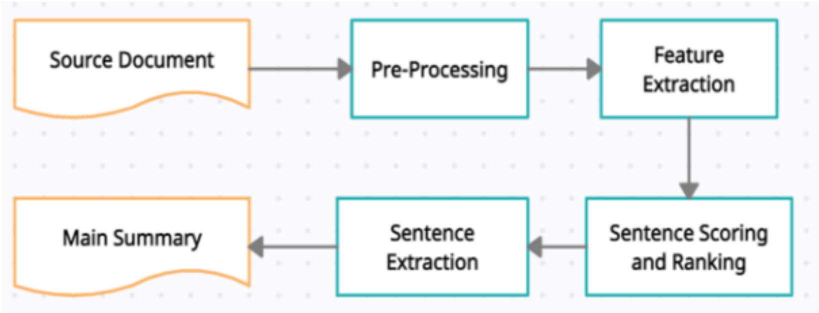
Model:

Here in this we can use our main **model for training** the machine which takes previous stage data as input in vector space and creates the summary. Generally **models are pre-trained** and may not work specifically according to our task then **we need to train them on top of their pre-training** so that they can give accurat result.

Evaluation method:

Right now we have two methods available for checking our accuracy :

- **Bleu** measures precision: how much the words (and/or n-grams) in the machine generated summaries appeared in the human reference summaries.
- **Rouge** measures recall: how much the words (and/or n-grams) in the human reference summaries appeared in the machine generated summaries.



These are basically steps needed for reaching the final stage but in the process we try to furnish the model use advance approaches/new methods to make our model robust.