* 1. Text summarization
     1. Many methods used to summarize the texts by assigning weightage to each of the sentence
        1. Cue method: sentence with the words (which are in cue dictionary) will have more weightage
        2. Title method: sentence with more words same as title name will have more weightage
        3. Location method: sentences appearing in the beginning of document and beginning of a paragraph will have more weightage
     2. In general, there are two different approaches for automatic summarization: extraction and abstraction.
        1. Extraction will extract meaning full information in terms of existing sentences only
        2. Abstraction will summarize the input data in new or critical terms
        3. Even though human approach for text summarization is more of abstractive, still most of the best performing approaches belongs to extraction concept.
     3. Extractive summarization:
        1. Intermediate representation:
           1. Topic representation: explains the topics discussed in document
           2. Indicator representation: explains important features of a sentence
        2. Sentence score: a sentence score will be generated which is based on how well a sentence is explaining a topic
        3. Summary sentences selection: k importance sentences will be selected based on sentence score or some optimization methods (which will tell us that what combination of sentences explain a particular topic in best way)
     4. Topic representation approaches:
        1. Topic words:
           1. In an advanced approach, log-likelihood test is used to identify explanatory words (topic signature)
           2. Here, sentence score can be measured in two ways:

Number of test signatures in a sentence: (can be biased for large sentences)

Proportion of test signature terms

* + - 1. Frequency driven approaches:
         1. Using word probability (average tf-score)
         2. Using tf-idf score:

It deals with drawback of word probability approach

If we don’t remore stop words in tf-idf approach then also, it will take care of that problem

In word probability it creates issue, because we are not sure which words should be included in stop words list.

After finding tf-idf, important words will be considered and cluster of sentences will be formed and then some metrics are defined to put each sentence’s score

* + - 1. Latent semantic analysis (LSA):
         1. It’s an unsupervised method for extracting a representation of text semantics
         2. A word-document matrix with tf-idf score will be estimated.
         3. Then SVD is applied then abiove matrix will be transformed into 3 matrixes.

A = U Σ(V )T, where U is term-topic matrix, Σ is topic-topic matrix and (V )T is topic-sentence matrix.

Or D= Σ(V )T represents topic-sentence matrix

Now, we can explain using some summarization approaches that

Scores for topics a sentences can explain

Scores for sentences explaining a topic

* + - 1. Bayesian topic models:
         1. Generally used for multi-document summarization
         2. Most of the existing models consider that sentences are independent of each other (drawback)
         3. Also, most of the existing models don’t have probabilistic interpretations
         4. Bayesian topic models cover all that information which were not there in non-probabilistic approaches.
         5. Probabilistic topic models have gained dramatic attention in recent years in various domains
         6. Latent Dirichlet allocation (LDA) model is the state of the art unsupervised technique for extracting thematic information (topics) of a collection of documents.
    1. Knowledge bases and automatic summarization:
       1. A step towards building more accurate summarization systems is to combine summarization techniques with knowledge bases (semantic-based or ontology-based summarizers).
    2. The impact of context in summarization:
       1. Web summarization
       2. Scientific articles summarization
       3. Email summarization:
          1. Newman et al. [47] describe a system to summarize a full mailbox rather than a single thread by clustering messages into topical groups and then extracting summaries for each cluster.
    3. Indicator representation approaches:
       1. Graph methods for summarization:
          1. Graph-based methods can be used for single as well as multi-document summarization.
          2. Since they do not need language specific linguistic processing other than sentence and word boundary detection, they can also be applied to various languages
          3. Nonetheless, using TFIDF weighting scheme for similarity measure has limitations, because it only preserves frequency of words and does not take the syntactic and semantic information into account.
          4. Thus, similarity measures based on syntactic and semantic information enhances the performance of the summarization system
       2. Machine learning for summarization:
          1. Naive Bayes approach:

The probability a sentence to belong to the summary is the score of the sentence.

* + - * 1. Other approaches like decision trees, support vector machines, Hidden Markov models and Conditional Random Fields are among the most common machine learning techniques used for summarization
        2. Methods explicitly assuming the dependency between sentences such as Hidden Markov model [18] and Conditional Random Fields [65] often outperform other techniques.
        3. One of the primary issues in utilizing supervised learning methods for summarization is that they need a set of training documents (labeled data) to train the classifier, which may not be always easily available. Researchers have proposed some alternatives to cope with this issue:

*Annotated corpora creation:* Creating annotated corpus for summarization greatly benefits the researchers, because more public benchmarks will be available which makes it easier to compare different summarization approaches together.

*Semi-supervised approaches:*

In semi-supervised learning we utilize the unlabeled data in training.

They co-trained two classifiers iteratively to exploit unlabeled data. In each iteration, the unlabeled training examples (sentences) with top scores are included in the labeled training set, and the two classifiers are trained on the new training data

* + - * 1. Machine learning methods have been shown to be very effective and successful in single and multi-document summarization, specifically in class specific summarization where classifiers are trained to locate particular type of information such as scientific paper summarization and biographical summaries
    1. EVALUATION
       1. Evaluation of a summary is a difficult task because there is no ideal summary for a document or a collection of documents and the definition of a good summary is an open question to large extent.
       2. It has been found that human summarizers have low agreement for evaluating and producing summaries. Additionally, prevalent use of various metrics and the lack of a standard evaluation metric has also caused summary evaluation to be difficult and challenging.
       3. Evaluation of automatically produce summaries:
          1. In order to be able to do automatic summary evaluation, we need to conquer three major difficulties:

It is fundamental to decide and specify the most important parts of the original text to preserve.

Evaluators have to automatically identify these pieces of important information in the candidate summary, since this information can be represented using disparate expressions. T

The readability of the summary in terms of grammaticality and coherence has to be evaluated.

* + - 1. Human Evaluation:
         1. The factors that human experts must consider when giving scores to each candidate summary are grammaticality, non- redundancy, integration of most important pieces of information, structure and coherence.
      2. Automatic evaluation methods:
         1. ROUGE is the most widely used metric for automatic evaluation.

There are several variations of ROUGE, and here we just mention the most broadly used ones:

ROUGE-n:

This metric is recall-based measure and based on comparison of n-grams. a series of n-grams (mostly two and three and rarely four) is elicited from the reference summaries and the candidate summary (automatically generated summary). Let p be "the number of common n-grams between candidate and reference summary", and q be "the number of n-grams extracted from the reference summary only".

The score is computed as: ROUGE-n = p/q

ROUGE-L

ROUGE-SU