

31-12-2020

PAGE NO. \_\_\_\_\_

## Chapter - 6

### Text Summarization and Text Generation.

#### Objectives :

- Describe automated text summarization and its benefits.
- Describe Text Rank Algorithm.
- Implement ~~Text~~ text summarization using gensim.
- Implement text summarization using word frequency.
- Generate text using markov chains.

## Text summarization :

Reducing the content of the text while retaining the central ideas of the content.

## Benefits of Automated text summarization :

- Sampling
- Searching
- Indexing
- Reading time
- Answering questions

⇒ High Level View of Text Summarization

⇒ Purpose

- Generic Summarization
- Domain specific Summarization
- Query based Summarization

⇒ Input :-

- Single document
- Multiple documents.

⇒ Output :-

- Extractive
- Abstractive

⇒ Different types of output

i) Extractive text summarization :

- From a text, we select some sentences which are important in describing its true meaning.

• Steps :

- i) Rank the sentences according to
  - How much a sentence contributes to the ~~stage~~ overall meaning of text
- ii) Select the highest ranked sentences and form a summary

Sentences can be ranked by  $\rightarrow$

- Word frequency
- Sentence similarity
- Clustering centrality

ii) Abstractive text summarization :

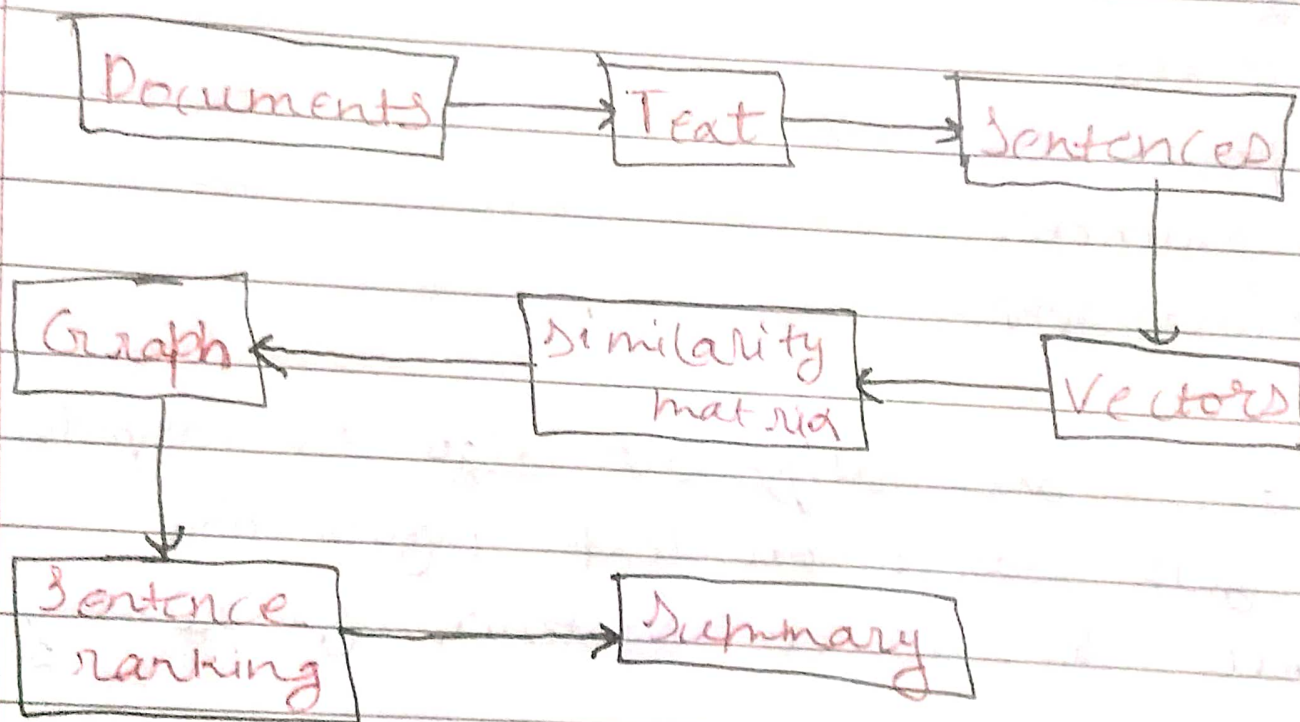
- ~~Creates~~ Creates an abstract summary, i.e., it does not have the exact words as used in text, but has the same meaning.

- Uses deep learning

eg: Sequence to sequence, encoder to decoder.



⇒ Text Rank



# Text Summarization using TextRank

Algo -  
Youtube (RSree tech)

## ⇒ Text Summarization :

- Summary of text represents the the most relevant information within the content.
- Extractive / Abstractive / Hybrid

## ⇒ Extractive Text Summarization :

- Select important sentences from a document / documents to form a summary
- Different text information units :
  - words
  - phrases
  - sentences
  - Paragraphs.
- There are different algo to compute weights based on text information units to select sentences for summary.
- Grammatically correct, but sequences of sentences may not be suitable for smooth reading.

## Abstractive Text Summarization:-

- Builds an intermediate semantic representation of the doc.
- Summary is synthesized from the intermediate semantic representation.
- Summary may not be having the original content and may be using paraphrasing technique.
- Challenging problem to generate grammatically and semantically correct summaries.



## Text Rank Algorithm

- For the task of automated Summarization, Text Rank models any document as a graph, using sentences as nodes.
- A function to compute the similarity of sentences is needed to build edges in between. The function is used to weight the graph edges.
- The higher the similarity, the more important the edge b/w them will be in the graph.

- TextRank determines the relation of similarity b/w two sentences based on the content that both share.
- The overlap is calculated simply as the number of common lexical tokens, divided by the length of each to avoid promoting long sentences.

7-1-2021

⇒ Summarizing text using word frequency.

1. Ignore stop words
2. Determine top words - most frequently occurring words
3. Select top words = A subset of top words used for scoring
4. Select top sentences: Sentences are scored on the basis of top words they contain.

## Math. Markov Chains →

- Mathematical system for transitioning between states based on probabilities.
- For a given present state, ~~we predict the~~ the possible future states are predetermined.
- Markov chain operates in time steps. After each timestep the system selects a new state.
- This new state is chosen based on the current state and probabilities of future state.

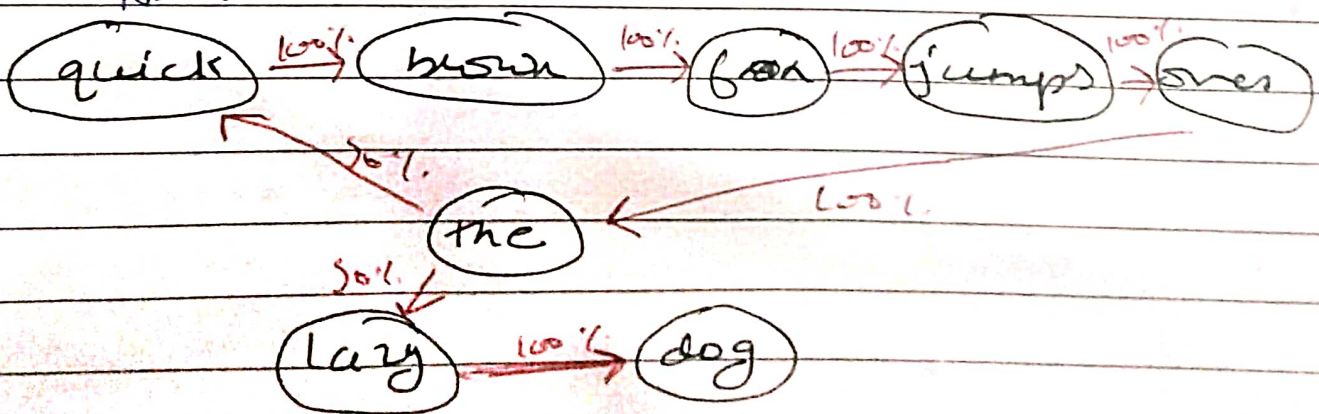


In text Generation,

Let's suppose the only sentence in  
our vocab is

"The quick brown fox jumps over the  
lazy dog"

Then the markov model generated  
is &





- each word is a state and arrow defines the next state with a prob.
- In this case the newly generated word ~~can~~ is ~~a~~ only dependent on the previous word.  
(This may not give meaningful sentences)