Text Summarizer and Question Answering System using NLTK by Anmol Parmar (102003615)

((Text Summarizer))

Text summarization is the process of shortening long pieces of text while preserving key information content and overall meaning, to create a subset (a summary) that represents the most important or relevant information within the Text.

Here we are using NLTK for making the text summarizer. Now the question arises here what is NLTK?

NLTK: -

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP).

It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes graphical demonstrations and sample data sets as well as accompanied by a cook book and a book which explains the principles behind the underlying language processing tasks that NLTK supports.

Features of Text Summarizer: -

1.) tf-idf matrix

2.) cue Phrases

3.) Numerical Data Comparing

4.) Sentence Length

5.) Sentence Position

6.) Upper Case

7.) Number of Proper Noun

8.) Heading Match

9.) Named Entity Recognition

1.) tf-idf matrix: -

Term frequency works by looking at the frequency of a *particular term* you are concerned with relative to the document. Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus.

We are using tf-idf matrix in our text summarizer to summarize the key intuition motivating TF-IDF is the importance of a term is inversely related to its frequency across documents.TF gives us information on how often a term appears in a document and IDF gives us information about the relative rarity of a term in the collection of documents.

Then, we calculate the score out of it and then add it to the total_score.

2.) cue Phrases: -

Cue phrases are linguistic expressions such as 'now' and 'well' that may explicitly mark the structure of a discourse. For example, while the cue phrase 'incidentally' may be used SENTENTIALLY as an adverbial, the DISCOURSE use initiates a digression.

Cue Phrases we are using: - anyway, by the way, furthermore, first, second, then, now, thus, moreover, therefore, hence, lastly, finally.

Then, we calculate the score out of it and then add it to total_score.

3.) Numerical Data: -

Here, we calculate the numerical data scoring based on how much it contains the numerical value and then add the score to the total_score.

4.) Sentence Length: -

Here, we find the score based on what is the length of the sentence and then add It to the total_score.

5.) Sentence Position: -

Here, we calculate the score based on the where the sentence is positioned and then add it to the total_score.

6.) upper case: -

In this we compare the text for the upper cases and then calculate the score out of it and then add it to the total_score.

7.) Number of Proper Noun: -

Proper Noun is a noun that designates a particular being or thing, does not take a limiting modifier, and is usually capitalized in English. So, we calculate the number of proper Noun and add it to the total score.

8.) Heading Match: -

Here, we compare it with the first line of the text file as it is considered as the heading and then calculate the score out of it and add it to the total_score.

9.) Named Entity Recognition: -

Named Entity Recognition is a natural language processing (NLP) technique that automatically identifies named entities in a text and classifies them into predefined categories. Entities can be names of people, organizations, locations, times, quantities, monetary values, percentages, and more. Here it highlights the text which is it like is it the person, place or anything else.

At last, we add all the scores and then out of that score we calculate the summary.

((Question Answering System))

Question answering is a critical NLP problem and a long-standing artificial intelligence milestone. QA systems allow a user to express a question in natural language and get an immediate and brief response. QA systems are now found in search engines and phone conversational interfaces, and they're fairly good at answering simple snippets of information. On more hard questions, however, these normally only go as far as returning a list of snippets that we, the users, must then browse through to find the answer to our question.

Step-1: - Removing stop words and searching it in the Summary

Firstly we remove the stopwords both from the question and summary and then search the question in the summary given by text summarizer.

Step-2: - Named Entity Recognition: -

Named Entity Recognition is a natural language processing (NLP) technique that automatically identifies named entities in a text and classifies them into predefined categories. Entities can be names of people, organizations, locations, times, quantities, monetary values, percentages, and more. Here it highlights the text which is it like is it the person, place or anything else.

---THANKS---