Text Summarization in Python

AGENDA

1. What is text Summarization?
2. Why text summarization should be used?
3. Use cases of text summarization
4. Methods of summarizing a text
5. Terminologies.
6. Let us implement our own text summarizer in Python
    a. Methodology.
    b. Pseudocode.
7. Task for you..

What is text summarization?

Text summarization is what its name suggests.

It is a technique in computer science to take a piece of text and return the desired length of text as a summary of the input text.

Why

Why should we bother about text summarization?

We are constantly getting bombarded with the text information like never before. News articles, social media, books etc. How are we going to consume all of this in our limited 24 hours day?

Use Cases

1. TextRank like PageRank (by Google)
2. There is a news application called Inshorts which summarizes the news content like no one does. They use AI powered methods to accomplish such a feat.
3. Automatic Book summaries.
4. Automatic Script Summaries
5. And so on….

Methods of Summarizing text

Broadly, the methods of summarizing the text can be classified into following two types:

1. Traditional Statistical Approach (Word Freq - Sentence weight based).
2. Modern more powerful, Seq2Seq neural networks models.

Some Terminologies

1. Tokenization
    a. Word based
    b. Sentence based
2. Web scraping
3. Text cleaning using Regular Expressions.
4. Stopwords.
5. And so on...

Let us implement our own text summarizer in python

Methodology (Algorithm)

1. Take a piece of large text = t
2. Clean that text and get a new text = t′
3. Get individual words called words and individual called sentences from t′
4. Make a frequency dictionary using the words list.
5. Normalize the dictionary to get the weights in the range 0-1
6. Then, use the dictionary to give sentences_score to each of the sentences.
7. Finally, sort the sentences with the highest scores as the one which will appear first.

Pseudocode

```
# Step. 1 Import necesary modules
# Like requests, BeautifulSoup, Pandas, nltk

# Step. 2 Get access to the text content of wikipedia page
url = ''
soup = BeautifulSoup(requests.get(url).text)

# Step. 3 use nltk to tokenize using word and sentences
words = []
sentences = []

# Step. 4 Loop through words and make a frequency dictionary
freq_dict = {} # Key will be word and then their frequency

# Step. 5 Loop through freq_dict and divide each value by max
# This will bring the values in the range 0 - 1 and call it weight

weight_dict= {}

# Step.6 Make a list sentences score and compute score of each
# sentence by adding the weight of each word in the weight_dict

sentences_score = []

#Finally, sort the sentences based on sentences_score in descending order
```

What can you do?

I have just touched the tip of the iceberg. The field of text summarization is vast. So, if you are interested, then you have plenty of things to explore. One of things that I would recommend you personally is to explore Seq2Seq models for text summarization as they have been proven to produce wonderful results.

The other things that you can do is to try to improve the text summarizer that I have implemented in this hands on tutorial. See, what is it that can be improve about this summarizer.

Thank you