Wikipedia Article Summarizer *(Python project based on NLP techniques)*

Text Summarization with NLTK

EmreYbs

(NLTK, or Natural Language Toolkit, is a Python package that you can use for NLP.)

Since I like F-Secure and wishing to attend their trainings, I search for them and with this simple Wikipedia Article summarizer, I also practise NLP and Python, meanwhile learning more about F-Secure, its history, culture, etc.

In [ ]:

```
# Install these as requirements if you need. You may also try "pip3 install b
# $ pip install beautifulsoup4
# $ pip install lxml
```

Scrapping Wikipedia Article

In [ ]:

```
# https://github.com/emreYbs, 3.10.2021

import bs4 as bs
import urllib.request
import re

# Normally, in Jupyter Notebooks, you may prefer to give a fixed URL, change
# and not ask for user input.But I wanted to see which articles,
# I can get a better summary and when the NLTK does "so so":)
userLink = input("Which Wikipedia article would you want me to summarize: ")
# Provide the Wikipedia URL like this: https://
raw_data = urllib.request.urlopen(userLink)
document = raw_data.read()

parsed_document = bs.BeautifulSoup(document,'lxml')

article_paras = parsed_document.find_all('p')

scrapped_data = ""

for para in article_paras:
    scrapped_data += para.text
```

In [ ]:

```
print(scrapped_data[:1500]) #You may increase or reduce the first 1000
# or 1500 characters of the scraped text, etc
```

```
Information Security, sometimes shortened to InfoSec, is the practice of prot
ecting information by mitigating information risks. It is part of information
risk management.[1] It typically involves preventing or reducing the probabil
ity of unauthorized/inappropriate access to data, or the unlawful use, disclo
sure, disruption, deletion, corruption, modification, inspection, recording,
or devaluation of information.[2] It also involves actions intended to reduce
the adverse impacts of such incidents. Protected information may take any for
m, e.g. electronic or physical, tangible (e.g. paperwork) or intangible (e.g.
knowledge).[3][4] Information security's primary focus is the balanced protec
tion of the confidentiality, integrity, and availability of data (also known
as the CIA triad) while maintaining a focus on efficient policy implementatio
n, all without hampering organization productivity.[5] This is largely achiev
ed through a structured risk management process that involves:
To standardize this discipline, academics and professionals collaborate to of
fer guidance, policies, and industry standards on password, antivirus softwar
e, firewall, encryption software, legal liability, security awareness and tra
ining, and so forth.[7] This standardization may be further driven by a wide
variety of laws and regulations that affect how data is accessed, processed,
stored, transferred and destroyed.[8] However, the implementation of any stan
dards and guidance within an entity may have li
```

Text Cleaning

In [ ]:

```
scrapped_data = re.sub(r'\[[0-9]*\]', ' ',  scrapped_data)
scrapped_data = re.sub(r'\s+', ' ',  scrapped_data)
```

In [ ]:

```
formatted_text = re.sub('[^a-zA-Z]', ' ', scrapped_data)
formatted_text = re.sub(r'\s+', ' ', formatted_text)
```

Finding Word Frequencies

In [ ]:

```
import nltk #if you don't have it, then>> python3 -m pip install nltk
all_sentences = nltk.sent_tokenize(scrapped_data)
```

In [ ]:

```
# Stop Words are the words that you will most probably ignore, so we filter t
stopwords = nltk.corpus.stopwords.words('english')

word_freq = {}
for word in nltk.word_tokenize(formatted_text):
    if word not in stopwords:
        if word not in word_freq.keys():
            word_freq[word] = 1
        else:
            word_freq[word] += 1
```

In [ ]:

```
max_freq = max(word_freq.values())

for word in word_freq.keys():
    word_freq[word] = (word_freq[word]/max_freq)
```

Finding Sentence Scores

In [ ]:

```
sentence_scores = {}
for sentence in all_sentences:
    for token in nltk.word_tokenize(sentence.lower()):
        if token in word_freq.keys():
            if len(sentence.split(' ')) <25:
                if sentence not in sentence_scores.keys():
                    sentence_scores[sentence] = word_freq[token]
                else:
                    sentence_scores[sentence] += word_freq[token]
```

Printing Summaries

In [ ]:

```
import heapq
selected_sentences= heapq.nlargest(5, sentence_scores, key=sentence_scores.ge

text_summary = ' '.join(selected_sentences)
print(text_summary)
```

```
Information security must protect information throughout its lifespan, from t
he initial creation of the information on through to the final disposal of th
e information. During its lifetime, information may pass through many differe
nt information processing systems and through many different parts of informa
tion processing systems. Information Security, sometimes shortened to InfoSe
c, is the practice of protecting information by mitigating information risks.
In the realm of information security, availability can often be viewed as one
of the most important parts of a successful information security program. In
information security, confidentiality "is the property, that information is n
ot made available or disclosed to unauthorized individuals, entities, or proc
esses."
```