LIS452

Ting Zhan

Github: https://github.com/tingzhan19921001/ischool-lis452/blob/master/Ting_Zhan_LIS452.py

Final Project: The News Abstract Tool

The purpose of The News Abstract Tool

The purpose of this tool is to create an abstract of a news if a new link is given. In the process of programming, the test link are: http://www.chinadaily.com.cn/business/2017-05/11/content_29300055.htm
http://www.chinadaily.com.cn/business/tech/2017-05/11/content_29303225.htm

My basic idea is simple: the sentence with the most keyword is the most important sentence. I will calculate each word's importance. The calculation equation is first to figure each word's frequency and then use this frequency to divide the total words, thereby we get each word's importance value. Sentence value is calculated based on the total value addition of word's value (the stopwords are not included during the addition). After sorting out the sentences according to the number of keywords they have, take ones that rank top, then aggregate into our summary. Here we use the (len(sentences)) to divide 10, every 10 sentence contributes to one sentence more to the summary.

So my work can be divided into the following steps: given the words that appear in the article, calculate the importance of each word according to the algorithm mentioned above, frequency of word/total word. First of all, we have to count out the number of words in the article, word frequency(which is done in Week8's assignment) , we can know the number of occurrences of the largest number of occurrences and then we use it to calculate the "important coefficient" of each word. Highly important words are our keywords. Calculate the total score of each sentence according to the importance of the words in the sentence. Now every word (except stopwords) has a value that is "important" as a quantified description. What we need now is the importance of the words in a sentence. It is only necessary to superimpose the importance of each word in the sentence. Sort out the sentences in the article according to the total score of the sentence. Take out the first n sentences as a summary and we get the results. I would add more and try do something more complicated and could include more of what we study in the process.

How to use The News Abstract Tool

Since I have the library ‘nltk’ in my programming, I hope that when users use my tool, it would be better if they have installed ‘nltk’ in python first.

With everything ready, now let’s start and give it a try.

There are several test news links I would like to offer:

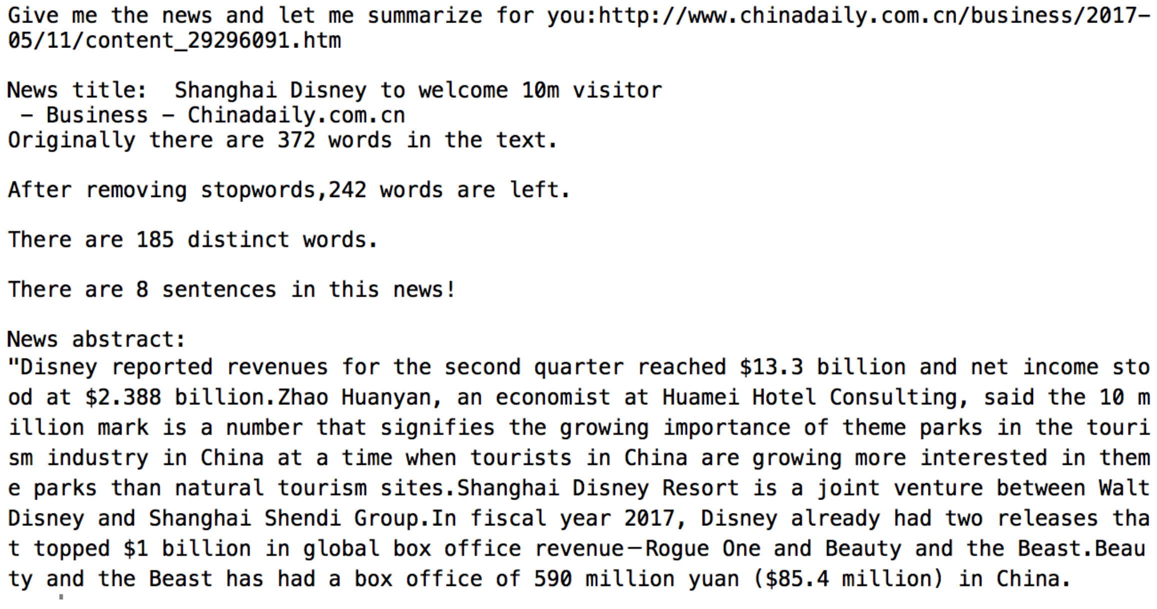
http://www.chinadaily.com.cn/business/2017-05/11/content_29300055.htm

http://www.chinadaily.com.cn/business/tech/2017-05/11/content_29303225.htm

http://www.chinadaily.com.cn/business/2017-05/11/content_29296091.htm

Pick anyone that you like.

For example, I use the third link and the output would look like this:



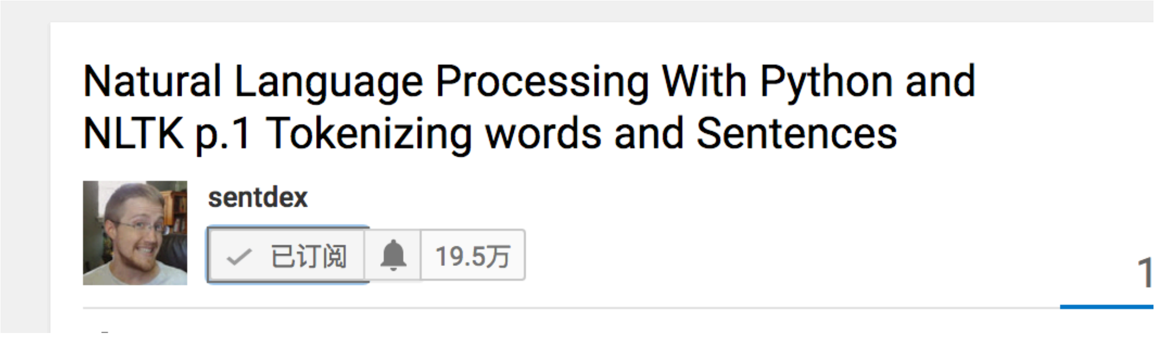
```
Give me the news and let me summarize for you:http://www.chinadaily.com.cn/business/2017-
05/11/content_29296091.htm

News title:  Shanghai Disney to welcome 10m visitor
 - Business - Chinadaily.com.cn
Originally there are 372 words in the text.

After removing stopwords,242 words are left.

There are 185 distinct words.

There are 8 sentences in this news!

News abstract:
"Disney reported revenues for the second quarter reached $13.3 billion and net income sto
od at $2.388 billion.Zhao Huanyan, an economist at Huamei Hotel Consulting, said the 10 m
illion mark is a number that signifies the growing importance of theme parks in the touri
sm industry in China at a time when tourists in China are growing more interested in them
e parks than natural tourism sites.Shanghai Disney Resort is a joint venture between Walt
Disney and Shanghai Shendi Group.In fiscal year 2017, Disney already had two releases tha
t topped $1 billion in global box office revenue—Rogue One and Beauty and the Beast.Beau
ty and the Beast has had a box office of 590 million yuan ($85.4 million) in China.
```

Run the program and give it a test link. It will automatically give you the abstract.

Challenges

As a matter of fact, there are too many challenges and difficulties during the whole coding process. Some challenges are too hard that I spent several days to solve by watching online tutorials or consulting classmates, friends and professors. Luckily, I have managed to solve them. I would like to name a few challenges I encounter during this project.

1. Nltk is a very new library to me. I am really grateful that John has given me some hint about this library. I found it very useful if applied in my project. I also want thank a Youtube blogger. He is very patient and made 21 tutorials about nltk and upload them online. They are all for free. I watched all of them and then start my project. Here is the tutorial link. https://www.youtube.com/watch?v=FLZvOKSCkxY&list=PLQVvvaa0QuDf2JswnfiGkliBInZnIC4HL

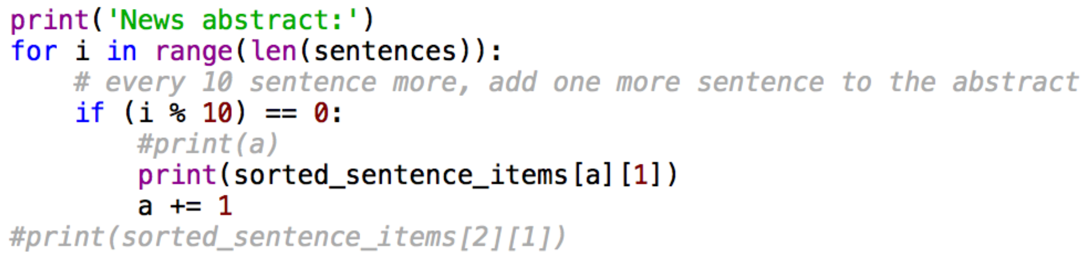
   Natural Language Processing With Python and NLTK p.1 Tokenizing words and Sentences

   sentdex

2. At first, I could not get rid of the stopwords. They keep appearing. Then I employ form a list of stopwords by using stopwords = set(stopwords.words('english')+list(punctuation)) and then use new_list=[x for x in words if x not in stopwords] to remove all the stopwords.

3. At first, it report an error saying *“UnicodeEncodeError: 'ascii' codec can't encode character '\uff0d' in position 450: ordinal not in range(128)”*. I try use ‘utf-8’ code by using f.write(str_body.encode('uft-8')). Another error is reported LookupError: unknown encoding: uft-8. The correct way is ‘UTF8’. Type error comes up: TypeError: write() argument must be str, not bytes. I transform bytes into string using ‘byte_body.decode(encoding='UTF-8')’. I successfully solved this problem.

4. When it comes to create the abstract part, at the beginning, I use the magic number 10,20 as the separation. I assume that, if the total sentence is less than 10, then take the first sentence with the highest sentence value. And if the total sentence is more than 10 but less than 20, the abstract should consist of the first 2 sentence that has the highest sentence value. However, this is kind of sloppy and not smart enough. I got a hint from the binary processing file, therefore I use the following code to fulfill my purpose.

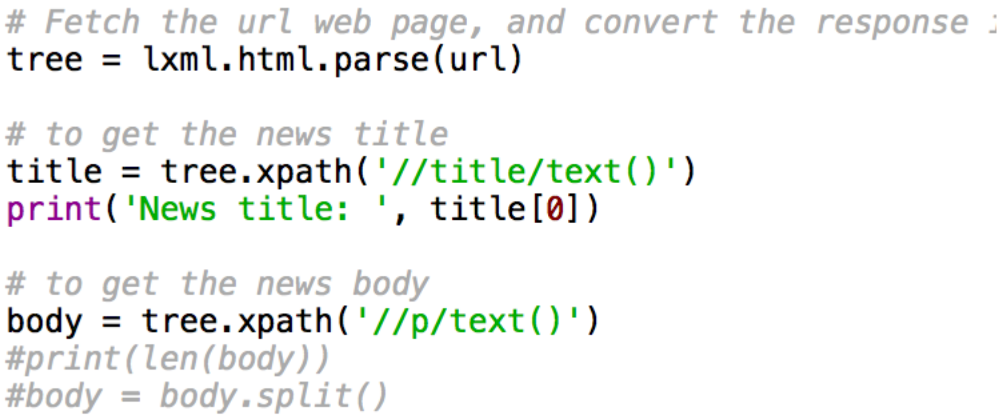


```
print('News abstract:')
for i in range(len(sentences)):
    # every 10 sentence more, add one more sentence to the abstract
    if (i % 10) == 0:
        #print(a)
        print(sorted_sentence_items[a][1])
        a += 1
#print(sorted_sentence_items[2][1])
```

Limitation

There are several major limitations of the news abstract tool. And it needs more input of efforts and improvement on coding.

1. Honestly speaking, the tool is not intelligent at all. The sentence value is based on simple addition of each word’s value. And the word’s value is calculated by using the frequency of each word divided by total words. In essence, this is not intelligent. I have read a paper written by Rada Mihalcea and Paul Tarau and the article’s name is TextRank: Bringing Order into Texts. It gives me a lot of hints about how to bring more intelligence into my programming and I will keep on exploring the tool over the summer.
2. I am afraid that my way to get news body is not able to tackle every news webpage. My code looks like this:



```
# Fetch the url web page, and convert the response
tree = lxml.html.parse(url)

# to get the news title
title = tree.xpath('//title/text()')
print('News title: ', title[0])

# to get the news body
body = tree.xpath('//p/text()')
#print(len(body))
#body = body.split()
```

If the webpage is more complicated and if the script is organized in a complex way, then my code might not be able to actually get the news. Therefore, this is another major limitation I should continue work on with.

Conclusion

In this final project, I have applied Natural Language Toolkit and the lxml to retrieve the new. I have not only got the precious chance to review the whole semester materials, but I also encourage myself to learn new knowledge. It is great fun in doing so. Finally, I want to extend my most heartfelt gratitude to the most respectful professor, John Weible, and TA, Shubhanshu Mishra. Great thanks to both of you for giving us such patient and beneficial guidance. I have a lot fun and you two do inspire my great enthusiasm towards coding. I am really lucky to have you as my mentors as I first start coding. Thank you and see you next semester.