NLP Application
Auto Summarize News Articles using Python

Himank Gupta
101512020
SEM 1

Outline

- Objective
- Procedure
- Work Flow Diagram
- Snapshots of Code
- Output

Objective

- Take in the URL of a newspaper article (from the Washington Post) , and Automatically Summarize it in 3 sentences.

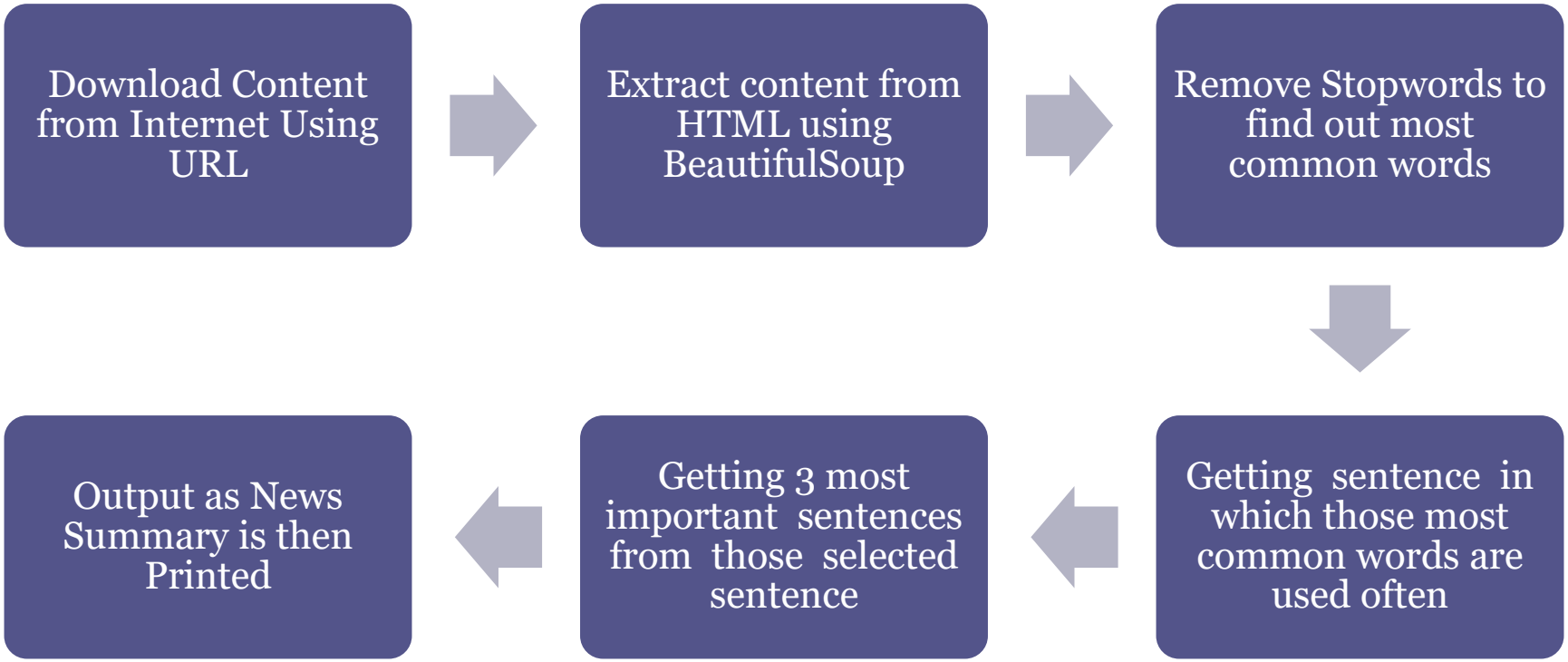
- It would be done using:
  - Natural Language processing
  - Python

Procedure

1. Download the contents of the URL.
2. Extract the Article from all the other HTML that is in the webpage.
3. Figure out which the 3 most important sentences in the article.

- The above mentioned steps will use following :
  - NLTK
  - BeautifulSoup Library

Work Flow Diagram



Snapshots of Code

```python
#HIMANK GUPTA 101512020 SEM1
#TextSummarizer

from nltk.tokenize import sent_tokenize,word_tokenize
from nltk.corpus import stopwords
from collections import defaultdict
from string import punctuation
from heapq import nlargest

class FrequencySummarizer:
    def __init__(self, min_cut=0.1, max_cut=0.9):

        self._min_cut = min_cut
        self._max_cut = max_cut
        self._stopwords = set(stopwords.words('english') + list(punctuation))

    def _compute_frequencies(self, word_sent):
        freq = defaultdict(int)

        for s in word_sent:
          for word in s:
            if word not in self._stopwords:
                freq[word] += 1

        m = float(max(freq.values()))

        for w in freq.keys():
            freq[w] = freq[w]/m
            if freq[w] >= self._max_cut or freq[w] <= self._min_cut:
                del freq[w]

        return freq

    def summarize(self, text, n):

        sents = sent_tokenize(text)
        assert n <= len(sents)
        word_sent = [word_tokenize(s.lower()) for s in sents]
        self._freq = self._compute_frequencies(word_sent)
        ranking = defaultdict(int)
```

```python
    def summarize(self, text, n):

        sents = sent_tokenize(text)
        assert n <= len(sents)
        word_sent = [word_tokenize(s.lower()) for s in sents]
        self._freq = self._compute_frequencies(word_sent)
        ranking = defaultdict(int)

        for i,sent in enumerate(word_sent):
            for w in sent:
                if w in self._freq:
                    ranking[i] += self._freq[w]
        sents_idx = nlargest(n, ranking, key=ranking.get)
        return [sents[j] for j in sents_idx]

import urllib2
from bs4 import BeautifulSoup

def get_only_text_washington_post_url(url):

    page = urllib2.urlopen(url).read().decode('utf8')
    soup = BeautifulSoup(page,"html.parser")
    text = ' '.join(map(lambda p: p.text, soup.find_all('article')))
    soup2 = BeautifulSoup(text,"html.parser")
    if soup2.find_all('p')!=[]:
        text = ' '.join(map(lambda p: p.text, soup2.find_all('p')))
    return soup.title.text, text

someUrl = "https://www.washingtonpost.com/news/the-switch/wp/2015/08/06/why-kids
textOfUrl = get_only_text_washington_post_url(someUrl)
fs = FrequencySummarizer()
summary = fs.summarize(textOfUrl[1], 3)
print(summary)
```

Output

Python 2.7.14 Shell

File Edit Shell Debug Options Window Help

```
Python 2.7.14 (v2.7.14:84471935ed, Sep 16 2017, 20:19:30) [MSC v.1500 32 bit (In
tel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
================== RESTART: C:\Users\Hp\Desktop\project.py ==================
[u'"The digital world has taken its place alongside school and friends\' houses
and extracurriculars as a place where teens go to make and strengthen friendship
s," said Amanda Lenhart, author of the report "Teens, Technology & Friendships"
and an associate\xa0director of research at Pew.', u'"Young people are very awar
e that people have highly curated images and that text fights can quickly go out
 of control and they are trying to sort it all out," said Rosalind Wiseman, auth
or of "Queen Bees and Wannabes" and speaker on youth issues.', u"Nearly half of
those surveyed say they've at least occasionally seen posts about events that th
ey were invited to; and\xa085 percent said they think social media users present
 a carefully crafted image of themselves online that may not be authentic."]
>>>
```

THANK YOU !!!