65070501037
Paweekorn Soratyathorn

Summarizing BBC news

In this lab, you will self-study two unsupervised graph-based summarization methods, namely LexRank and TextRank, and apply them to summarize news data.

First of all, download data and extract files.

```
# importing required modules
from zipfile import ZipFile

with ZipFile('bbc-fulltext.zip', 'r') as zip:
    # printing all the contents of the zip file
    zip.printdir()

    # extracting all the files
    print('Extracting all the files now...')
    zip.extractall()
    print('Done!')
```

```
File Name                                             Modified             Size
bbc/                                           2015-04-05 16:29:08            0
bbc/entertainment/                             2010-03-30 00:45:20            0
bbc/entertainment/289.txt                      2010-03-30 00:45:20         2261
bbc/entertainment/262.txt                      2010-03-30 00:45:20         4810
bbc/entertainment/276.txt                      2010-03-30 00:45:20         2127
bbc/entertainment/060.txt                      2010-03-30 00:45:20         1046
bbc/entertainment/074.txt                      2010-03-30 00:45:20         1586
bbc/entertainment/048.txt                      2010-03-30 00:45:20         2121
bbc/entertainment/114.txt                      2010-03-30 00:45:20         1481
bbc/entertainment/100.txt                      2010-03-30 00:45:20         1821
bbc/entertainment/128.txt                      2010-03-30 00:45:20         1238
bbc/entertainment/316.txt                      2010-03-30 00:45:20         3092
bbc/entertainment/302.txt                      2010-03-30 00:45:20          956
bbc/entertainment/303.txt                      2010-03-30 00:45:20         2516
bbc/entertainment/317.txt                      2010-03-30 00:45:20         1034
bbc/entertainment/129.txt                      2010-03-30 00:45:20         1449
bbc/entertainment/101.txt                      2010-03-30 00:45:20         1946
bbc/entertainment/115.txt                      2010-03-30 00:45:20         1534
bbc/entertainment/049.txt                      2010-03-30 00:45:20         1601
bbc/entertainment/075.txt                      2010-03-30 00:45:20         2198
bbc/entertainment/061.txt                      2010-03-30 00:45:20         4178
bbc/entertainment/277.txt                      2010-03-30 00:45:20         1448
bbc/entertainment/263.txt                      2010-03-30 00:45:20         3401
bbc/entertainment/288.txt                      2010-03-30 00:45:20         1981
bbc/entertainment/275.txt                      2010-03-30 00:45:20         3926
bbc/entertainment/261.txt                      2010-03-30 00:45:20         1449
bbc/entertainment/249.txt                      2010-03-30 00:45:20          862
bbc/entertainment/088.txt                      2010-03-30 00:45:20         1586
bbc/entertainment/077.txt                      2010-03-30 00:45:20         1899
bbc/entertainment/063.txt                      2010-03-30 00:45:20         1985
bbc/entertainment/103.txt                      2010-03-30 00:45:20         2317
bbc/entertainment/117.txt                      2010-03-30 00:45:20         1643
bbc/entertainment/301.txt                      2010-03-30 00:45:20         1773
bbc/entertainment/315.txt                      2010-03-30 00:45:20         1626
bbc/entertainment/329.txt                      2010-03-30 00:45:20         1468
bbc/entertainment/328.txt                      2010-03-30 00:45:20         1896
bbc/entertainment/314.txt                      2010-03-30 00:45:20         6100
bbc/entertainment/300.txt                      2010-03-30 00:45:20         2998
bbc/entertainment/116.txt                      2010-03-30 00:45:20         1354
bbc/entertainment/102.txt                      2010-03-30 00:45:20         1338
bbc/entertainment/062.txt                      2010-03-30 00:45:20         1551
bbc/entertainment/076.txt                      2010-03-30 00:45:20         1376
bbc/entertainment/089.txt                      2010-03-30 00:45:20          878
bbc/entertainment/248.txt                      2010-03-30 00:45:20         2848
bbc/entertainment/260.txt                      2010-03-30 00:45:20         1119
bbc/entertainment/274.txt                      2010-03-30 00:45:20         1039
bbc/entertainment/258.txt                      2010-03-30 00:45:20         2080
bbc/entertainment/270.txt                      2010-03-30 00:45:20         1139
bbc/entertainment/264.txt                      2010-03-30 00:45:20         3294
bbc/entertainment/099.txt                      2010-03-30 00:45:20         1420
bbc/entertainment/072.txt                      2010-03-30 00:45:20         2018
bbc/entertainment/066.txt                      2010-03-30 00:45:20         1824
bbc/entertainment/106.txt                      2010-03-30 00:45:20         1164
bbc/entertainment/112.txt                      2010-03-30 00:45:20          987
bbc/entertainment/338.txt                      2010-03-30 00:45:20         1275
bbc/entertainment/304.txt                      2010-03-30 00:45:20         1361
bbc/entertainment/310.txt                      2010-03-30 00:45:20         1015
```

Below, Politics news is selected. *(Note that you are free to use other categories as you would like i.e. tech, sports, business, and entertainment.)*

In the Politics category, there are 417 news articles. The goal is to summarize **each news article**, at least 10 news. The compression ratio should be within 25%-30%.

```
# !pip install path
from path import Path

documents = []
documents_dir = Path('bbc/politics')
for file_path in documents_dir.files('*.txt'):
    with file_path.open(mode='rt', encoding='utf-8') as fp:
        documents.append(fp.readlines())
```

Use sentences in one of the news *as an example*.

```
sentences = documents[0]
print(sentences)
```

```
['Labour plans maternity pay rise\n', '\n', 'Maternity pay for new mothers is to rise by £1,400 as part of new proposals announced b
```

LexRank

TODO #1: Study an algorithm of LexRank and describe how it works.

TODO #2: Use the LexRank library to summarize data as shown in the example below.

Note: Make sure that, in your final summary the selected sentences must be ordered chronologically.

Reference: LexRank library

Run LexRank to summarize input document.

```
# !pip install lexrank
from lexrank import STOPWORDS, LexRank
lxr = LexRank(documents, stopwords=STOPWORDS['en'])
```

Get scores of each sentence.

```
# 'fast_power_method' speeds up the calculation, but requires more RAM
scores_cont = lxr.rank_sentences(sentences,
                                 threshold=None,
                                 fast_power_method=False,)
print(scores_cont)
```

```
[1.10540489 1.         1.05086576 1.         1.08395518 1.
 1.11241192 1.         1.04556705 1.         0.60179519]
```

Print high-ranked sentences.

```
summary = lxr.get_summary(sentences, summary_size=2, threshold=.25)
print(summary)
```

```
['Ms Hewitt also stressed the plans would be paid for by taxpayers, not employers. But David Frost, director general of the British
```

```
# get summary with continuous LexRank
summary_cont = lxr.get_summary(sentences, threshold=None)
print(summary_cont)
```

```
['Ms Hewitt said: "We have already doubled the length of maternity pay, it was 13 weeks when we were elected, we have already taken
```

TextRank

TODO #3: Study an algorithm of TextRank and describe how it works.

TODO #4: Use the TextRank library to summarize data as shown in the example below.

Note: Make sure that, in your final summary the selected sentences must be ordered chronologically.

Reference: [TextRank library]

Join all sentences into one piece of text.

```
text = ' '.join(sentences)
print(text)
```

```
Labour plans maternity pay rise

Maternity pay for new mothers is to rise by £1,400 as part of new proposals announced by the Trade and Industry Secretary Patricia

It would mean paid leave would be increased to nine months by 2007, Ms Hewitt told GMTV's Sunday programme. Other plans include let

Ms Hewitt said: "We have already doubled the length of maternity pay, it was 13 weeks when we were elected, we have already taken i

She said the Conservatives would announce their proposals closer to the General Election. Liberal Democrat spokeswoman for women Sa

Ms Hewitt also stressed the plans would be paid for by taxpayers, not employers. But David Frost, director general of the British C
```

```
from summa.summarizer import summarize
print(summarize(text, ratio=0.25))
```

```
Maternity pay for new mothers is to rise by £1,400 as part of new proposals announced by the Trade and Industry Secretary Patricia H
The Tories dismissed the maternity pay plan as "desperate", while the Liberal Democrats said it was misdirected.
"We are going to extend the pay to nine months by 2007 and the aim is to get it right up to the full 12 months by the end of the nex
We will definitely extend the maternity pay, from the six months where it now is to nine months, that's the extra £1,400." She said
```

Lab part

```
lab_docs = []
documents_dir = Path('bbc/tech')
for file_path in documents_dir.files('*.txt'):
    with file_path.open(mode='rt', encoding='utf-8') as fp:
        lab_docs.append(fp.readlines())
```

1. Study an algorithm of LexRank and describe how it works.
   Ans. It is the extractive-based text summarization algorithm which construct a graph data structure and using eigenvector concept for choosing an important sentences to be the result. Start by calculating TF-IDf for each words in document then construct a matrix with |s| × |s| where |s| = no. of sentences in the document. After construct a matrix, it calculating the idf-modified-similarity for each poisition in the matrix. At this point, we will get the relationship of every pair of sentences. Finally, it calculate the centrality probability to determine which sentence is important for document context and the threshold is the average probability of all sentences.
2. Use the LexRank library to summarize article.\

- I use continuous LexRank because I want algorithm to concern about the context.

```
def lexRankSummary(num):
    result = []
    for i in range(num):
        print(f'Article {i + 1}')
        sentences = lab_docs[i]
        scores_cont = lxr.rank_sentences(sentences,
                                         threshold=None,
                                         fast_power_method=True,)
        print(scores_cont)

        summary = lxr.get_summary(sentences, threshold=None)
        summary = [i.replace('\n', '') for i in summary]
        sum_text = ' '.join(summary)
        print(sum_text, '\n')

        result.append(f'## Article {i+1}\n{' '.join(sentences)}\n### Summary\n{sum_text}\n\n')

    return result

```

```
20
21 lexRank = lexRankSummary(10)
```

```
Article 1
[1.07991332 1.         0.9151488  1.         0.81387248 1.
 0.78406742 1.         1.09394175 1.         1.05256816 1.
 0.76561659 1.         1.33278971 1.         1.16208175 1.
 1.        ]
The author of one such article began a petition drive against the use of the ink. The greatest part of the opposition to ink has

Article 2
[0.78798173 1.         1.01448895 1.         1.04296871 1.
 1.02844597 1.         1.12611465]
Net cafes are hugely popular in China because the relatively high cost of computer hardware means that few people have PCs in the

Article 3
[1.01564315 1.         1.19006563 1.         1.14842376 1.
 0.5620037  1.         1.08386376]
Microsoft is investigating a trojan program that attempts to switch off the firm's anti-spyware software.

Article 4
[0.95926489 1.         1.12180566 1.         1.03642977 1.
 1.09740204 1.         0.78509763]
Nicholas Negroponte, chairman and founder of MIT's Media Labs, says he is developing a laptop PC that will go on sale for less th

Article 5
[1.17875307 1.         0.92447435 1.         1.20778099 1.
 0.83562351 1.         0.82113454 1.         0.71387239 1.
 1.31836113]
Technology as a way of unleashing creativity has massive potential, not least because it gives people something to do with their

Article 6
[1.00547859 1.         1.19581057 1.         0.95165797 1.
 0.73731704 1.         1.05323028 1.         1.16315545 1.
 1.15530593 1.         0.73804416]
A network of community computer centres, linked by wireless technology, is providing a helping hand for poor farmers in Peru.

Article 7
[0.82272506 1.         1.11288435 1.         0.94122872 1.
 1.12316186]
One of the critical patches Microsoft has made available is an important one that fixes some IE flaws. Stephen Toulouse, a Micros

Article 8
[1.1159979  1.         0.9496307  1.         1.05021751 1.
 0.79239631 1.         1.09175758]
Virus poses as Christmas e-mail

Article 9
[1.86063727 1.         2.01222463 1.         1.22385108 1.
 1.02097262 1.         1.8413613  1.         0.87482224 1.
 0.76090405 1.         0.5584398  1.         0.52560314 1.
 1.34338455 1.         1.01292418 1.         0.99733168 1.
 1.55096263 1.         1.         1.         0.77006478 1.
 1.25359404 1.         0.96696017 1.         0.89613042 1.
 1.40963352 1.         0.72379796 1.         0.57824214 1.
 0.84591855 1.         1.23170064 1.         0.4423852  1.
 0.96578546 1.         1.20701621 1.         0.50007819 1.
 1.40962004 1.         1.03088278 1.         0.4475659  1.
 0.72184205 1.         0.9444816  1.         0.69788429 1.
 0.50943346 1.         0.86356344]
```

```
1 with open('LexRank.txt', 'w', encoding='utf-8') as f:
2     for line in lexRank:
3         f.write(f"{line}")
```

3. Study an algorithm of TextRank and desctribe how it works

   Ans. This algorithm uses graph-based structure same as LexRank but it calculates the similarity between the sentences to determine the important of each sentences. Finally, the algorithm ranks the sentences from graph which each node represents the sentences from document.

4. Use the TextRank to summarize article.

```
1 def textRankSummary(num):
2     result = []
3     for i in range(num):
4         sentences = lab_docs[i]
5         clean_sent = [sentence.replace('\n', '') for sentence in sentences]
6         text = ' '.join(clean_sent)
7
8         sum_text = summarize(text, ratio=0.25)
9         print(sum_text, '\n')
```

```
        record = f'## Article {i + 1}\n {' '.join(sentences)}\n### Summary\n{sum_text}\n\n'
        result.append(record)

    return result

textRank = textRankSummary(10)
```

```
Ink helps drive democracy in Asia  The Kyrgyz Republic, a small, mountainous state of the former Soviet republic, is using invisi
In an effort to live up to its reputation in the 1990s as "an island of democracy", the Kyrgyz President, Askar Akaev, pushed thr
The use of ink is only one part of a general effort to show commitment towards more open elections - the German Embassy, the Sorc
At the entrance to each polling station, one election official will scan voter's fingers with UV lamp before allowing them to ent
The other common type of ink in elections is indelible visible ink - but as the elections in Afghanistan showed, improper use of
The use of "invisible" ink is not without its own problems.
The use of ink and readers by itself is not a panacea for election ills.

China net cafe culture crackdown  Chinese authorities closed 12,575 net cafes in the closing months of 2004, the country's governm
China has long been worried that net cafes are an unhealthy influence on young people.
This is not the first time that the Chinese government has moved against net cafes that are not operating within its strict guide
Laws on net cafe opening hours and who can use them were introduced in 2002 following a fire at one cafe that killed 25 people.

Microsoft said it did not believe the program was widespread and recommended users to use an anti-virus program.
The program attempts to disable or delete Microsoft's anti-spyware tool and suppress warning messages given to users.

He said the child could use the laptop like a text book.
However, Mr Negroponte has adapted the idea to his own work in Cambodia where he set up two schools together with his wife and ga
"We put in 25 laptops three years ago , only one has been broken, the kids cherish these things, it's also a TV a telephone and a

Technology gets the creative bug  The hi-tech and the arts worlds have for some time danced around each other and offered creativ
The idea, says BT, is to shape a "21st Century model" which will help cement the art, technology, and business worlds together.
"We are hoping to understand the creative industry that has a natural thirst for broadband technology," said Frank Stone, head of
Using a 3D graphics engine, the type commonly used in gaming, Bafta-winning artists Langlands & Bell have created a virtual, stor
But collaboration between art and digital technology is by no means new, and many keen coders, designers, games makers and animat
The art world is "fantastically rich", said Mr Stone, with creative people and ideas which means traditional companies like BT wa
The partnership between artists and technologists is part of trying to understand the creative potential of technologies like bro
It is about both industries borrowing strategies and creative ideas together which can result in better business practices for cr

Wi-fi web reaches farmers in Peru  A network of community computer centres, linked by wireless technology, is providing a helping
The Agricultural Information Project for Farmers of the Chancay-Huaral Valley also provides vital links between local organisatio
The Board of Irrigation Users which runs the computer centres, aims to make the network self-sustainable within three years, thro
One of the key elements of the project is the Agricultural Information System, with its flagship huaral.org website.
"Throughout the last three years, the people have provided a vital thrust to the project; they feel it belongs to them," said Mr
But we have also had a great feedback when we trained 40 or 50-year old women, who were seeing a computer for the first time in t

Microsoft releases bumper patches  Microsoft has warned PC users to update their systems with the latest security fixes for flaws
One of the critical patches Microsoft has made available is an important one that fixes some IE flaws.
Often, when a critical flaw is announced, spates of viruses follow because home users and businesses leave the flaw unpatched.

The Zafi.D virus translates the Christmas greeting on its subject line into the language of the person receiving infected e-mail.
Anti-virus firm Sophos said that 10% of the e-mail currently on the net was infected with the Zafi virus.
The virus' subject line says "Merry Christmas" and translates this into one of 15 languages depending of the final suffix of the

Apple laptop is 'greatest gadget'  The Apple Powerbook 100 has been chosen as the greatest gadget of all time, by US magazine Mob
The 1991 laptop was chosen because it was one of the first "lightweight" portable computers and helped define the layout of all f
The magazine specified that gadgets also needed to be a "self-contained apparatus that can be used on its own, not a subset of an
The oldest "gadget" in the top 100 is the abacus, which the magazine dates at 190 A.D., and put in 60th place.
Other pre-electronic gadgets in the top 100 include the sextant from 1731 (59th position), the marine chronometer from 1761 (42nd
The Tivo personal video recorder is the newest device to make the top 10, which also includes the first flash mp3 player (Diamoun
The most popular gadget of the moment, the Apple iPod, is at number 12 in the list while the first Sony transistor radio is at nu
Karl Elsener's knife, the Swiss Army Knife from 1891, is at number 20 in the list.
Gadgets which could be said to feature surprisngly low down in the list include the original telephone (23rd), the Nintendo GameB
A laptop computer is not a gadget!
Surely the most important gadget of the modern age is the mobile phone?
From outside the modern age, the marine chronometer is the single most important gadget, without which modern transportation syst
Of the electronic gadgets, thousands of journalists in the early 1980s blessed the original noteboook pc - the Tandy 100.
```

```
with open('textRank.txt', 'w', encoding='utf-8') as f:
    for line in textRank:
        f.write(f"{line}")
```

```
Start coding or generate with AI.
```