Assignment 3: Trích xuất dữ liệu văn bản

Tổng quan:

- Trong bài này chúng ta sẽ lần lượt thực hành trích xuất dữ liệu văn bản từ các tài liệu pdf, docx, JSON
- Sau đó ta sẽ làm sạch (clean) các text thu được sử dụng biểu thức chính quy
- Bài tập yêu cầu các kiến thức về lập trình python với các thư viện: PyPDF2, docx, json, re

Câu hỏi 1: trích xuất dữ liệu từ file pdf

```
# Cài đặt thư viện PyPDF2
# !pip install PyPDF2  # bỏ dấu "#" đầu dòng để chạy cài đặt
```

```
# import thư viện
import PyPDF2
from PyPDF2 import PdfFileReader
```

Hoàn thiện đoạn chương trình sau:

- Đọc nội dung file và các thông tin về số trang của file "News1.pdf"
- Nội dung file được lưu vào biến dạng string

```
#### YOUR CODE HERE ####

#Creating a pdf file object
pdf = open("News1.pdf","rb")
#creating pdf reader object
pdf_reader = PyPDF2.PdfFileReader(pdf)
#checking number of pages in a pdf file
print(pdf_reader.numPages)
#creating a page object
page = pdf_reader.getPage(0)
#finally extracting text from the page
pdf_text = page.extractText()
print(pdf_text)
#closing the pdf file
pdf.close()

#### END YOUR CODE #####
```

```
3
The firm, which is now one of the biggest investors in Goog
le, benefited from sales of
high-speed internet connections and higher advert sales. TimeWarner said fourth
quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one
```

```
-off gains
which offset a profit dip at Warner Bros, and less users for
 AOL.
 Time Warner said on Friday that it now owns 8% of search
-engine Google. But its own
internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth
quarter profits were lower than in the preceding three quarters. However, the comp
any
said AOL's underlying profit before exceptional items rose 8% on the back of stronger
internet advertising revenues. It hopes to increase subscribers by offering the online
service free to TimeWarner internet customers and will try to sign up AOL's exi
sting
customers for high
-speed broadband. TimeWarner also has to restate 2000 and 2003
results following a probe by the US Securities Exchange Commission (SEC), which is
close to concluding.
 Time Warner's fourth quarter profits were slightly better than an
alysts' expectations. But
its film division saw profits slump 27% to $284m, helped by box
-office flops Alexander
and Catwoman, a sharp contrast to year
-earlier, when the third and final film in the Lord
of the Rings trilogy boosted results. For the full
-year, TimeWarner posted a profit of
$3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn.
"Our financial performance was strong, meeting or exceeding all of our full
-year
objectives and greatly enhancing our flexibility," chairman
and chief executive Richard
Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around
5%, and also expects higher revenue and wider profit margins.
 TimeWarner is to restate its accounts as part of efforts to resolve an inquiry in
to AOL by
US market regulators. It has already offered to pay $300m to settle charges, in a deal
that is under review by the SEC. The company said it was unable to estimate the
amount it needed to set aside for legal reserves, which it previously set at $5
00m. It
intends to adjust the way it accounts for a deal with German music publisher
Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising
revenue. It will now book the sale of its stake in AOL Europe as a loss on the value
of that stake.
 nd Alan Greenspan highlighted the US government's willingness to curb spending and
rising household savings as factors which may help to reduce it. In late trading in New
York, the dollar reached $1.2871 against the euro, from $1.2974 on Thu
rsday. Market
concerns about the deficit has hit the greenback in recent months. On Friday, Federal
Reserve chairman Mr Greenspan's speech in London ahead of the meeting of G7
finance ministers sent the dollar higher after it had earlier tumbled on the bac
k of worse
-than-expected US jobs data. "I think the chairman's taking a much more sanguine view
on the current account deficit than he's taken for some time," said Robert Sinche, head
of currency strategy at Bank of America in New York. "He's taking a long
er-term view,
laying out a set of conditions under which the current account deficit can improve this
```

Câu hỏi 2: trích xuất dữ liệu văn bản từ file Word

Hoàn thiện đoạn chương trình sau:

- Trích xuất dữ liệu từ file "News.docx"

- Nội dung lưu vào một biến string

```
# Cài đặt thư viện docx
#!pip install docx  # bỏ dấu "#" đầu dòng để chạy cài đặt


#Import library
from docx import Document


#### YOUR CODE HERE ####

#Creating a word file object
doc = open("News2.docx","rb")
#creating word reader object
document = Document(doc)
docx_text = ""
for para in document.paragraphs:
       docx_text += para.text
#to see the output call docu
print(docx_text)

#### END YOUR CODE #####
```

```
     Revised figures indicated growth of just 0.1% - and a similar-sized contraction in the previous q
```

Câu hỏi 3: trích xuất văn bản từ file json

```
# import thư viện
import json
```

Hoàn thiện đoạn chương trình sau:

- Đọc dữ liệu từ file "News3.json"
- Sau đó nối các bản tin (News) lại, và lưu vào một biến string, mỗi bản tin trên một dòng.

```
#### YOUR CODE HERE ####

#json from "https://quotes.rest/qod.json"
with open('News3.json') as f:
    data = json.load(f)
f.close()
json_text = "\n".join(data.values())
print(json_text)

#### END YOUR CODE #####
```

```
     The Telegraph Group says the cuts are needed to fund an £150m investment in new printing facilitie
     The National Union of Journalists said it stood strongly behind the journalists and did not rule o
```

```
Some broadsheet newspapers - especially those which have not moved to a tabloid format - have suf
The Guardian is hedging its bets, planning a larger tabloid format like those popular in continen
Millions of Indonesians use kerosene for basic cooking, and prices have been heavily subsidised fo
Indonesia pays subsidies to importers in order to stabilise domestic fuel prices, but higher oil p
```

Câu hỏi 4: Xử lý dữ liệu đã trích chọn được

Câu hỏi 4.1: từ các dữ liệu trích chọn từ các câu hỏi 1, 2, 3. Hãy nối chúng lại với nhau vào một biến string

```
#### YOUR CODE HERE ####

news_text = "\n".join([pdf_text, docx_text, json_text])
print(news_text)

#### END YOUR CODE #####
```

```
     The firm, which is now one of the biggest investors in Goog
     le, benefited from sales of
     high-speed internet connections and higher advert sales. TimeWarner said fourth
     quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one
     -off gains
     which offset a profit dip at Warner Bros, and less users for
      AOL.
      Time Warner said on Friday that it now owns 8% of search
     -engine Google. But its own
     internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth
     quarter profits were lower than in the preceding three quarters. However, the comp
     any
     said AOL's underlying profit before exceptional items rose 8% on the back of stronger
     internet advertising revenues. It hopes to increase subscribers by offering the online
     service free to TimeWarner internet customers and will try to sign up AOL's exi
     sting
     customers for high
     -speed broadband. TimeWarner also has to restate 2000 and 2003
     results following a probe by the US Securities Exchange Commission (SEC), which is
     close to concluding.
      Time Warner's fourth quarter profits were slightly better than an
     alysts' expectations. But
     its film division saw profits slump 27% to $284m, helped by box
     -office flops Alexander
     and Catwoman, a sharp contrast to year
     -earlier, when the third and final film in the Lord
     of the Rings trilogy boosted results. For the full
     -year, TimeWarner posted a profit of
     $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn.
     "Our financial performance was strong, meeting or exceeding all of our full
     -year
     objectives and greatly enhancing our flexibility," chairman
     and chief executive Richard
     Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around
     5%, and also expects higher revenue and wider profit margins.
      TimeWarner is to restate its accounts as part of efforts to resolve an inquiry in
```

```
to AOL by
US market regulators. It has already offered to pay $300m to settle charges, in a deal
that is under review by the SEC. The company said it was unable to estimate the
amount it needed to set aside for legal reserves, which it previously set at $5
00m. It
intends to adjust the way it accounts for a deal with German music publisher
Bertelsmann's purchase of a stake in AOL Europe, which it had reported as advertising
revenue. It will now book the sale of its stake in AOL Europe as a loss on the value
of that stake.
 nd Alan Greenspan highlighted the US government's willingness to curb spending and
rising household savings as factors which may help to reduce it. In late trading in New
York, the dollar reached $1.2871 against the euro, from $1.2974 on Thu
rsday. Market
concerns about the deficit has hit the greenback in recent months. On Friday, Federal
Reserve chairman Mr Greenspan's speech in London ahead of the meeting of G7
finance ministers sent the dollar higher after it had earlier tumbled on the bac
k of worse
-than-expected US jobs data. "I think the chairman's taking a much more sanguine view
on the current account deficit than he's taken for some time," said Robert Sinche, head
of currency strategy at Bank of America in New York. "He's taking a long
er-term view,
```

Câu hỏi 4.2 Hoàn thiện hàm xử lý chuỗi bản tin

Mô tả hàm: hàm này nhận tham số đầu vào là một chuỗi, và trả ra chuỗi đã được xử lý, công việc chính trong hàm như sau:

- Thay thế các ký tự "^A-Za-z0-9(),!?'`" bằng dấu " "
- Thay thế "'s" thành " 's"
- Thay thế "'ve" thành " 've"
- Thay thế "n't" thành " n't"
- Thay thế "'re" thành " 're"
- Thay thế "'d" thành " 'd"
- Thay thế "'ll" thành " 'll"
- Thay thế "," thành " , "
- Thay thế "!" thành " ! "
- Thay thế "(" thành " ("
- Thay thế ")" thành ") "
- Thay thế "?" thành " ? "
- Thay thế "\s{2,}" thành " "
- Cắt bỏ các khoảng trắng ở đầu
- Chuyển văn bản thành chữ thường

Gợi ý: sử dụng re.sub()

```
# import thư viện regular expression
import re


#### YOUR CODE HERE ####
```

```
def clean_str(string):
    """
    Tokenization/string cleaning for all datasets except for SST.
    Original taken from https://github.com/yoonkim/CNN_sentence/blob/master/proce
    """
    string = re.sub(r"[^A-Za-z0-9(),!?\'\`]", " ", string)
    string = re.sub(r"\'s", " \'s", string)
    string = re.sub(r"\'ve", " \'ve", string)
    string = re.sub(r"n\'t", " n\'t", string)
    string = re.sub(r"\'re", " \'re", string)
    string = re.sub(r"\'d", " \'d", string)
    string = re.sub(r"\'ll", " \'ll", string)
    string = re.sub(r",", " , ", string)
    string = re.sub(r"!", " ! ", string)
    string = re.sub(r"\(", " \( ", string)
    string = re.sub(r"\)", " \) ", string)
    string = re.sub(r"\?", " \? ", string)
    string = re.sub(r"\s{2,}", " ", string)
    return string.strip().lower()

#### END YOUR CODE #####
```

Kiểm tra kết quả với hàm vừa viết trên dữ liệu đã trích xuất

```
print (clean_str(news_text))
```

```
     the firm , which is now one of the biggest investors in goog le , benefited from sales of high sp
```