

Web Mining (CSE3024)

Lab Assignment 10

Name: Kritika Mishra

Registration Number: 16BCI0041

Slot: L15+L16

Date: 24th October 2018

Question:

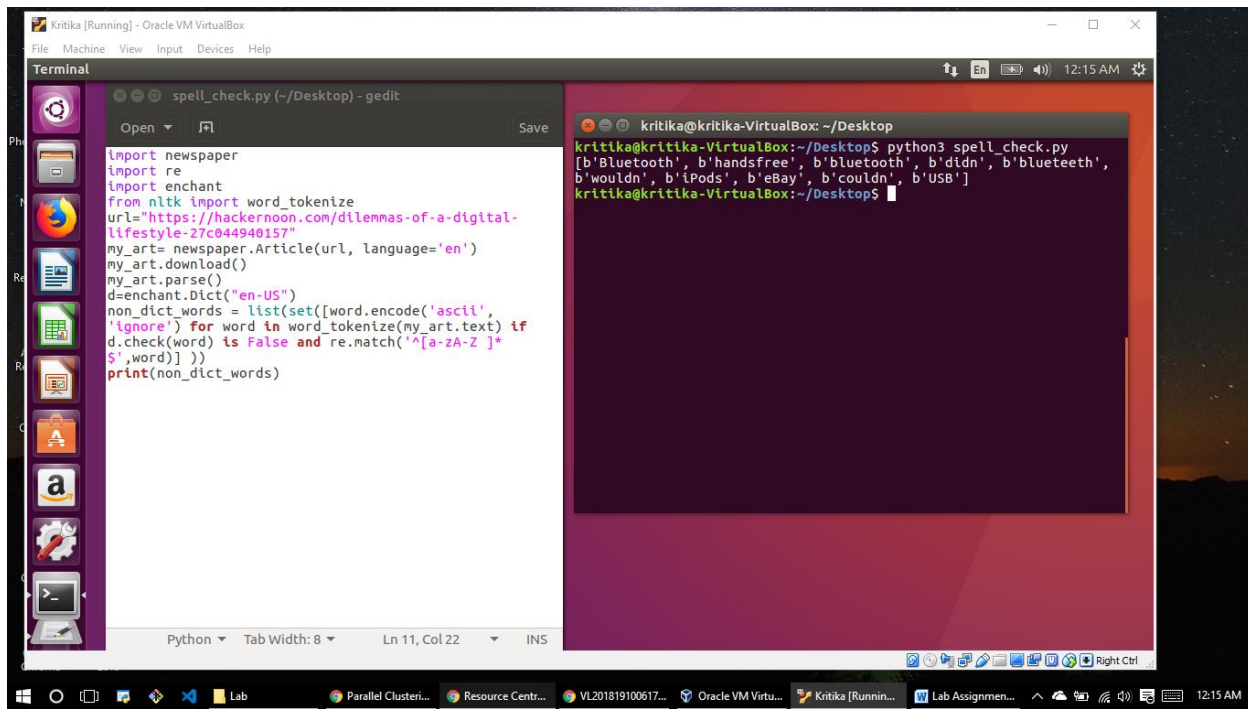
Write a python code to find the miss-spell words in a web page. Use appropriate packages to perform the following

1. Input: Define url (link) from whose spelling must be checked
2. Extract: Tokenize (split the complete article into bag of words)
3. Match: Cross-validate the extracted words against English dictionary words
4. Output: List down the words that didn't match (those are mis-spelt / non-dictionary words)

Code:

```
import newspaper
from nltk import word_tokenize
import enchant
import re
url = 'https://hackernoon.com/dilemmas-of-a-digital-
lifestyle-27c044940157' #input URL
my_article = newspaper.Article(url, language='en')
my_article.download()
my_article.parse()
d = enchant.Dict("en_US")
print(list(set([word.encode('ascii', 'ignore') for word in
word_tokenize(my_article.text) if d.check(word) is False
and re.match('^[a-zA-Z ]*$',word)] )))
```

Output:



The screenshot displays a virtual machine window titled "Kritika [Running] - Oracle VM VirtualBox". Inside the VM, a terminal window is open, showing the execution of a Python script named `spell_check.py`. The script is located at `~/Desktop` and is being edited in a text editor. The script's code is as follows:

```
import newspaper
import re
import enchant
from nltk import word_tokenize
url="https://hackernoon.com/dilemmas-of-a-digital-lifestyle-27c044940157"
my_art= newspaper.Article(url, language='en')
my_art.download()
my_art.parse()
d=enchant.Dict("en-US")
non_dict_words = list(set([word.encode('ascii',
'ignore') for word in word_tokenize(my_art.text) if
d.check(word) is False and re.match('[a-zA-Z ]*$',word)] ))
print(non_dict_words)
```

The terminal output shows the results of running the script:

```
kritika@kritika-VirtualBox: ~/Desktop$ python3 spell_check.py
[b'Bluetooth', b'handsfree', b'bluetooth', b'didn', b'blueteeth',
b'wouldn', b'iPods', b'eBay', b'couldn', b'USB']
kritika@kritika-VirtualBox: ~/Desktop$
```

The terminal window is titled `kritika@kritika-VirtualBox: ~/Desktop`. The background of the VM desktop is a dark, abstract image. The taskbar at the bottom of the VM shows several open applications, including "Parallel Clusters...", "Resource Centr...", "VL201819100617...", "Oracle VM Virtu...", "Kritika [Runnin...", and "Lab Assignmen...". The system clock in the bottom right corner indicates the time is 12:15 AM.