Web Mining Lab Assignment

Name: Om Ashish Mishra

Registration Number: 16BCE0789

Slot: F2

Q.

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') # ### Extracting the Article Content

my\_article.download()

my\_article.parse()

my\_article.text

d = enchant.Dict("en\_US")

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:**

