Web Mining Lab Assignment 10

## Name: Om Ashish Mishra

## Registration Number: 16BCE0789

## Slot: F2

# The 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 / nondictionary words)

# The Answer

Code:

**from** **urlparse** **import** urlparse

**>>>** o = urlparse('http://www.cwi.nl:80/*%7E*guido/Python.html')

**>>>** o *# doctest: +NORMALIZE\_WHITESPACE*

ParseResult(scheme='http', netloc='www.cwi.nl:80', path='/%7Eguido/Python.html',

params='', query='', fragment='')

**>>>** o.scheme

'http'

**>>>** o.port

80

**>>>** o.geturl()

'http://www.cwi.nl:80/%7Eguido/Python.html'