Série TP n°2

Chapitre 1: introduction

Module	**TAL** trait. Auto. Lang. natur.	
Filière	**Master GSI**	**1ère Année**

1- pratique : exécuter le code suivant sous python

Utiliser un fichier text appelé « text.txt »

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

# importer la bibliothèque sys
# on veut l'utiliser pour quiter le programme en cas d'erreur
import sys
# lire du texte à partir d'un fichier

def treat_line(line):
    """
    traiter une ligne de texte.
    """
    return line

def read_file(filename):
    try:
        myfile = open(filename)
    except:
        print "Can't open file ", filename
        sys.exit()
    # lecture des lignes de texte
    return myfile.readlines();

def treat_text(lines):
    for line in lines:
        print line


# La fonction main
if __name__ == '__main__':
    DATA_FILE = "text.txt"
    lines = read_file(DATA_FILE)
    treat_text(lines);
```

2- Tester la fonction tokenize()

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

# regular expression
import re
def tokenize(text):
    """ extraire les mots """
    # diviser la lignes par les espaces
    list_word =  text.split(" ")
    return list_word

# La fonction main
if __name__ == '__main__':
    text = "I'm Very Hungry, I want to eat something."
    tokens = tokenize(text);
    print tokens
```

3- On veut proposer plusieurs implémentation de la fonction tokenize(), tester les différentes fonctions :

```
# tokenize par des expression régulière simple
def tokenize_regex_punct(text):
    """ extraire les mots"""
    tokens = re.split("[.,:; ]+", text)
    return tokens


# tokenize par des expression régulière simple, en gardant la ponctuation
def tokenize_regex_punct_keep(text):
    """ extraire les mots"""
    tokens = re.split("([.,:; ]+)", text)
    return tokens

# tokenize par des expression régulière
def tokenize_regex(text):
    """ extraire les mots"""
    tokens = re.split("\W+", text)
    return tokens

# tokenize par des expression régulière, en gardant la ponctuation
def tokenize_regex_keep_punct(text):
    """ extraire les mots"""
    tokens = re.split("(\W+)", text)
    return tokens
```

Noter les différences entre les différentes implémentation de la fonction tokenize

4- Appliquer la fonction tokenize() sur le texte lu à partir du fichier.

Travail à domicile :

1- Ecrire un programme python qui permet de calculer la fréquence des mots dans un texte donné.

On veut traiter les textes normalisés. La normalisation est la suppression des diacritiques en arabe, les accents en français.

Choisir au moins une langue pour

a- Le fichier contient un long texte en français, le calcul de la fréquence doit ignorer les accents. Par exemple les deux mots « donnee » et « donnée » sont indexé dans la même entrées.

b- Le fichier contient un long text en arabe, le calcul de la fréquence ignore les diacritiques 'Harakat'.

Et les Hamza affectées au lettres Yeh, Alef, Waw, c'est-à-dire : ئ devient ي. Les lettres أآإ devient ا.

ؤ devient و..