Série TP n°4

Chapitre 1: niveau lexical

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

Le codage

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

import re
# la liste des suffixes
SUFFIX_LIST=['ique', 'ation', 'tion', 'é', 'er','eur', 'ien']

def stemming(word, suffix_list):
    """ stem a word"""
    stem_list =[]
    for suffix in suffix_list:
        if word.endswith(suffix):
            stem = re.sub(suffix+'$','', word)
            stem_list.append((word, stem, suffix))
    return stem_list;


if __name__ == "__main__":
    text = """
    Soundex est un algorithme phonétique d'indexation de noms par
leur prononciation en anglais britannique"""
    list_word = text.split(' ')
    for word in  list_word:
        print stemming(word, SUFFIX_LIST)
```

En utilisant la fonction « stemming » :

- Segmenter en mots (tokenize) un texte lu à partir d'un fichier texte.
- Lemmatiser (stemming) les mots en ces mots d'origine.
- Enrichir la liste des suffixes afin de lemmatiser les mots de votre texte.
- Indexer les mots par leurs lemmes.

Série TP n°4

Chapitre 1: niveau lexical

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

Le codage

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

import re
# la liste des suffixes
SUFFIX_LIST=['ique', 'ation', 'tion', 'é', 'er','eur', 'ien']

def stemming(word, suffix_list):
    """ stem a word"""
    stem_list =[]
    for suffix in suffix_list:
        if word.endswith(suffix):
            stem = re.sub(suffix+'$','', word)
            stem_list.append((word, stem, suffix))
    return stem_list;


if __name__ == "__main__":
    text = """
    Soundex est un algorithme phonétique d'indexation de noms par
leur prononciation en anglais britannique"""
    list_word = text.split(' ')
    for word in  list_word:
        print stemming(word, SUFFIX_LIST)
```

En utilisant la fonction « stemming » :

- Segmenter en mots (tokenize) un texte lu à partir d'un fichier texte.
- Lemmatiser (stemming) les mots en ces mots d'origine.
- Enrichir la liste des suffixes afin de lemmatiser les mots de votre texte.
- Indexer les mots par leurs lemmes.