|  |  |
| --- | --- |
|  | TP1 NLP |
|  |  |
| 14/10/2017 | Pierre VALENTIN |
|  | [Tapez le résumé du document ici. Il s'agit généralement d'une courte synthèse du document. Tapez le résumé du document ici. Il s'agit généralement d'une courte synthèse du document.] |

Table des matières

[Introduction 2](#_Toc495771140)

[Tâche 1 – Extraire des informations du world factbook 2](#_Toc495771141)

[Hymne National 2](#_Toc495771142)

[Pourcentage d’alphabétisation 2](#_Toc495771143)

TP1 NLP

Pierre VALENTIN

# Introduction

Dans ce document, je vais préciser comment j'ai extrait de l'information et réaliser de la détection de langue automatique sur de fichiers texte. Je vais ici détailler l'application des techniques vus en cour sur les expressions régulières et les modelé nGram

# Tâche 1 – Extraire des informations du world factbook

Dans cette partie, nous devions extraire les informations des pays demandés depuis le world factbook.

Je vais dans un premier temps détailler mes expressions régulières et enfin finir par revenir sur la précision de mon code. Vous trouverez mon code à cette adresse \url{github.com/Focom/NLPWork1} ou attaché à ce document.

Dans cette partie, rendez-vous dans le répertoire Part1 pour voir le code. Toutes les regex se trouve dans info.py

## Hymne National

J'ai pu identifier une ligne spécifique proche de l'information. Je parcours le fichier et essaye de faire matcher cette regex :

pattern\_all\_file = re.compile("<span class=\"category\_data\" style=\"font-weight:normal; vertical-align:bottom;\">\"")

Lorsque on a une correspondance, il n’y a plus qu’à extraire l’information. Deux cas sont possible, soit l’hymne est déjà en anglais dans ce cas on prend le texte entre les balises. Sinon on doit prendre le texte situé dans les parenthèses.

On essaye les deux regex et on revoit celle qui donne une occurrence. Ces deux ligne s’en charge :

if(result\_non\_english):

return result\_non\_english.group(4)

if(result\_english):

return result\_english.group(2)

## Pourcentage d’alphabétisation

Ici je cherche la ligne qui contient « age 15 and over »

pattern\_all\_file = re.compile("age 15 and over")