BONDON Thibault
GREAU Alexandre

TD - Information Extraction from Wikipedia pages

EXERCICE 1 :

- Nom & prénom

Comme dans l'exemple de Catherine Deneuve dans l'énoncé nous avons annoté toutes les occurrences de l'entité dans son article.
Pour cela nous considérons le titre comme le nom et prénom de l'entité, puis nous recherchons les occurrences à l'aide d'un scanner qui lira le texte mot à mot. L'analyseur reconnait ainsi bien les noms et prénoms de l'entité, cependant pour qu'une séquence "Catherine Deneuve" ne soit pas coupée en deux nous ajoutons une close lors de la détection du prénom afin de savoir si le prochain mot est le nom de l'entité ou pas.

- Pronoms

Le premier problème est de savoir le sexe de l'entité, doit on chercher un "he", un "she ? nous avons remarqué que le premier he/she rencontré fait référence de façon correcte a l'entité. A partir de cette hypothèse il nous fallait détecter le premier he/she et chercher toute ses occurrences dans le texte. Dans le cas des villes il n'y a pas de pronoms.

EXERCICE 2 :

La date de naissance des entités, si elle existe, est la première date suivant la première parenthèse ouvrante. Celle-ci se trouvant toujours dans la première phrase nous avons split le texte au niveau des parenthèse ouvrantes et scannons les String obtenues, la première date trouvée interrompt la recherche. Dans notre programme, un jour est un [0-9]?[0-9], une année [0-9]{4}, pour le mois, l'expression [A-Za-z]{3,9} n'étant pas assez restrictive, nous avons énuméré les différents mois.
Nous avons ensuite construit la balise à l'aide d'un String Builder.

EXERCICE 3 :

Le type des entités se trouve à la fin de la première phrase. Nous splittons donc le texte une première fois sur le "." pour obtenir la première phrase mais étant donné qu'il peut arriver que des points se trouve dans la partie sur la prononciation de l'entité nous effectuons un split préliminaire sur la parenthèse fermante afin d'obtenir la fin de la première phrase où se trouve ce qui nous intéresse. Ensuite nous recherchons un mot clé tels que is ou was car ce qui suit correspond au type que nous recherchons.