

Big Data Analytics

Prétraitements

Des sources variées :

- Wikipedia
- Articles de journaux
- Littérature
- User Generated Content
 - Blogs
 - Commentaires
 - Réseaux sociaux

Une source \Rightarrow un “web scraper”

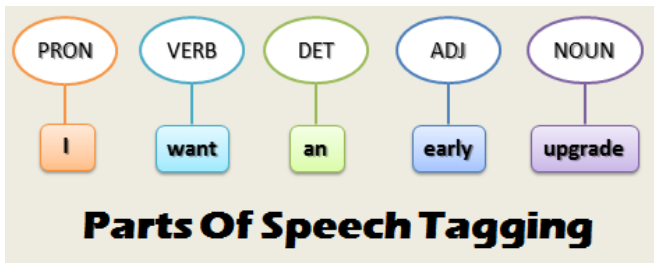
Séparer une chaîne de caractères en token n'est pas trivial :

Le Dr. Pond élève des poules. L'éleveur les sur-exploite.

(en phrases ou en mots)

Prétraitements : POS-Tagging

Étiquetage Morpho-Syntaxique



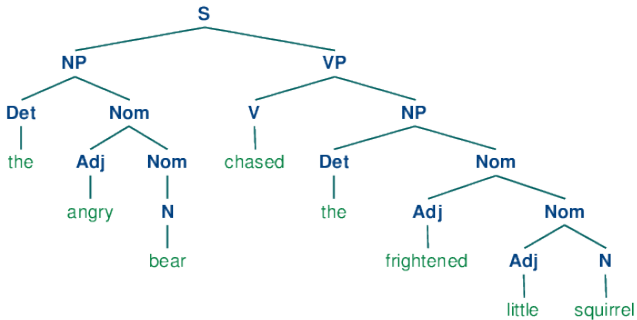
Reconnaissance d'Entités Nommées



Figure 1: An example of NER application on an example text

Prétraitements : Parsing tree

Arbre Syntaxique



Prétraitements : Lemmatisation

Exprimer les mots sous leur forme canonique :

jouant ⇒ jouer

ont été jouées ⇒ jouer

étoiles ⇒ étoile

claires ⇒ clair

noire ⇒ noir

...

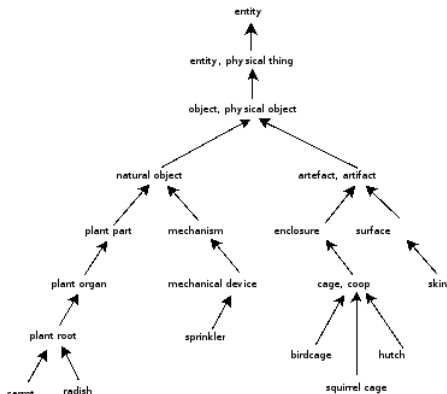


Figure 1. "is a" relation example

Projet sur le français : WOLF (Wordnet Libre du Français)

Outils : DBpedia

