Big Data Analytics

Prétraitements

Prétraitements : Collecte

Des sources variées :

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

Une source $\Rightarrow$ un “web scraper”

Prétraitements : Tokenisation

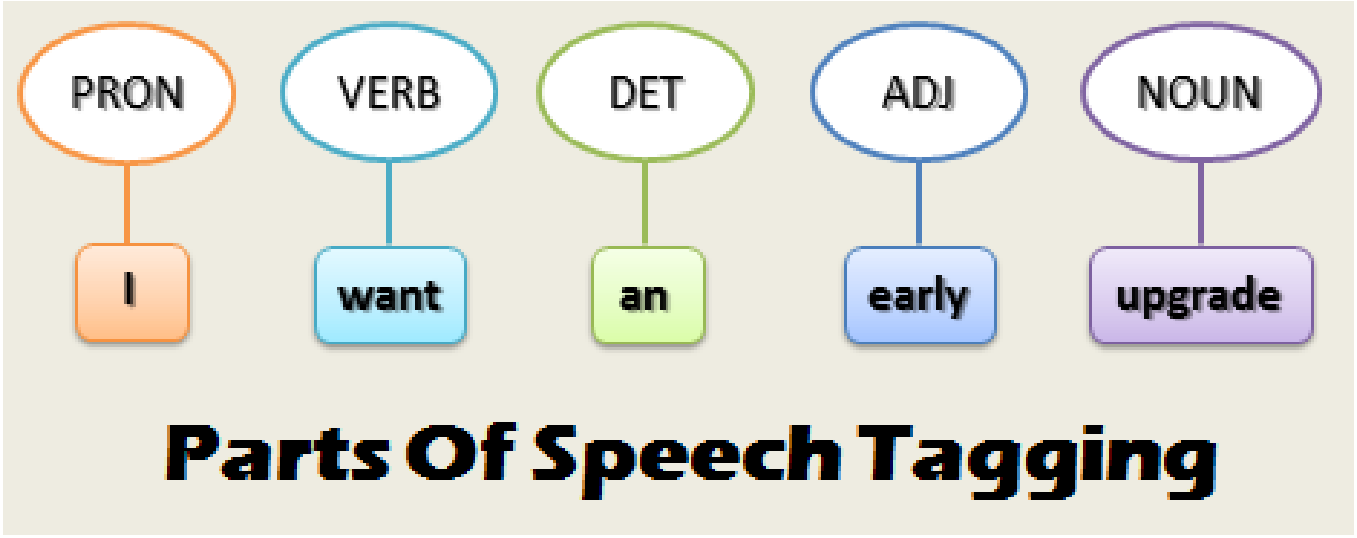
Séparer une chaine de caractères en token n'est pas trivial :

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

(en phrases ou en mots)

Prétraitements : POS-Tagging

Étiquetage Morpho-Syntaxique



Prétraitements : NER

Reconnaissance d'Entités Nommées

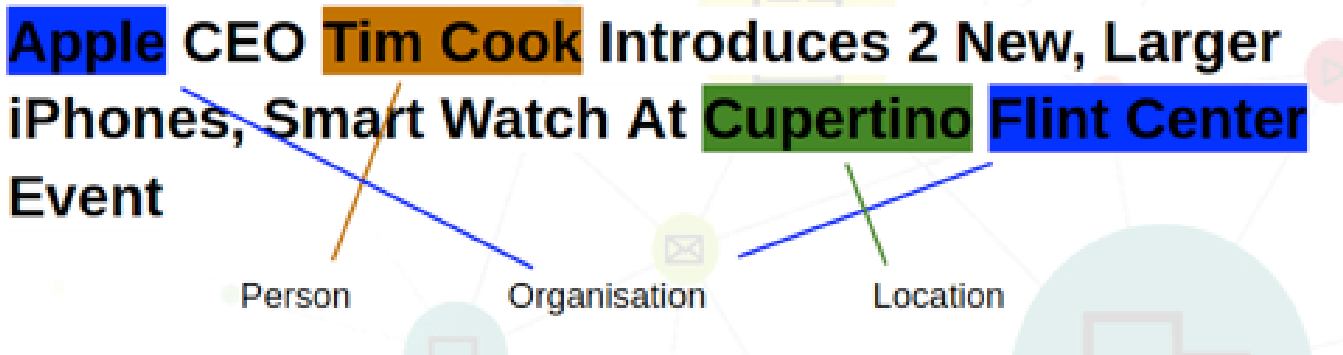
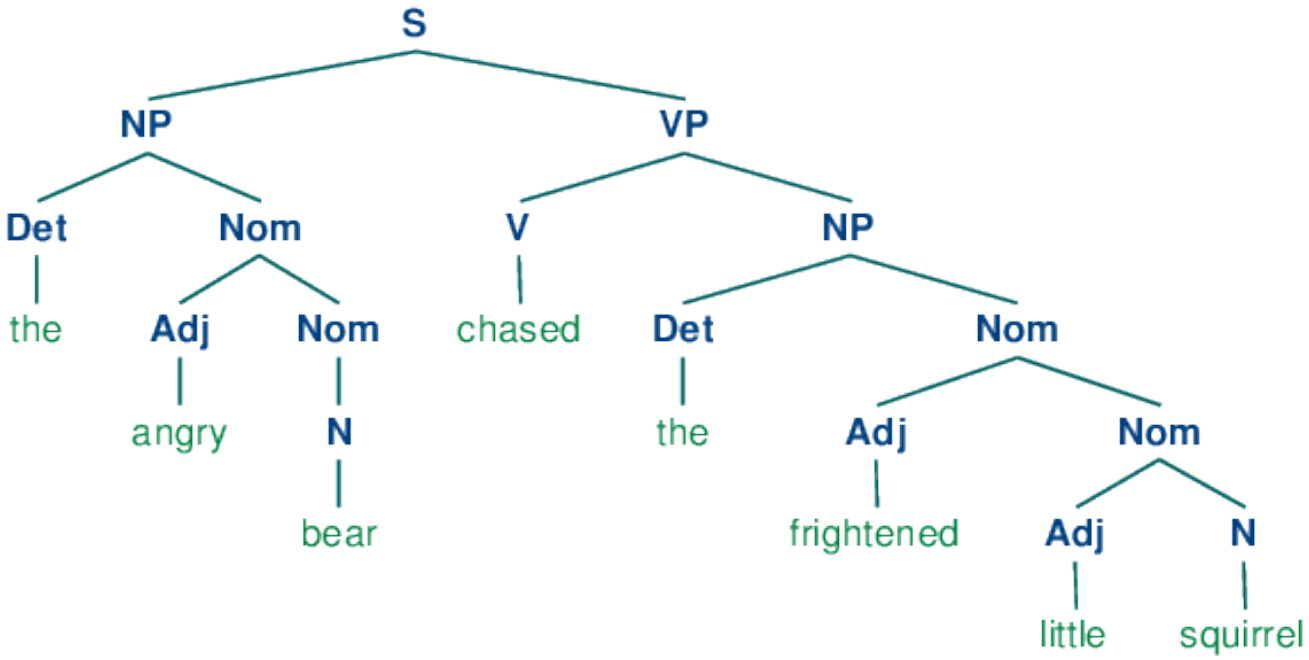


Figure 1: An example of NER application on an example text

Prétraitements : Parsing tree

Arbre Syntaxique



Prétraitements : Lemmatisation

Exprimer les mots sous leur forme canonique :

jouant $\Rightarrow$ jouer
ont été jouées $\Rightarrow$ jouer
étoiles $\Rightarrow$ étoile
claires $\Rightarrow$ clair
noire $\Rightarrow$ noir
...

Outlis : WordNet

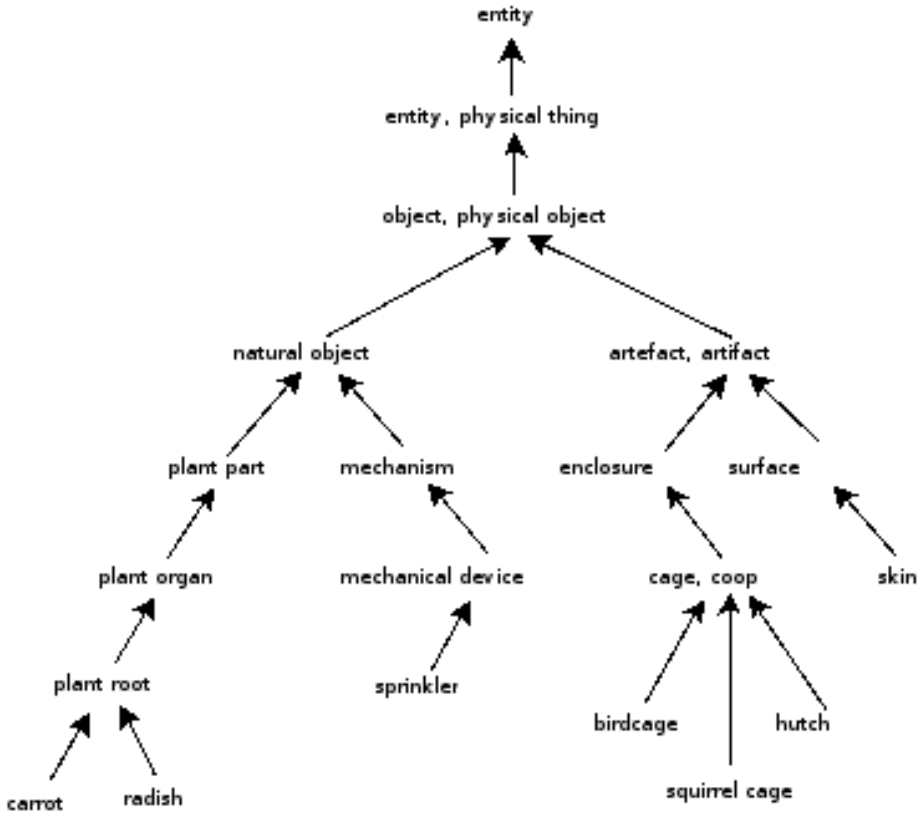
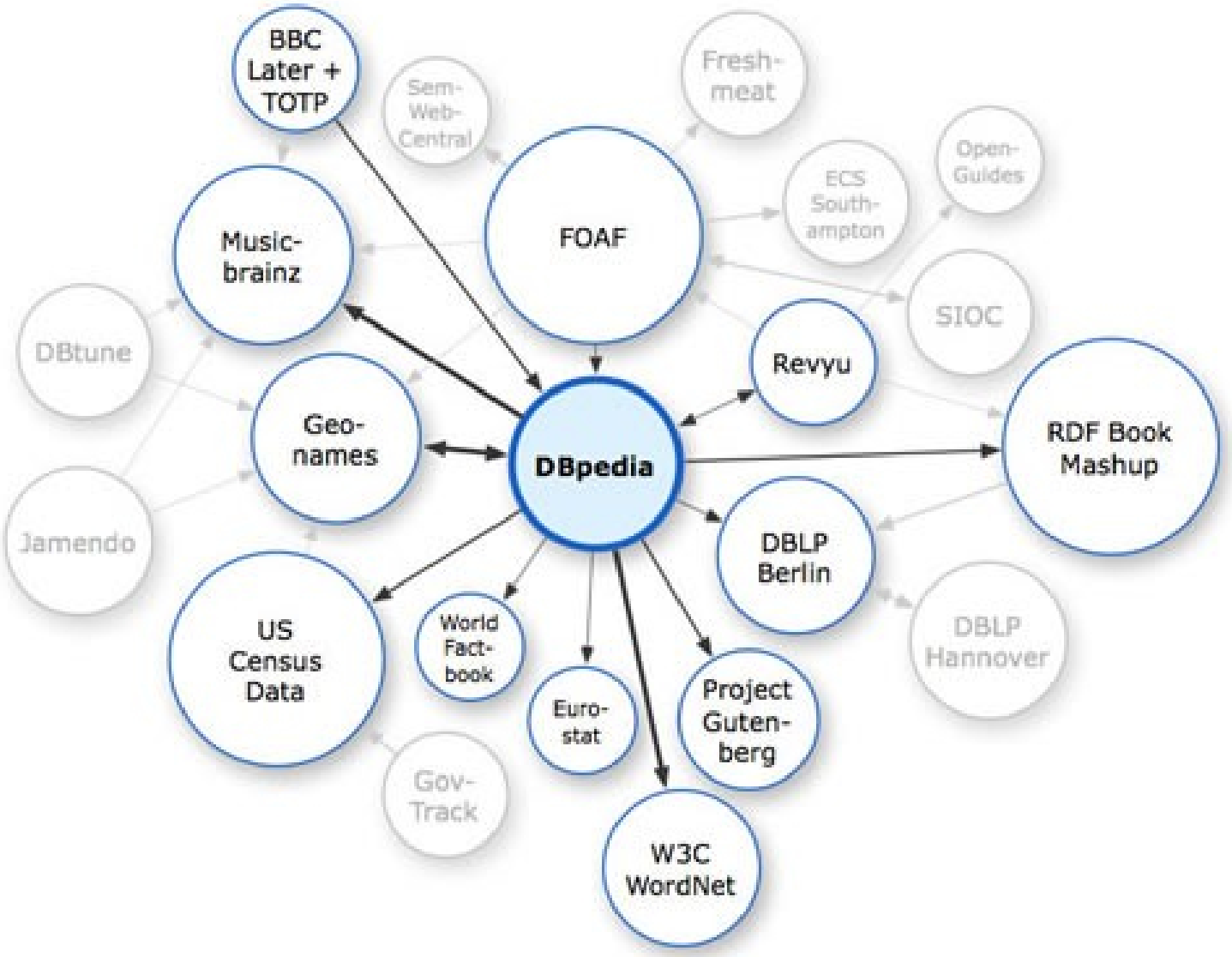


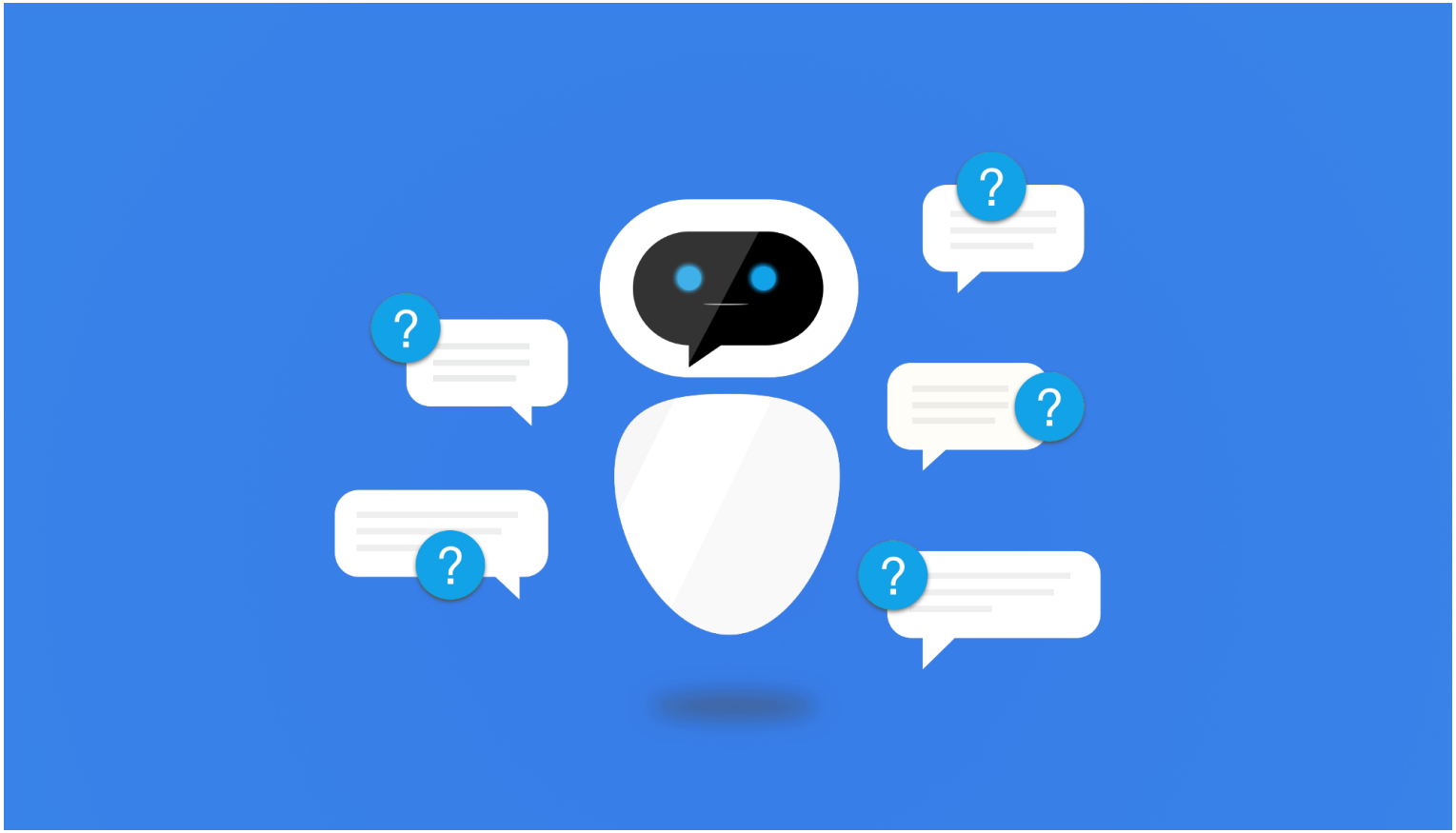
Figure 1. "is a" relation example

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

Outils : DBpedia



Outils : DBpedia



Outils : DBpedia