Extraction automatique de relations sémantiques dans Wikipédia

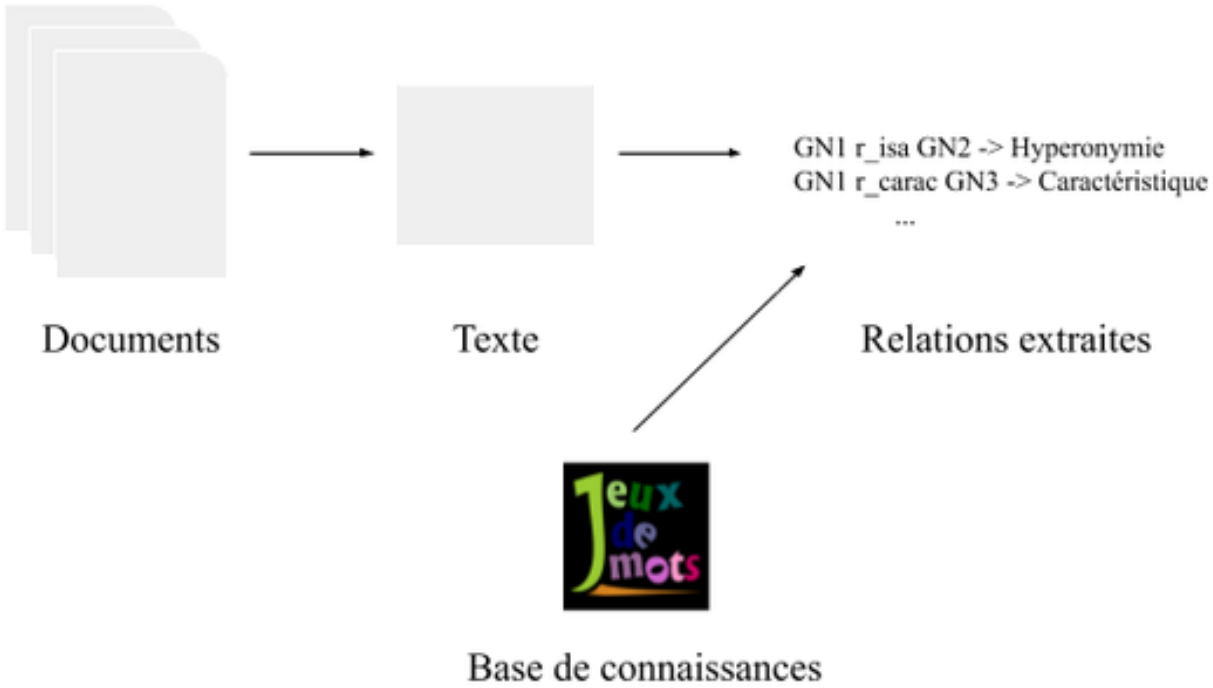
Miassa ABBOUTE
Thierno BARRY
Karim DAHDOUH

DEPARTEMENT INFORMATIQUE
DE LA FACULTE DES SCIENCES

Mai 2020

Sommaire

Problématique



Objectif

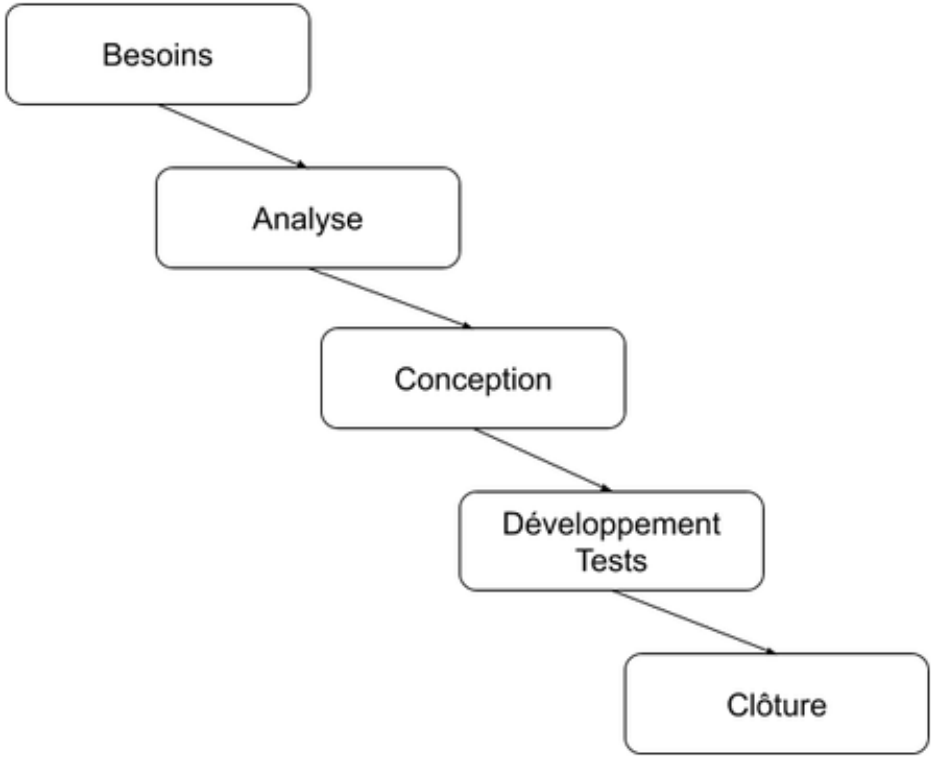
Extraire de façon automatique des relations sémantiques à partir d'articles médicaux issus de Wikipédia

Plan de travail

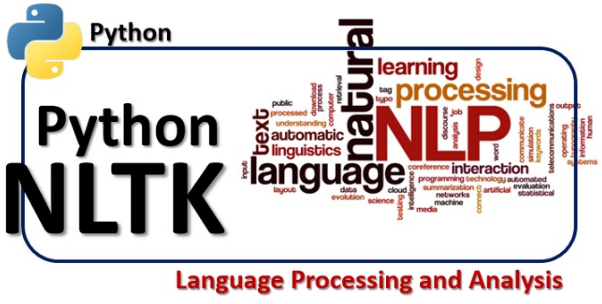
		Name	Perce...	Duration	Start	Finish	Resource Names
1		**⊟Projet TER**	**98%**	**69.25 da...**	**2/5/20, 8:00 AM**	**5/12/20, 10:00 AM**	
2	✔	**⊟Etude des besoins**	**100%**	**6.25 days**	**2/5/20, 8:00 AM**	**2/13/20, 10:00 AM**	
3	✔	Accord sur les objectifs	100%	3 days	2/5/20, 8:00 AM	2/7/20, 5:00 PM	Miassa Abboute;Encad...
4	✔	Spécification des besoins	100%	2 days	2/10/20, 8:00 AM	2/11/20, 5:00 PM	Miassa Abboute;Encad...
5	✔	Définition des contraintes	100%	1.25 days	2/12/20, 8:00 AM	2/13/20, 10:00 AM	Miassa Abboute;Encad...
6	✔	**⊟Analyse du projet**	**100%**	**10.75 days**	**2/13/20, 10:00 AM**	**2/27/20, 5:00 PM**	
7	✔	Recherche de travaux existants	100%	5 days	2/13/20, 10:00 AM	2/20/20, 10:00 AM	Miassa Abboute
8	✔	Etude de relations sémantiques	100%	4.125 days	2/13/20, 12:00 PM	2/19/20, 2:00 PM	Zainoul Barry
9	✔	Exploration de la base de connaissance	100%	5 days	2/21/20, 8:00 AM	2/27/20, 5:00 PM	Karim Dahdouh
10	✔	**⊟Conception**	**100%**	**10.125 d...**	**2/28/20, 8:00 AM**	**3/13/20, 9:00 AM**	
11	✔	Identification de la gouvernance	100%	2 days?	2/28/20, 8:00 AM	3/2/20, 5:00 PM	Encadrant
12	✔	Listing des tâches	100%	1.5 days?	3/3/20, 8:00 AM	3/4/20, 1:00 PM	
13	✔	Identification des moyens matériels et l...	100%	3 days?	3/3/20, 9:00 AM	3/6/20, 9:00 AM	Miassa Abboute;Encad...
14	✔	Choix du langage de développement	100%	2.625 da...	3/3/20, 9:00 AM	3/5/20, 3:00 PM	Zainoul Barry;Karim D...
15	✔	Aclimatation des moyens	100%	5 days	3/6/20, 9:00 AM	3/13/20, 9:00 AM	Miassa Abboute
16	✔	**⊟Développement et Tests**	**100%**	**30.875 d...**	**3/13/20, 9:00 AM**	**4/24/20, 5:00 PM**	
17	✔	Proposition d'une première version de l'...	100%	5 days	3/13/20, 9:00 AM	3/20/20, 9:00 AM	Zainoul Barry;Karim D...
18	✔	Création fichier de règles	100%	3 days?	3/20/20, 9:00 AM	3/25/20, 9:00 AM	Miassa Abboute
19	✔	Implémentation + Test de la première v...	100%	2 days	3/25/20, 9:00 AM	3/27/20, 9:00 AM	Zainoul Barry
20	✔	Création d'une base de données lexico-s...	100%	2 days	3/25/20, 9:00 AM	3/27/20, 9:00 AM	Karim Dahdouh
21	✔	Retour de l'encadrant + Recueil des su...	100%	6 days	3/30/20, 8:00 AM	4/6/20, 5:00 PM	Miassa Abboute;Encad...
22	✔	Améliorations de l'algorithme + Tests	100%	2.5 days?	3/30/20, 8:00 AM	4/1/20, 1:00 PM	Karim Dahdouh
23	✔	Modification et remplissage du fichier d...	100%	3 days	4/3/20, 8:00 AM	4/7/20, 5:00 PM	Miassa Abboute
24	✔	Remplissage de la base de données lexi...	100%	7 days	4/9/20, 8:00 AM	4/17/20, 5:00 PM	Zainoul Barry;Karim D...
25	✔	Echange avec une doctorante sur le sujet	100%	5 days	4/20/20, 8:00 AM	4/24/20, 5:00 PM	Miassa Abboute
26		**⊟Clôture**	**89%**	**11.25 days**	**4/27/20, 8:00 AM**	**5/12/20, 10:00 AM**	
27	✔	Rédaction rapport final	100%	9.125 days	4/27/20, 8:00 AM	5/8/20, 9:00 AM	Miassa Abboute;Zaino...
28		Préparation slides présentation	0%	2.125 days	5/8/20, 9:00 AM	5/12/20, 10:00 AM	Miassa Abboute;Zaino...

Projet TER - Organisation

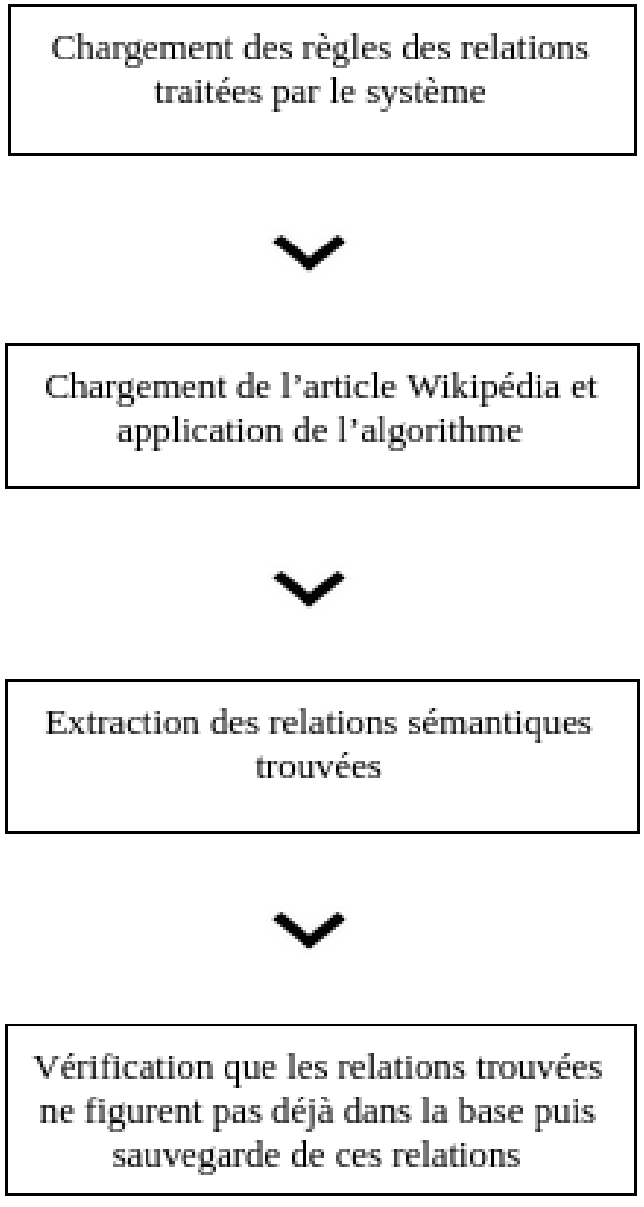
Plan de travail



Outils utilisés



Méthode d'extraction adoptée



Relations sémantiques et règles d'extraction

- **r_isa**

```
nc1 être det nc2 punc => nc1 r_isa nc2
det nc1 être det nc2 => nc1 r_isa nc2
det nc1 être det nc2 punc => nc1 r_isa nc2
```

- **r_own**

nc1 avoir det nc2 punc => nc1 r_own nc2

det nc1 avoir det nc2 => nc1 r_own nc2

det nc1 avoir det nc2 punc => nc1 r_own nc2

- **r_carac**

det nc1 p det nc2 être adj => nc1 r_carac adj

det nc1 p det nc2 être adj punc => nc1 r_carac adj

det nc1 p det nc2 être adj => nc1 r_carac adj

Implémentation

```
import nltk
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.tag.stanford import StanfordPOSTagger
root_path="nltkTools"
pos_tagg = StanfordPOSTagger(root_path + "/french.tagger", root_path + "/stanford-postagger.ja
r",encoding='utf8') #instance de la classe StanfordPOSTagger en UTF-8
```

```
for l in lines:

    l = l.strip()
    if len(l) > 10:
        rl = word_tokenize(l)
        rel =rl[-3]+" "+rl[-2]+" "+rl[-1]
        rel =rel.replace("nc1",ncs[0])
        rel =rel.replace("nc2",ncs[1])
        if len(adj) !=0 :
            rel =rel.replace("adj",adj[0])
        if len(ncs) >2:
            rel =rel.replace("nc3",ncs[2])
        relation = rel


    l = l.replace(" ","")
    if len(l) != 0:
        regle=l
        token = word_tokenize(regle)
        r = token[0]
        r = r.replace("=","")
        if(r in tag):
            if(r_carac in relation):
                    filer_carac.write(relation+"\n")
                    resultat = relation
            if(r_isa in relation):
                    filer_isa.write(relation+"\n")
                    resultat = relation
            if(r_own in relation):
                    filer_own.write(relation+"\n")
                    resultat = relation
```

Exemple

- Lézard possède queue $\Rightarrow$ Lézar *r_haspart* queue
- Etudiant possède ordinateur $\Rightarrow$ Etudiant *r_own* ordinateur

Test et démonstration (1/2)

- Texte étudié

> *"Le cancer est une maladie provoquée par la transformation de cellules qui deviennent anormales et prolifèrent de façon excessive. Ces cellules déréglées finissent parfois par former une masse qu'on appelle tumeur maligne. Les cellules cancéreuses ont tendance à envahir les tissus voisins et à se détacher de la tumeur initiale. Elles migrent alors par les vaisseaux sanguins et les vaisseaux lymphatiques pour aller former une autre tumeur (métastase)."*

- Résultat obtenu

$$\text{cancer } r\_\text{isa } \text{maladie}.$$

Test et démonstration (2/2)

https://github.com/TAZBY/TER

Conclusion

- Objectif
- État actuel
  - Cahier des charges respecté
- Perspectives
  - Enrichissement du fichier de règles
  - Gestion de nouvelles relations sémantiques
- Acquis

Merci ...

... de votre attention