**Plan du rapport :** 5 à 7 pages

1) Introduction :

* Problématique : l’objectif est de comparer/expérimenter deux framework d’analyse linguistique : l’un basé sur des règles (lima), l’autre sur des statistiques (standford core mlp).
* Expérimenter deux plateformes d’analyse linguistique
* Evaluer et Comparer

2) Etat de l’art :

* Décrire brièvement les différentes approches d’analyse linguistique existantes (½ page environ) (regarder des articles scientifiques qui décrivent lima et core mlp (cf. partie bibliographie à la fin))
* Dire que l’on a choisi lima et standford mlp au lieu des autres parce que ce sont des plateformes libres (licences publiques) alors que les autres sont privés (i.e : commerciales), et en plus ils sont intéressant à étudier car ils ne sont pas basés sur les mêmes méthodes (un sur les règles, l’autre sur les statistiques) et donc il est intéressant de les comparer

3) Description des deux plateformes d’analyse linguistique (brièvement, résumer les articles (regarder abstract, intro, et conclusion des articles scientifiques sur ces deux plateformes))

* CEA LIST LIMA
* Stanford Core NLP

4) Description des expérimentations

* Résultats de chaque plateforme
* Comparaison des résultats (que POStag et entités nommés)

Pour comparer les systèmes, il faut les mettre au format universel

***Comparer par rapport au POS tag :***

Etapes :

1)Transformer la phrase analysée par lima sous la forme “John\_NP ate\_V …”

(ex : le lemme(i.e : forme canonique) John devient John\_NP, le ate devient ate\_V)

2)Créer une table de conversion des étiquettes Lima et Standford en étiquettes universelles

3) Créer un programme pour appliquer 2) à 1) pour avoir des étiquettes universelles après les \_ (ex ; John\_NP devient John\_N)

4) Comparer les performances pour le POS tag (les tag)

**Comparer par rapport aux Entités nommées** (LOCATION, etc.)

Etapes :

Idem, transformer avec \_ (ex : Boca Raton\_LOCATION)

Attention : utiliser un caractère entre Boca et Raton pour éviter l’ambiguïté du

-) (ex : &, ABAABA, comme on veut quoi)

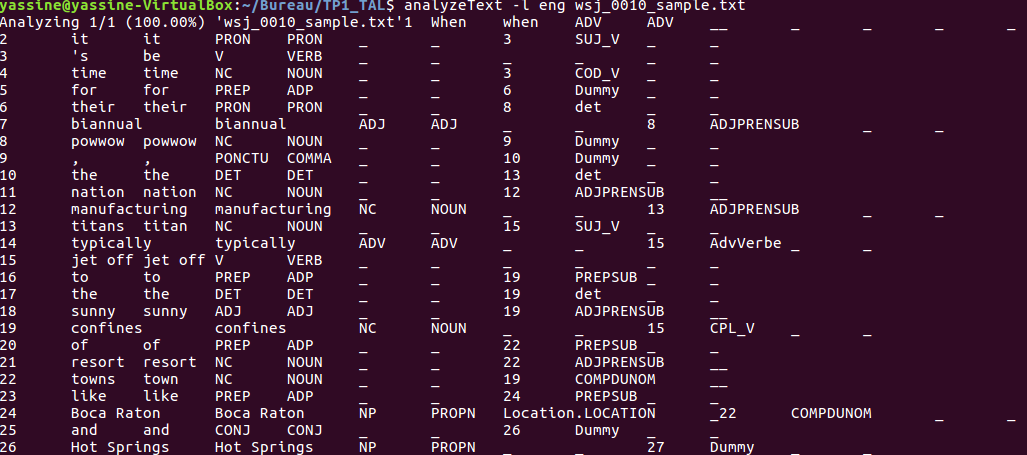
Dans le cas des entitées nommées, les étiquettes universelles sont appelés des modèles.

* Analyse des résultats (donner des exemples)

5) Conclusion

* Limitations
* Perspectives pour améliorer la performance de ces deux plateformes

Exercice 1

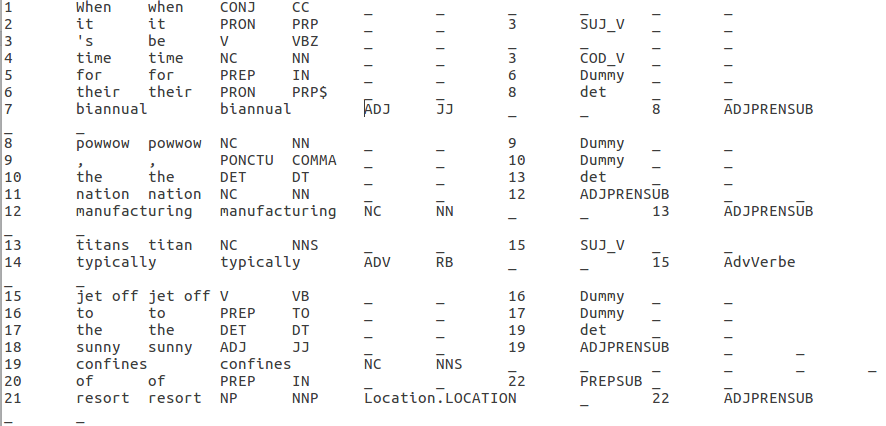


c. Cette analyse ressemble à une analyse à 5 étapes :

* tokenisation : les chaînes de caractères du texte sont bien découpés en mots (cf. première colonne de l’output ci-dessus)
* analyse morphologique : à chaque mot découpé l’analyseur détecte bien si ce mot fait partie de la langue ou non, et lui associe une propriété syntaxique.
* analyse morpho-syntaxique : les mots sont désambiguïsés puisque l’on reconnaît bien si un mot est un pronom, un nom ou un verbe
* analyse syntaxique : on voit bien les relation entre les mots (par exemple un pronom PRON est relié à un SUJ\_V)
* reconnaissance des entités nommées : les dates, lieux, etc. sont bien identifiés. Par exemple, “Boca Raton” est bien détecté comme une Location.LOCATION.

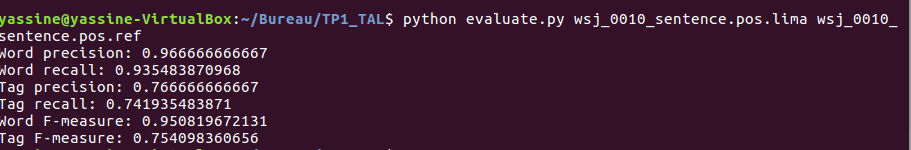
e. Deux nouvelles colonnes sont apparues :

(CC PRP VBZ, etc. et JJ,NN,RB, NNS etc.)



4. Evaluation de l’analyse morpho-syntaxique

1.



III )

Exemple :

John\_NNP ate\_VBD dicious\_JJ pizza\_NN with\_IN friends\_NNS .\_.