LETTRES SORBONNE UNIVERSITÉ

Modèles de Linguistique Computationnelle
CM 6 : segmentation avec python

M1 Langue et Informatique

Crédits: Gaël Lejeune, Karën Fort, Iana Atanassova, Djame Seddah, Eleni Kogkitsidou, Olga Seminck

Alice Millour prenom.nom@sorbonne-universite.fr
Yoann Dupont prenom.nom@sorbonne-universite.fr

2020-2021

Sorbonne-Université

Organisation

Séance d'aujourd'hui

- 30 minutes max : travail sur fichier `Erreurs_en_python.ipynb`
- CM 6 : approches de tokénisation / segmentation avec python
- Fin du TD 4 : Implémentation d'un agent conversationnel

Plan

Les erreurs en python

Segmentation

Importance des messages d'erreur

faire des erreurs de python est **normal**

les **messages** d'erreur sont avant tout des **messages**

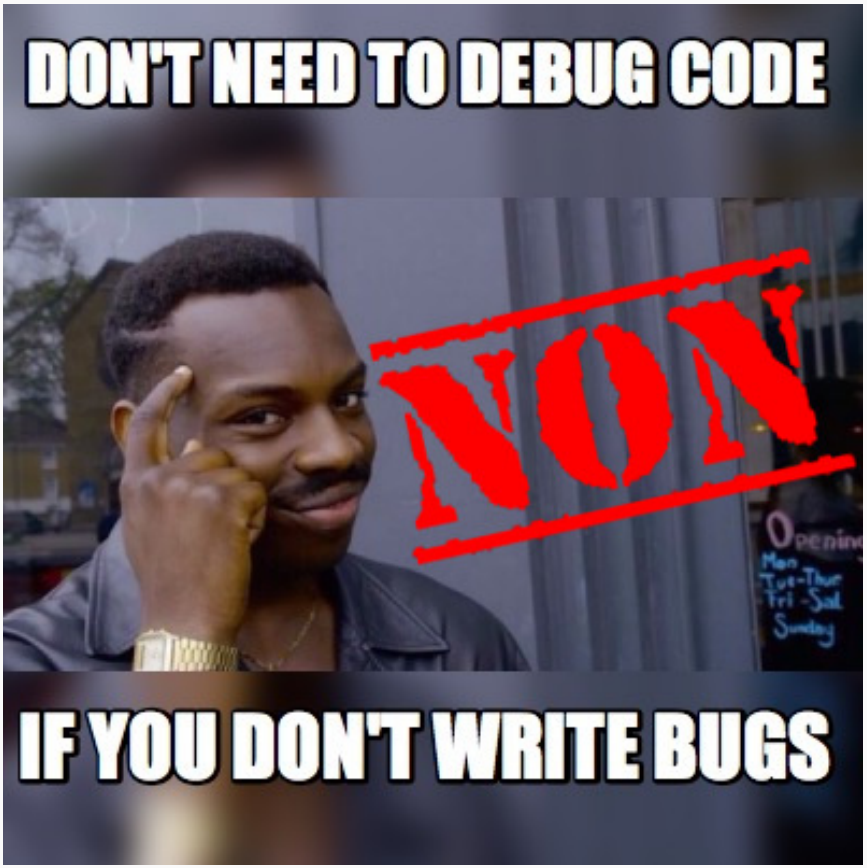
```
#######   CODE 2    ######
def fonction(b) :
    b = [x for x in a if x = 4]
    print(b)
```

```
  File "<ipython-input-5-619d9de4560f>", line 11
    b = [x for x in a if x = 4]
                         ^
SyntaxError: invalid syntax
```

il est donc **indispensable** :

1. de les lire
2. de les comprendre (avec l'aide d'internet si nécessaire)
3. de se servir de l'information qu'il nous donnent pour corriger le code

Stratégies de *debugging*



Plan

Les erreurs en python

Segmentation

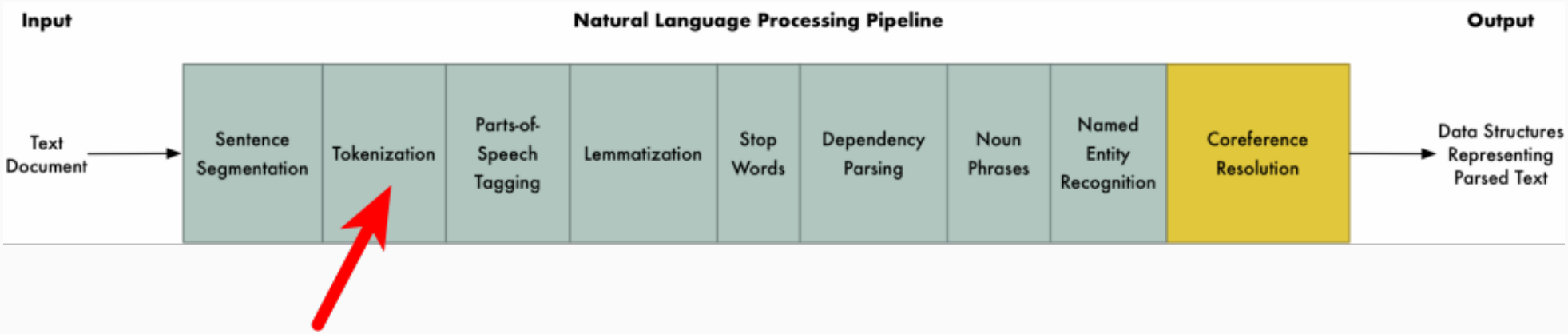
La segmentation, c'est quoi ?

Chercher un grain d'analyse pour une tâche particulière.

Mais qu'est-ce qu'un grain ?

- Le corpus ;
- Le document/le texte ;
- Les titres, les légendes, les paragraphes ;
- Les phrases ;
- les "mots" (on préfèrera tokens)...

Exemple de chaîne de traitement



segmentation : découpage en unités pertinentes

Source : https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e

Segmentations : exercice introductif, les tokens

Marchepied à plate forme sécurisée 2.41m hauteur travail max. 3.16m 5 marches et garde corps fixe PROFORT TUBESCA

- Combien trouvez vous de tokens ?
- Quels problèmes avez-vous rencontré ?
- Quels choix avez-vous dû faire ?

- LIBRE OFFICE en trouve 18 ;
- GEDIT dit 20 ;
- l'outil "wc" de Linux dit 18.

Segmentations : approche simpl(ist)e pour les phrases

1. Segmentation en phrases (texte brut) :
   - séparateurs : ponctuation de n de phrase (".", " ;", " ?", " !") ;
   - autre séparateur : espace/retour à la ligne suivi d'une majuscule ;
   - les "..." forment une ponctuation unique.
2. Segmentation en tokens (tokenization) :
   - séparateurs : espace, tab, apostrophe, ponctuation, les "." ne sont pas des séparateurs (plutôt ". ") ;
   - les tirets si suivis d'un pronom (vient-il), et la séquence -t doit être effacée dans certains cas (envoie-t-elle, césure de n de phrase) ;
   - considérer des transformations locales : Au, du, cet, qu', l', m' ...

Segmentations : les questions

Définition théorique du segment :

- Qu'est-ce qu'une phrase ?
  - Ah ?
  - - ?
- Qu'est-ce qu'un token ?
  - Plate forme, marche-pied, marche pied, marchepied
  - Les marchands du temple : du ! de le ?
  - Time's up
- Ambiguïté des séparateurs :
  - Microsoft.com, 23.5, ... , C.G.T.
  - Aujourd'hui, 9'8, jusqu'à
  - H&M, R&D
  - Sépara-

    teur

Comment segmenter en tokens avec python ?

`nltk` : *Natural Language Toolkit*

= **la boîte à outil python du traitement des langues**

test du code `tokenize.py`

(consulter https://www.nltk.org/py-modindex.html > `nltk.tokenize` > `word_tokenize`)

Améliorer le tokéniseur par défaut

Écrivez un `RegexpTokenizer` à l'aide du canevas `tokenizer.py`. Idéalement, le tokeniseur devrait savoir gérer les règles suivantes :

- les mots comme "garde-manger" ne doivent pas être tokenisés ;
- les nombres comme "0.5" et "0,5" ne doivent pas être tokenisés ;
- les mots comme "l'" et "qu'" ne doivent pas être tokenisés ;
- les mots graphiques respectant le format "majuscule suivie d'un point" ("M.", "X.") ne doivent pas être tokenisés.