Langages de script
Loic Grobol

Documentation du Projet Python POStats Interface

Siyu WANG et Chen SUN
Master TAL M2
INALCO

Objectifs

Les objectifs de notre projet est de réaliser une interface qui, en chargeant un fichier de texte par l'utilisateur, est capable de nous dire en quelle langue est-il écrit ce fichier, de nous montrer tous les tokens et les bigrammes qui apparaîssent le plus dans ce fichier ainsi que la statistique telle que la répartition des mots en fonction de leurs étiquettes morpho-syntaxiques. Il est également capable de donner une présentation graphique de la distribution d'étiquettes POS et le wordcloud.

Nous l'avons conçus pour traiter six langues pour le moment : chinois, anglais, français, allemand, espagnol et italien.

Données de test

Tous les corpus qu'on a utilisé pour tester cette interface sont au format de fichier texte.

- Chinois:

Ce corpus provient de l'équipe NLP de l'Université de Fudan. Il se compose de plusieurs parties en fonction du sujet, tel que l'économie, la politique...etc.
Source:
http://www.nlpir.org/wordpress/2017/10/02/%e6%96%87%e6%9c%ac%e5%88%86%e7%b1%bb%e8%af%ad%e6%96%99%e5%ba%93%ef%bc%88%e5%a4%8d%e6%97%a6%ef%bc%89%e6%b5%8b%e8%af%95%e8%af%ad%e6%96%99/

- Français:

Ce corpus est un ensemble de fichier xml. Nous en avons choisi un vu que les fichiers sont tous de très grande taille. Et nous avons nettoyé le fichier en supprimant toutes les balises.

Source:
https://www.cnrtl.fr/corpus/estrepublicain/

- Anglais, Allemand, Espagnol, Italien:

Les corpus de ces quatre langues sont tous trouvés sur le site Gutenberg projet:
https://www.gutenberg.org/

Méthodologie

Pour la répartition de travail, le script est principalement écrit par Siyu WANG pour développer l'interface, puis Chen SUN a cherché les données de différentes langues pour tester l'interface, enfin nous avons rédigé la documentation ensemble.

❖ Étapes du projet :

- ➢ Environnement virtuel
- ➢ Définir les fonctionnements de l'interface
- ➢ Detection de langue
- ➢ Tokennization selons langues différentes
- ➢ POS tag selons langues différentes
- ➢ Tester l'interface avec des corpus de langues différentes
- ➢ Wordcloud
- ➢ Documentation

❖ Problème résolus:

➢ Le module tkinter

Ce n'était pas facile d'apprendre comment faire une GUI, et nous avons trouvé ce module qui semble être le plus 'débutant-friendly' pour les GUIers. Cependant on a pris beaucoup de temps pour appendre les fonctionnalités du module, et nous avons pu faire les fontions suivantes : faire entrer un fichier par utilisateur, vider la zone d'informations quand un nouveau fichier est chargé, afficher les graphes, afficher les messages de warning quand l'utilisateur n'a pas utilisé correctement (ex. 'no file loaded' ou 'language not supported yet').

➢ Tokennization chinois

Comme il n'y a pas d'espace entre les mots chinois, la tokennization est plus compliquée à réaliser. Nous avons donc utilisé le module 'jieba' qui a été dévéloppé par MIT pour faire la tokennization.

➢ Wordcloud chinois (problème de caractères)

En ce qui concerne la génération du wordcloud du chinois, le module wordcloud n'a pas bien marché comme pour les autres langues, car les caractères chinois ne s'affichent pas sur l'image générée.

D'après les recherches, ce problème a été résolu grace à l'ajout d'une ligne après l'import du module matplotlib et un fichier de font chinois téléchargé : .simfang.ttf

```
import matplotlib
matplotlib.use("TkAgg")
```

```
# if chinese
if language == 'zh':
    font = './simfang.ttf'
    wc = WordCloud(font_path=font,
```

Implémentations

➔ Modules utilisé :

- tkinter (ttk, filedialog, messagebox)
  Pour la construction de l'interface
- re
  Pour la detection de charactère chinois
- io (io.open)
  Pour l'ouverture des fichers du system
- os
  Pour manipuler les fichiers du system
- jieba(pseg pour POS tag)
  Pour tokennization du corpus chinois et 'pseg' pour le POS tagging
- nltk(.corpus stopwords, .tokenize)
  Pour la tokennization et l'utilisation de son corpus stopwords
- langdetect
  Pour la detection des langues
- spacy
  Pour le POS tagging du français, allemend, italien, espagnol
- matplotlib (pyplot)
  Pour la génération des graphes
- wordcloud
  Pour générer les wordcloud
- numpy
  Pour les labels de graphes de bar

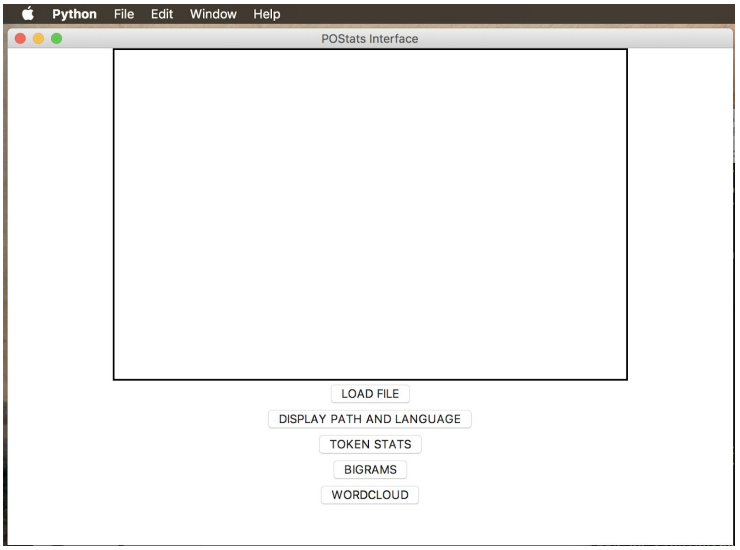
- collections Counter
  Pour compter les top tokens , top bigams...

➔ Langues prises en charge:

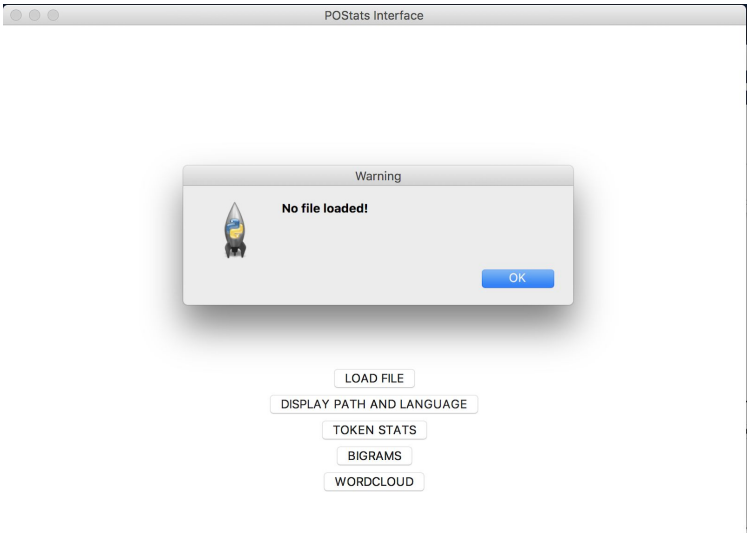
- Français
- Anglais
- Chinois
- Itatlien
- Espagnol
- Allemend

Les résultats

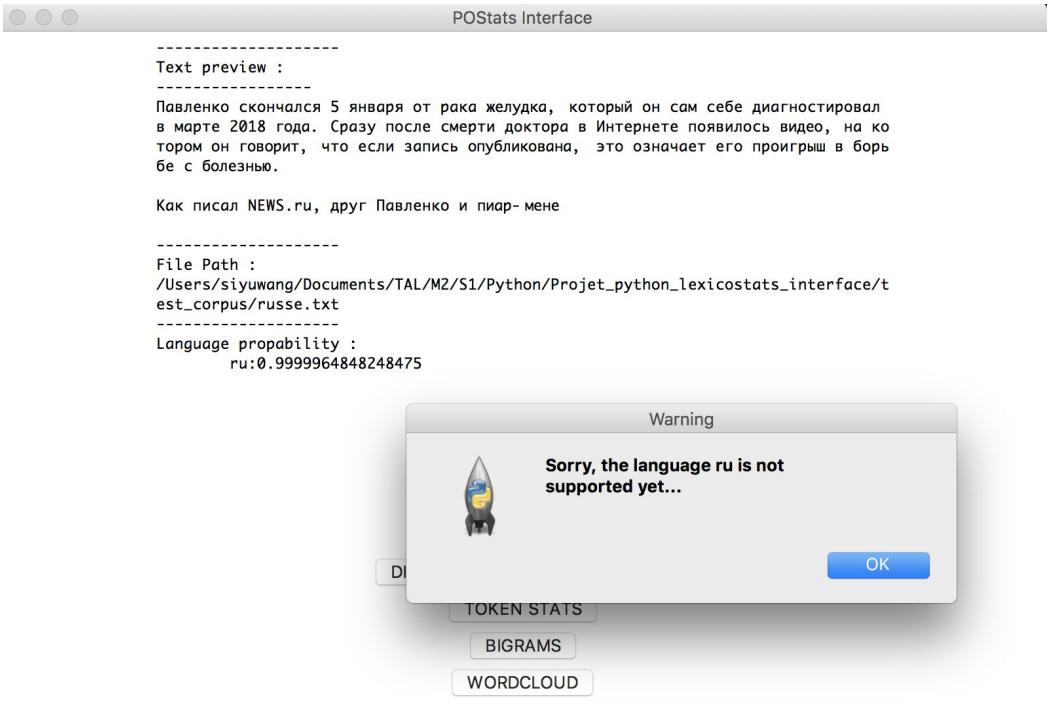
Visualisation :



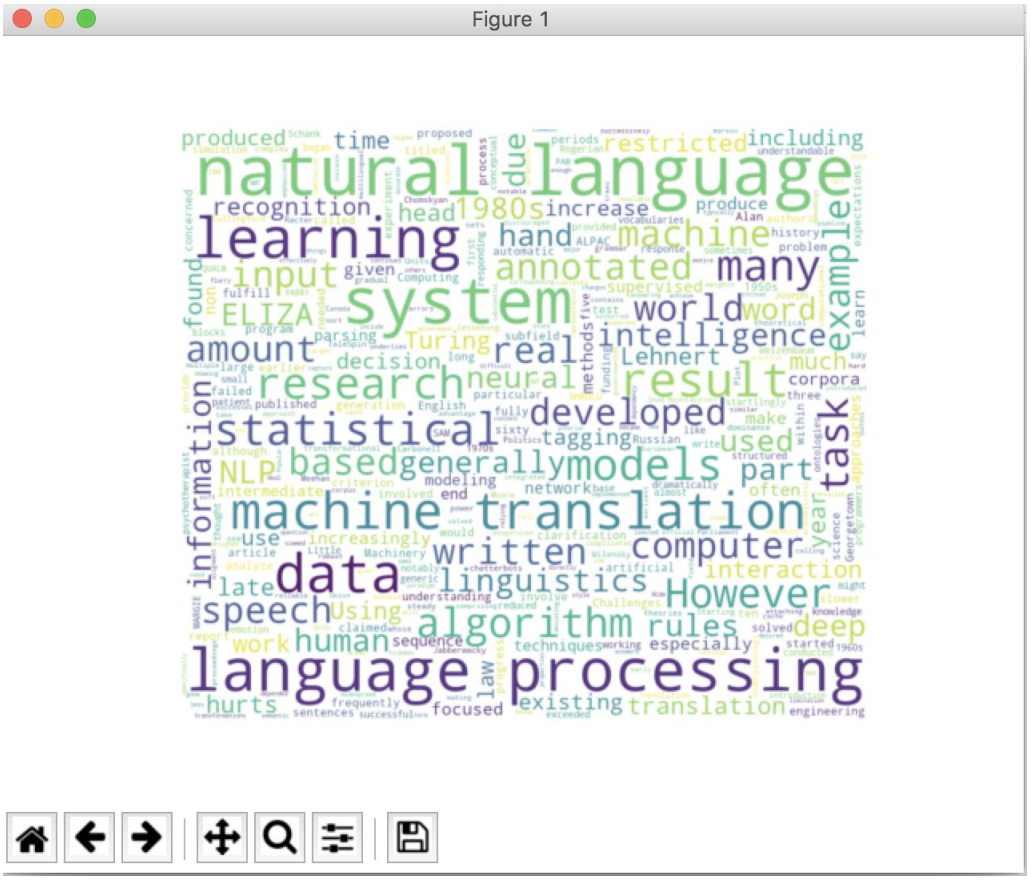
Message de fichier non-importé:



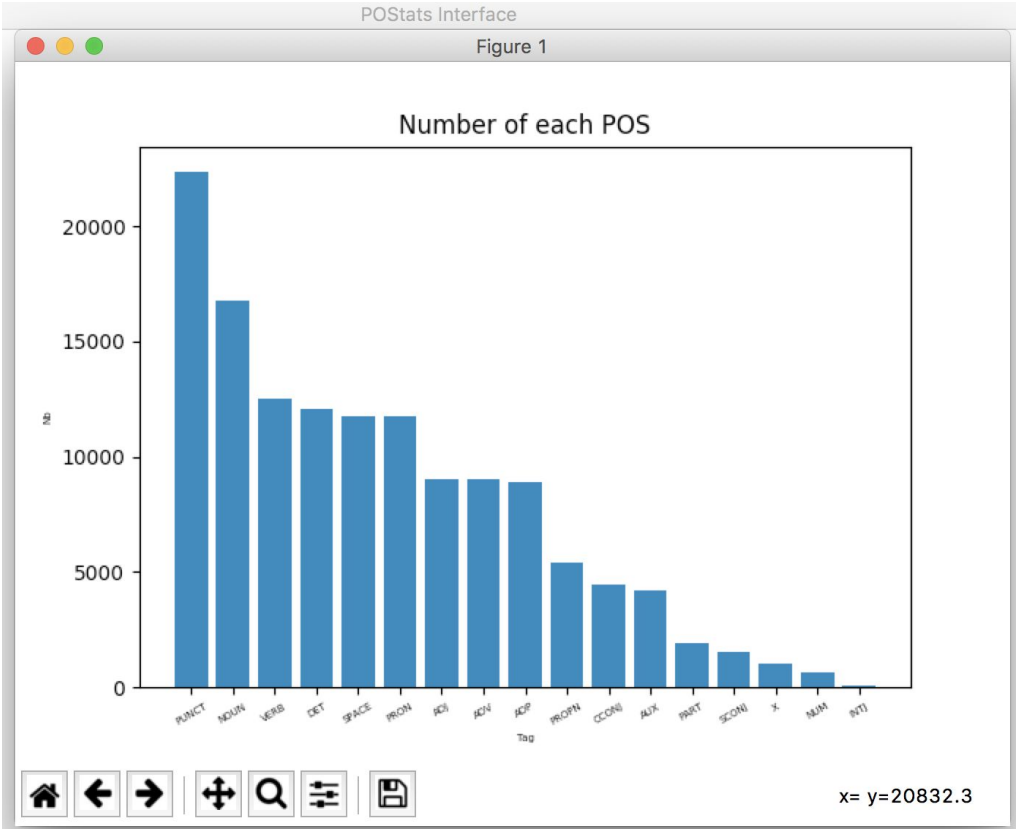
Message de langue non supporté:



Graphique wordcloud avec stopwords pris en compte:



Graphe stats:



Ce que nous aurions aimé faire :

- ➔ Décoration esthetique de l'interface
- ➔ Améliorer le temps d'ouverture du programme