INTERFACE GRAPHIQUE POUR LE TRAITEMENT STATISTIQUE SIMPLE « Sympa »

Par Alexandr Ivanov et Valentin Tinarrage

Voici la documentation pour l'interface graphique « Sympa » qui permet aux linguistes d'effectuer des traitements statistiques simples sur leurs textes directement et rapidement sans devoir avoir recours à plusieurs logiciels. Ce logiciel permet également d'effectuer des travaux sur le texte pour le normaliser ou pour tagger le corpus.

Pour l'utiliser et avoir accès à toutes les fonctionnalités, il faut également installer les bibliothèques suivantes:

- nltk library
- treetaggerwrapper pour Python

Objet du travail:

Le choix du sujet de ce projet vient d'une observation faite lors d'une enquête téléphonique des doctorants en sciences du langage de deux universités françaises , Paris-Sorbonne et Paul-Valéry, qui ont identifié une difficulté dans le domaine de la textométrie :
" les logiciels textométriques sont très performants et proposent des traitements assez complexes, mais souvent ne proposent aucune solution pour les traitements statistiques simples. "

En effet, lorsqu'ils travaillent sur des corpus, les linguistes se heurtent souvent à un manque de possibilités d'effectuer des mesures statistiques simples (compter la diversité/richesse du vocabulaire, compter le nombre de phrases etc). Ils sont souvent obligés d'avoir recours à plusieurs logiciels (excel et worddoc ou calculatrice etc) pour effectuer ces mesures sur un même texte.

L'objet de ce projet est de :

- regrouper quelques traitements statistiques simples dans un même module.
- mettre en place un tagger simple qui permet de visualiser le texte d'une manière claire.
- créer une interface graphique simple et facile à prendre en main pour des utilisateurs non spécialistes en informatique ou qui n'ont pas beaucoup de temps pour se familiariser avec le fonctionnement du logiciel.

Les enjeux que nous avons suivi pour répondre à cet objet ont été les suivants :

- Associer simplicité et fonctionnalité.
- Choisir les statistiques les plus communes pour des linguistes.
- Rendre les explications les plus claires et concises possible.
- Choisir une langue donnée.

Méthodologie:

Pour ce projet, nous avons suivi les étapes suivantes :

1) Identification des problèmes, et choix des solutions à apporter

Travail effectué par : Alexandre, Valentin.

Résultat : créer un logiciel qui associe les traitements statistiques simples avec une interface en PyQt. Le choix de PyQt est conditionné par le fait que Qt est un module libre multiplateforme qui permet d'associer la flexibilité de python avec les fonctionnalités de C++.

2) Répartition des tâches

Travail effectué par : Alexandre, Valentin.

Résultat : L'interface graphique est fait réalisée par Alexandre, les fonctionnalités sont rajoutées par Valentin.
Les deux membres sont libres de proposer des fonctionnalités ou des solutions graphiques supplémentaires n'importe quand pendant du déroulement du projet.

3) Création de l'interface

Travail effectué par : Alexandre

Résultat :

- création d'une fenêtre mainWindow
- ajout des composants graphiques (9 composants) :
    - TextEdit : permet d'éditer le texte
    - Line : permet d'écrire les mots qu'on cherche
    - Label(3 labels) : permet de donner les noms aux fenêtres ou expliquer ce qu'il faut faire(Choisissez vos stats, Choisissez vos traitements et Écrivez votre mot)
    - Button (2 buttons) : permet d'effectuer le traitement sur le texte ou l'effacer (Appliquer, Supprimer)
    - MenuBar : permet de sauvegarder et ouvrir les fichiers.

Difficultés :
Apprentissage de PyQt dès le début.
Difficile de relier les fenêtres les unes entre les autres et avec leurs comportements.
Difficile à dynamiser les contenu graphique

4) Fonctionnalités

Travail effectué par : Valentin, Alexandre

Résultat :

- création du module foncStats
- création des fonctions dans foncStats :
    - compter le nombre de caractères
    - compter le nombre de mots
    - compter le nombre de phrases
    - compter la fréquence d'un pattern choisi par l'utilisateur
    - calculer la longueur moyenne de la phrase
    - calculer la richesse du vocabulaire
    - compter la fréquence d'un mot choisi par l'utilisateur
    - enlever les espaces en trop
    - tagger le texte

Difficultés :
Associer les fonctions aux composants graphiques (il faut que chaque fonctionnalité soit strictement une fonction avec un seul retour str).
Associer le contenu de TextEdit et des Lines (une variable pour la valeur de chaque composant).

4) Test

Travail effectué par : Alexandre, Valentin

Résultat :
Toutes les fonctionnalités fonctionnent sur nos machines respectives.
Il faut que le combobox non utilisé ait la valeur vide.

5) Rédaction du rapport

Travail effectué par : Alexandre, Valentin

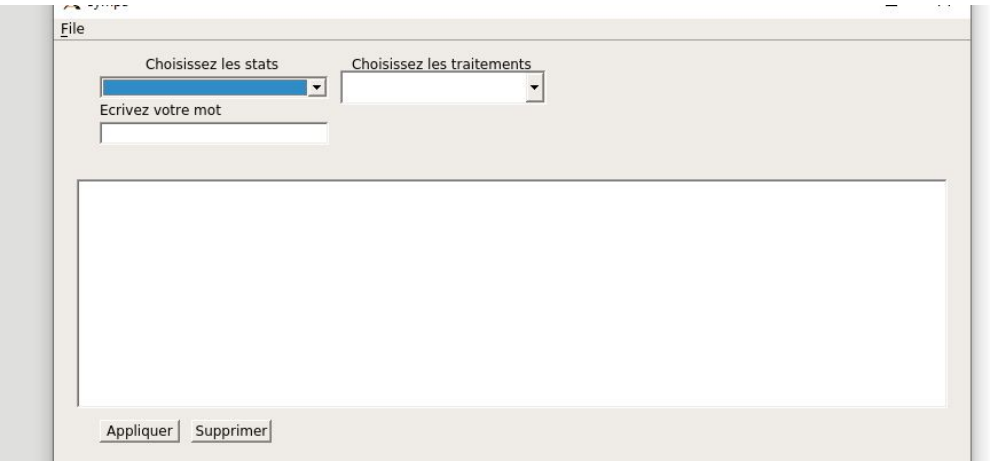
Données employées pour les tests:

Les données utilisées pour tester le script se composent de deux textes en format txt :

- un extrait de *'la vie de Marianne'* (marianne.txt)
- un extrait de *'Le ventre de Paris'* (le_ventre_de_paris.txt).
- Les données de simple saisie du texte par l'utilisateur dans la boîte de dialogue. Les textes en question ont ainsi été modifié manuellement (ajout d'espaces, de mots, etc.) lors du test du logiciel.

Les deux textes sont extraits de Wikisource :
https://fr.wikisource.org/wiki/La_Vie_de_Marianne_(%C3%A9d._Duviquet)
https://fr.wikisource.org/wiki/Le_Ventre_de_Paris

Résultats:



Le logiciel présente une fenêtre principale qui contient:

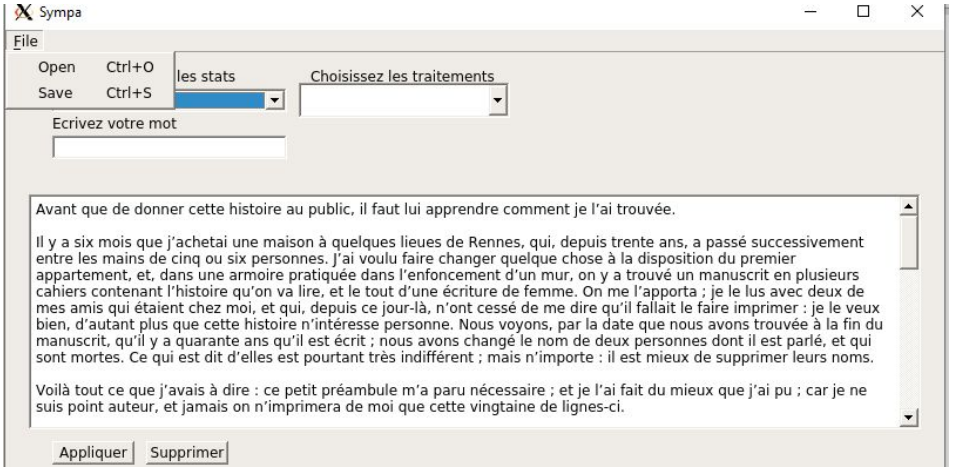
- un bouton File qui permet d'ouvrir des fichiers et de charger leur contenu dans l'éditeur.

- deux barres à déroulant qui proposent respectivement les calculs statistiques et les traitements textuels.

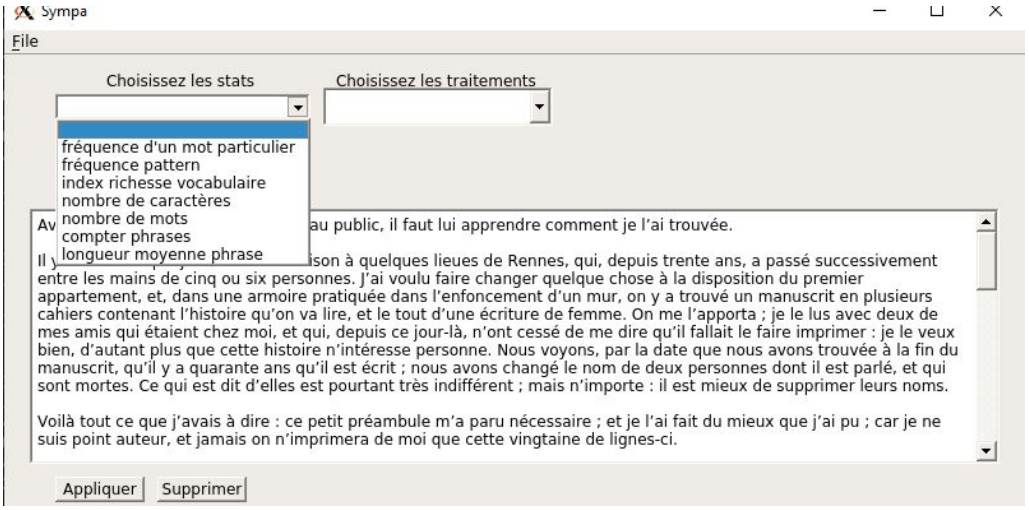
- un éditeur de texte pour lire et interagir avec le texte.

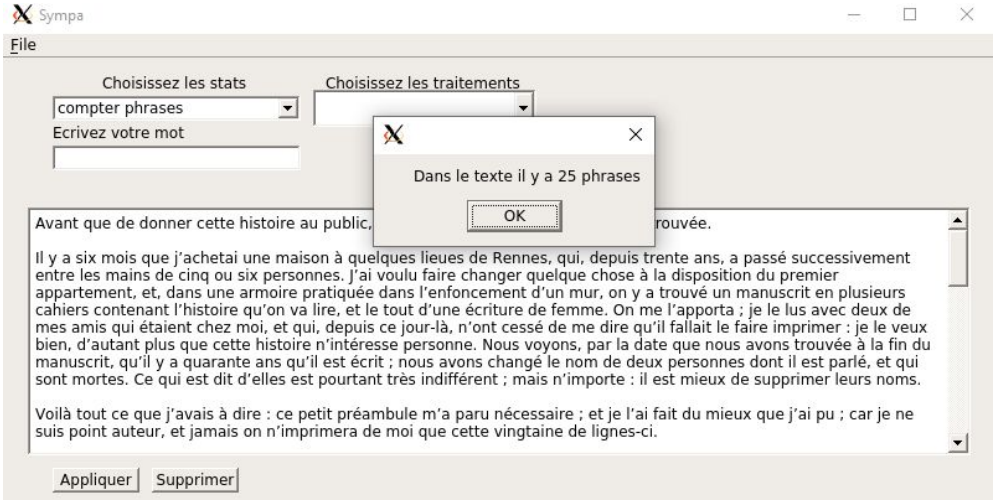
- et deux boutons:
  - Appliquer permet de procéder au traitement sélectionné (il faut veiller à n'en sélectionner qu'un).
  - Supprimer pour vider l'éditeur.

Le bouton File permet d'ouvrir des fichiers et de charger leur contenu dans l'éditeur:



La première barre propose les différents calculs statistiques :





Nous pouvons également tagger le texte :

Le logiciel peut ainsi être utilisé pour des textes variés en format .txt.
Il est instinctif et simple d'utilisation, et peut être maîtrisé en seulement quelques minutes.

Conclusion

Dans le cadre de ce travail, nous avons réalisé une interface graphique légère capable d'effectuer différentes tâches et calculs dans le but de répondre à un besoin des linguistes d'avoir une application simple et rapide à prendre en main.
Nous avons ainsi rempli notre objectif, en réussissant à fournir la plupart des fonctionnalités voulues tout en conservant une application instinctive et simple.

Pour poursuivre ce travail, il serait pertinent d'ajouter des fonctionnalités supplémentaires notamment :

- travailler sur plusieurs textes à la fois.
- effectuer une recherche par patron basée à la fois sur le texte et sur les part of speech.

Cependant, il est important de conserver la facilité d'utilisation que présente ce logiciel, c'est pourquoi ajouter de telles fonctionnalités présente le risque de sur-compliquer l'interface.