L3 Linguistique de Corpus TALA332A & Outils et méthodes de traitement de corpus TALA332B Devoir-Examen du premier semestre

P. Paroubek, LISN-CNRS-U. Paris-Saclay

le 20 décembre 2022

1 Consignes

A rendre au plus tard pour le lundi 09 janvier 2023, 24h par voie électronique en le déposant sur le site :
`https://mycore.core-cloud.net/index.php/s/3j9MbymrY7m3e42`
sous forme d'une archive compressée zip mentionnant votre nom et votre numéro d'étudiant. Notez que cette url n'est accessible qu'en mode téléversement, vous ne pouvez donc pas voir le fichier que vous y déposez ni les autres fichiers ; en cas de doutes sur votre téléversement contactez patrick.paroubek@lisn.upsaclay.fr.

Les résultats attendus sont d'une part un rapport décrivant les programmes (en utilisant le style disponible à l'URL

`https://perso.limsi.fr/pap/inalco/TNML3_2022_2023/template-memoire-inalco-2016-2017.zip`

,le code des programmes réalisés et les données produites (listes de fréquence, texte annoté etc.). Lorque vous utiliserez des bibliothèques Python ou des outils logiciels ou utilisez des algorithmes, méthodes ou approches issues de vos lectures (articles scientifiques, de vulgarisation, page web, blog etc.), vous devez donner leur référence dans la partie bibliographie du rapport. Rappel, si les discussion entre vous concernant le code et les bonnes pratiques sont fructueuses et à encourager, la rédaction du rapport et l'écriture des programmes doit rester personnelle. Dans la mesure du possible, lorsque vous présenterez des tableaux de résultats numériques (listes de fréquence etc.) pour des études comparatives, essayez de fournir aussi une illustration au moyen de graphique pour rendre les éléments saillants plus facilement visibles (par ex. avec une bibliothèque graphique comme `https://matplotlib.org/`.

2 Introduction

Vous allez effectuer une étude de lexicométrie comparative au moyen des outils de programmation python3 vus au premier semestre. Cette étude portera sur trois versions du roman de Jules Verne « Le tour du monde en 80 jours » :

1. une version françaises issue de la boîte à outils Unitex `80jours_unitex_gramlab_v3_2.txt`,
2. une autre version française issue du site de l'ABU `jv80jours_ABU_unformated_iso88591.txt`
3. et une version en anglais issue du projet Gutenberg `103-0_jv80_english_gutenberg_prj.txt`.

Les données (corpus et lexique) sont disponibles à l'url :

`https://perso.limsi.fr/pap/inalco/TNML3_2022_2023/data_devoir_L3S1_22_23.zip`

et ce sujet est disponible à l'url :

`https://perso.limsi.fr/pap/inalco/TNML3_2022_2023/devoir_exam_L3_S1_20221220.pdf`

3 Mise au format des données

Normalisez les contenus de manière à ce que tous les fichiers utilisent le même encodage des caractères en utf-8, et soient organisés de la même façon. C'est à dire dans votre organisation des données, vous séparerez pour les 3 version de la même façon les données (le texte de l'auteur) des meta-données (les informations de source, de date etc.). Décrivez les étapes de traitement que vous appliquerez expliquant pourquoi vous les avez effectuées, en indiquant en particulier comment vous avez traité les éléments de structuration du texte. Les étapes seront réalisées à l'aide de scripts écrits en Python que vous founirez avec les données normalisées.

4 Extraction des distributions

Pour toutes les versions vous extrairez les distributions de caractères, que vous organiserez par ordre décroissant de fréquence. Ensuite, après avoir expliqué quel est votre algorithme de segmentation en unités tokens vous extrairez les distributions de tokens pour les trois versions. Vous expliquerez en particulier comment vous avez traité les marques de ponctuation. Ensuite, vous nalyserez et commenterez les différentes distributions les unes par rapport aux autres, qu'en concluez-vous ? En particulier comparez les distributions de tokens des versions françaises et anglaise.

5 Filtrage des données

Utilisez le contenu du lexique `dimaju-4.1.1_utf8.txt` fourni avec les données, pour extraire des différentes versions les noms propres présents dans le lexique (indiqués par la présence d'une étiquette `SBP`). Effectuez des recherches sur Internet pour analyser et discuter de la couverture du résultat que vous obtenez avec la liste des personnages du Roman. Quel commentaire pouvez-vous faire sur la liste des pays mentionnés dans les versions françaises du roman ? Après traduction de cette liste en anglais, vous regarderez si elle se retrouve à l'identique dans la version anglaise.

6 Annotation

Extrayez à la main 3 phrases dont la taille est supérieure ou égale à 10 mots, que vous annoterez à la main avec des étiquettes morpho-syntaxique en précisant quel liste d'étiquettes vous utilisez, celui de la campagne GRACE vu en cours, ou bien par ex. celui proposé par le projet Universal Dependencies à l'url `https://universaldependencies.org/fr/pos/index.html`. Ecrivez un script python pour annoter ces trois phrases avec les informations du lexique `dimaju-4.1.1_utf8.txt`, puis évaluez à la main votre performance avec la mesure de `précision`. Analysez votre résultat.

7 Annotation II

En utilisant le lexique `dimaju-4.1.1_utf8.txt`, vous étiqueterez une des versions françaises en associant à chaque mot trouvé dans le lexique toutes ses étiquettes. Vous ferez ensuite une étude par fréquence des listes d'étiquettes, commentez sur les plus fréquentes et sur la fréquence de l'étiquette INCONNU qui correspond à un mot que vous n'aurez pas trouvé dans le lexique.