Vérification de la corrélation de Spearman dans un ensemble KSC

Alexandre Roulois (LLF Université Paris Cité/CNRS)

2023-12-05

Sujet (12 points)

Soit deux romans en anglais de veine fantastique, à savoir le *Frankenstein* de Mary Shelley et le *Dracula* de Bram Stoker que vous utiliserez pour constituer un ensemble de corpus de telle manière que le premier sera composé de 100 % du premier texte, le second de 90 % du premier texte et de 10 % du second texte, le troisième de 80 % du premier texte et de 20 % du second texte, et ainsi de suite jusqu'au dernier corpus qui sera lui formé uniquement de 100 % du second texte.

La méthode, connue sous le nom de *Known-Similarity Corpora* (Kilgarriff 2001), ou KSC, permet de dégager des assertions sur la similarité entre les corpus constitués et les textes d'où ils sont issus :

- Le corpus n°1 ressemble davantage au corpus n°2 qu'à tous les autres corpus ;
- le corpus n°2 ressemble davantage au corpus n°3 que le corpus n°1 ;
- le corpus n°3 ressemble davantage au corpus n°4 que les deux premiers corpus ;
- et ainsi de suite…

Votre objectif est de prouver ces assertions en utilisant le coefficient de corrélation des rangs de Spearman, une mesure statistique qui a le double avantage d'être non paramétrique (c'est-à-dire ne supposant a priori aucune loi de probabilité) et indépendant du type de variables. En effet, elle repose sur un calcul de la différence entre les rangs des données :

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Pour calculer le ρ de Spearman entre tous les corpus, vous utiliserez la mesure de fréquence d'occurrences des 500 mots-formes les plus fréquents. Considérons par exemple les corpus A et B ci-dessous pour lesquels nous avons préalablement établi que les 4 mots-formes les plus fréquents dans les textes dont ils sont issus étaient *le*, *mon*, *chat*, *petit* :

Mot-forme	Fréquence corpus A	Fréquence corpus B
le	38	31
mon	15	23
chat	27	18
petit	19	30

Représentons à présent les rangs des mots-formes dans chacun des corpus :

Mot-forme	Rang corpus A	Rang corpus B
le	1	1
mon	4	3
chat	2	4
petit	3	2

Le coefficient de corrélation de Spearman vaut ainsi pour les corpus A et B :

$$\begin{aligned}\rho &= 1 - \frac{6 \cdot ((1-1)^2 + (4-3)^2 + (2-4)^2 + (3-2)^2)}{4 \cdot (4^2 - 1)} \\ &= 1 - \frac{6 \cdot (0 + 1 + 4 + 1)}{4 \cdot 15} \\ &= 1 - \frac{36}{60} \\ &= 0,4\end{aligned}$$

Vous remettrez, par email et pour le 12 janvier 2024 au plus tard, votre code ainsi qu'un fichier CSV répondant à la structure ci-dessous, où les deux premiers champs représentent les numéros des corpus pour lesquels vous avez estimé la similarité et le dernier champ la corrélation de Spearman :

```
0;1;0.998735
0;2;0.975688
0;3;0.834120
```

Vous ne serez pas tant évalué · es sur le résultat que sur la qualité de votre code et la stratégie mise en place dans l'élaboration de votre programme.

Bibliographie

Kilgarriff, Adam. 2001. "Comparing Corpora." Journal Article. *International Journal of Corpus Linguistics* 6 (1): 97–133. https://doi.org/10.1075/ijcl.6.1.05kil.