Frantext intégral

Familiarisation avec le corpus

Allez dans Documentation > Didacticiel > Frantext 1, Présentation

Quel critère de sélection : **genre textuel, période** ?

But de recherche ?

Quelle taille ? Nombre de textes, nombre mots ?

Premier contact avec l'outil

Accédez au corpus intégral. Comme beaucoup d'outils, l'interface de Frantext s'organise en plusieurs étapes : (1) **sélection du corpus de travail**, (2) **recherche dans les textes**, (3) **calcul de fréquence**. Contrairement à d'autres, Frantext utilise aussi beaucoup des **listes de mots**.

Sélectionnez tout le corpus, puis effectuez des recherches dans les textes pour déterminer précisément ce qu'est un « mot » pour Frantext.

Les signes de ponctuation sont-ils des mots ?

La recherche est-elle sensible à la casse ?

« aujourd'hui », « parce que », « pomme de terre »… combien de mots ?

Les majuscules comportent-elles des diacritiques ?

Le corpus utilise-t-il les ligatures ?

Premières recherches

Sélectionnez les textes de Zola (n'oubliez pas de vider préalablement le corpus de travail !).

Recherchez **les flexions du verbe** *faire*, puis *taire*. Quelle est la limite de ce mode de recherche ?

Expressions régulières

N'importe quel mot : ^**.*$**

N'importe quel mot qui contient *blanc* : blanc

Le mot *blanc* exactement (il commence *et* finit par *blanc*) : ^blanc**$**

N'importe quel mot se terminant par -*âtre* : âtre**$**

N'importe quel mot commençant par *anti-* : ^anti

Le mot *blanc*, éventuellement suivi d'un *s* : ^blancs?$

Le mot *blanc*, éventuellement fléchi : ^blanc**(**s|he|hes**)?$** ou bien ^blanc**(**he**)?**s**?$**

… et aussi *blancheur(s)* : ^blanc**(**he|heur**)**s**?$**

Les parenthèses et les barres (...|...) fonctionnent en mode « expression régulière », mais aussi dans les autres modes.

Loi de Zipf

La loi de Zipf est une observation empirique concernant la fréquence des mots dans un texte.

> Dans les années 30, un scientifique de l'université de Harvard, G.K. Zipf, a montré qu'en classant les mots d'un texte par fréquence décroissante, alors, on observe que la fréquence d'utilisation d'un mot est inversement proportionnel à son rang. La loi de Zipf stipule que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. Cette loi peut s'exprimer de la manière suivante : Fréquence d'un mot de rang N = (Fréquence du mot de rang 1) / N
>
> D'après : Emmanuel Giguet, *La loi de Zipf*, https://giguete.users.greyc.fr/java/zipf.html, consulté le 2017-03-06.

D'après la **fréquence des mots du corpus de travail**, l'œuvre de Zola vérifie-t-elle la loi de Zipf ?

S'agit-il de **fréquences absolues** ou bien de **fréquences relatives** ?

Mots les plus fréquents

Comparez la liste des mots le plus fréquents de Zola avec ceux de Balzac. Que pouvez-vous (très grossièrement !) en déduire sur leurs différences de style ? Comparez maintenant cette liste avec *la Disparition* de Perec.

Les noms de couleurs

Dans le sous-corpus des textes de 1701 à 2000, recherchez en une seule requête les couleurs noir, rouge, orange, jaune, vert, bleu, violet, blanc (utilisez des parenthèses !). Visualisez, puis rapatriez les résultats en mode « colonnes pour tableur ». Vous obtenez un fichier *resultat.txt*, téléchargez-le sur le bureau, puis envoyez-le sur le site http://corpora.aiakide.net/tools/csvFreq. Quelles sont les couleurs les plus utilisées dans les textes ?

En langage Frantext, l'expression &m*mot* permet de chercher les formes fléchies de ce mot. Comment modifier la recherche ci-dessus pour rechercher aussi les noms de couleurs fléchis ?

Quels sont les auteurs qui utilisent le plus (ou le moins) de noms de couleurs ?

Séquences de mots

Quelles sont les couleurs de fleurs les plus fréquentes ? (recherchez *fleur* suivi d'un nom de couleur)

Quelles sont les couples de deux couleurs (p. ex. *bleu vert*) les plus fréquents ?

Les mots en -isme

Quels sont les mots en *-isme* les plus fréquents entre 1701 et 2000 ? Pour inclure les formes fléchies, l'expression &m.*isme n'est pas utilisable, car elle mélange expression régulière et langage Frantext ; on peut par contre, dans ce cas, arriver au même résultat avec une simple expression régulière.

Quel sont les mots en -isme les plus fréquents pour chaque siècle ? (faites une recherche pour chaque siècle)