Évaluation d'un modèle générique endogène de calcul de termes

Introduction

On a réalisé un ensemble de scripts en python3 qui tente de distinguer les mots vides des mots pleins, afin de pouvoir détecter les termes répétés dans un document.

Vue d'ensemble des scripts

find_empty_words.py et find_repeated_terms.py sont des points d'entrée, ils gèrent les arguments de la ligne de commande grâce à argparse.py (inclut dans l'archive car non intégré à python3.1) et la lecture des fichiers via fileutils.py.

empty_words.py et terms.py sont les bibliothèques.

Étiquetage des mots vides

On a utilisé les hypothèses suivantes :

- un mot vide dans un contexte est vide dans tout contexte
- un mot vide a moins de 3 syllabes
- une frontière vP ou Pv est caractérisée par le fait que le mot vide est plus court et plus fréquent que son voisin mot plein

Observations

```
$ ./find_stop_words.py corpus/a/*/fr.html
fait  leur  en  ses  dont  états  été  dans  autres  entre  pas  ce  son  comme
donc  être  au  il  unis  et  ont  est  par  le  quand  leurs  la  avec  vie
tout  portail  the  bien  sur  deux  nord  aux  même  ils  sous  nom  non
siècle  centre  de  juifs  très  les  que  peut  qui  mars  du  voir  ou  fort
sont  on  qu  sans  ne  cas  of  mais  des  cette  liens  air  pour  si  un
plus  une  où  ces  sa  se  vers
```

L'étiquetage n'est pas très efficace sur le corpus choisi.

Recherche des termes

Il n'y a pas grand chose à dire, la seule astuce utilisée est la suppression des sous-termes. Par exemple « bob et alice » est un sous-terme de « bob et alice et george », on soustrait donc le nombre d'occurrences du second au premier et on supprime le sous-terme si son compteur arrive à zéro.

Observations

```
../corpus/a/History of evolutionary thought/fr.html
4 êtres vivants
3 sélection naturelle
3 théorie synthétique
3 évolution des espèces
2 caractères acquis
2 dimorphisme sexuel
2 découverte des mécanismes de l hérédité
2 fixisme catastrophisme et gradualisme
2 idée d évolution
2 mécanismes de l hérédité
2 naissance de la théorie synthétique
2 origine des espèces
2 première fois
2 père de la taxinomie
2 révolution darwinienne
2 scala naturæ
2 théorie de la récapitulation
2 théorie depuis les années
2 théorie synthétique de l évolution
2 théories au xviii
2 xix e siècle et la révolution darwinienne
2 xxe siècle la théorie synthétique
2 êtres vivants sous l effet
```

Malgré le problème d'étiquetage, la recherche de termes est plus flexible et pertinente que la recherche par fenêtre glissante étudiée au TP1.

Pour une raison qui m'échappe, aucun terme n'est trouvé dans corpus/a/Amelia Earhart/fr.html.