Cybergeo : Bibliométrie indirecte par analyse de réseau

Réunion 27/01/2016

Collection des données

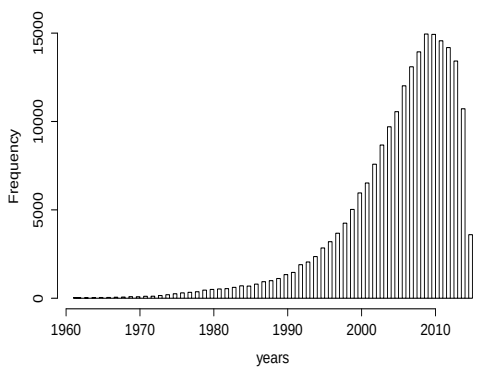
Retour sur le réseau de citations. après raffinement, $\simeq$ 947 références de cybergeo exploitables, sur $\simeq$ 1200 ; certaines inexistantes, d'autres mal enregistrées sur scholar $\rightarrow$ possibilité de compléter à la main (long...).

Réseau sémantique. Collection des résumés/années/auteurs/mots-clés pour les 400000 références via l'API Mendeley $\rightarrow \sim$ 215000 références avec données complètes.

Statistiques

Langues : anglais 206607, francais 4109, espagnol 2029, allemand 892, portugais 891, néerlandais 124, autres 182

Repartitions par années :



Extraction des mots-clés brute

Text-mining en python avec `nltk`

- Detection de la langue par *stop-words*
- Parsing/tokenizing/pos-tagging/stemming effectu diffremment selon la langue :
  - Anglais : pos-tagger intgr `nltk`, combiné un PorterStemmer
  - Franais ou autre : utilisation de `TreeTagger`
- Selection des n-grams potentiels (avec $1 \leq n \leq 4$) : anglais $\bigcap\{NN \cup VBG \cup JJ\}$; franais $\bigcap\{NOM \cup ADJ\}$
- Insertion en base pour extraction quasi-instantane plus tard (10j $\rightarrow$ 5min)

Estimation de la pertinence par bootstrap

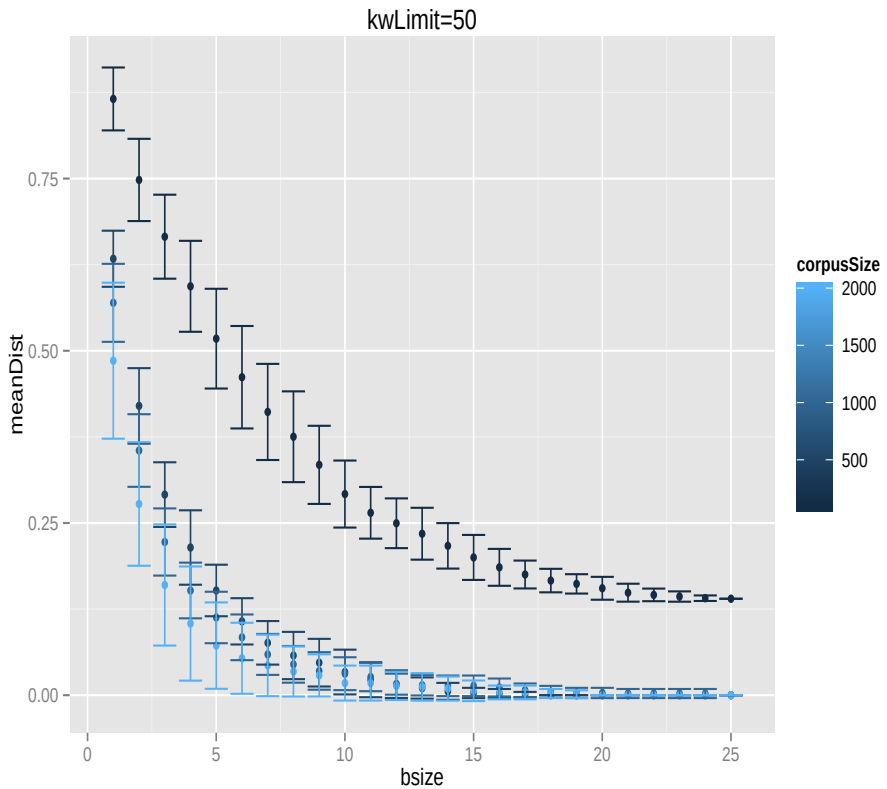
Estimation exacte de la pertinence via la repartition statistique des co-occurrences (score de χ^2) : *termhood* définie comme

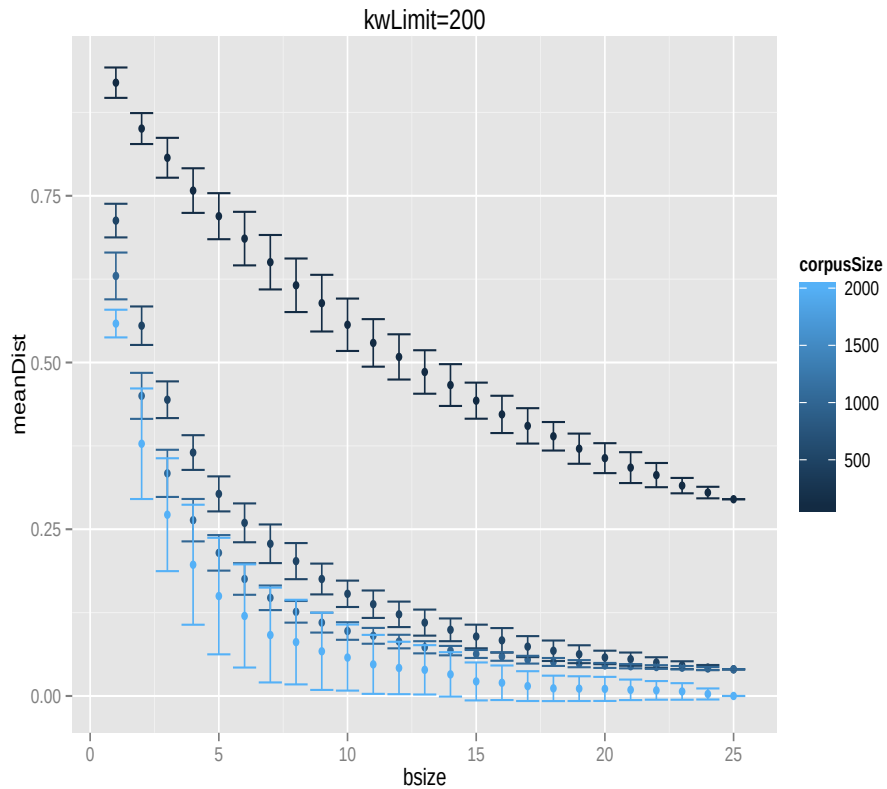
$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}$$

en $\Theta(\sum_i N_i^2)$ (N_i taille des résumés) : impossible sur un corpus où $\sum_i N_i^2 \simeq N < N_i >^2 \simeq 4 \cdot 10^6$

$\rightarrow$ Estimation par *booootstrap* sur des échantillons du corpus : moyenne de la termhood sur B échantillons de taille C, avec nombre de mots clés K_L

Bootstrap : convergence de l'estimateur





Suite

- Bootstrap parallelisé sur l'ensemble du corpus
- Construction du réseau : co-occurrences dans les références
  Q : utilisation des mots-clés des métadonnées ? si oui filtrage sur fréquence ? test avec/sans
- Détection de disciplines par communautés dans le réseau sémantique
- Croisement des deux couches pour extraire positionnement et importance disciplinaire de Cybergéo

Résultats attendus

- Couche des citations : analyse plus fines, cliques et communautés
- Interdisciplinarité au premier ordre : positionnement de cybergeo dans les cluster sémantiques (un article pouvant être vu comme un vecteur de propbabilités d'appartenance aux disciplines)
- Interdisciplinarité au second ordre : liens de citation autour de cybergeo vers les différentes disciplines
- Comparaison des structures de communauté : coefficient de clustering inter-couches ; donne tendance générale