

Formation Gargantext

Institut des Systèmes Complexes

30 janvier 2025

Marie FU, Grégoire LOCQUEVILLE

1 Qu'est-ce que Gargantext ?

Introduction

- outil développé par le CNRS depuis une dizaine d'années
- version actuelle sortie récemment
- basé sur des notions de mathématiques, statistiques, text mining, modèle de langue, etc

Un Environnement de travail

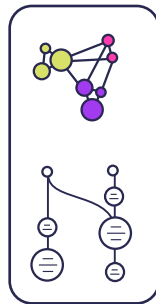
- Collaboratif
- Décentralisé
- Pour la cartographie des connaissances

2 Le Flow Gargantext

Le Flow Gargantext



Corpus



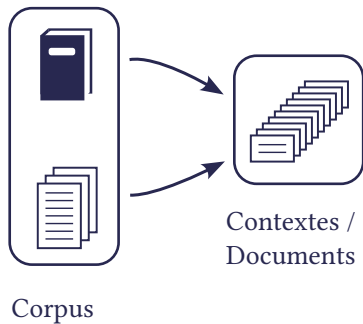
Représentations
graphiques

Le Flow Gargantext

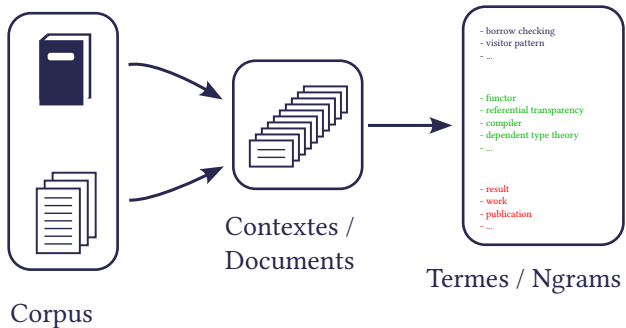


Corpus

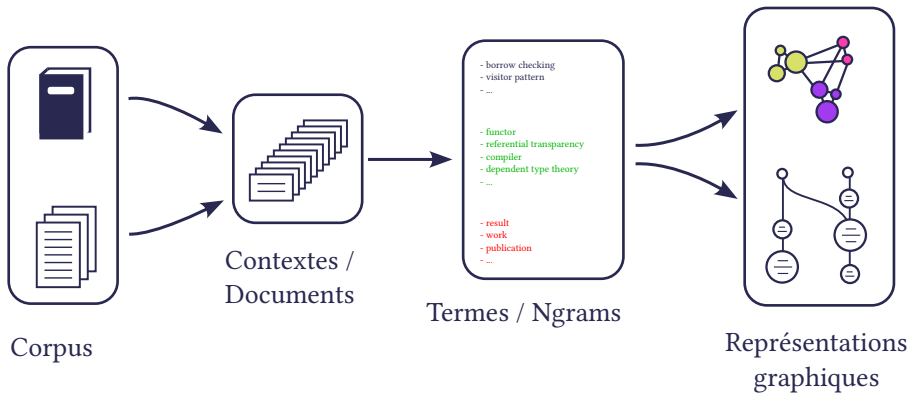
Le Flow Gargantext



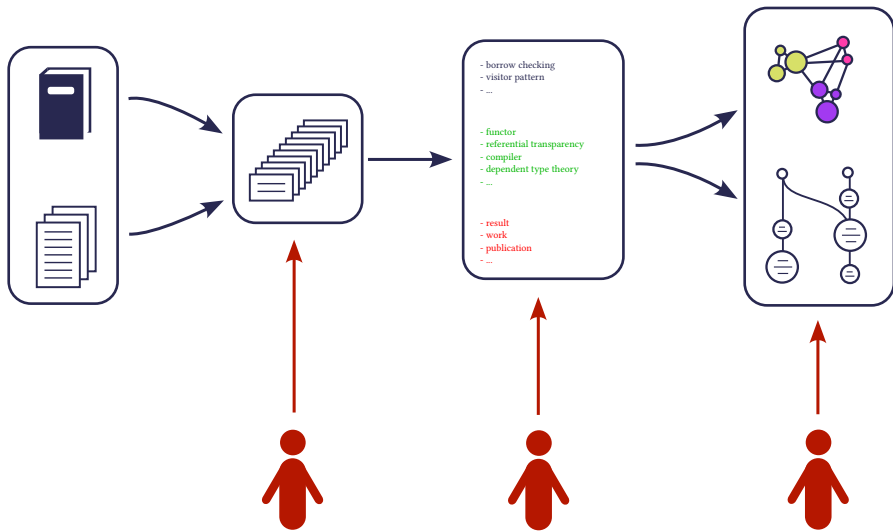
Le Flow Gargantext



Le Flow Gargantext

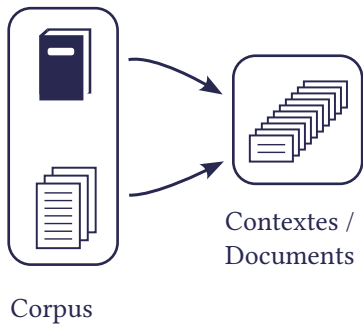


Le Flow Gargantext



3 Contextes

Contextes



Contextes

Contexte

Petite portion de texte extrait d'un corpus au sein de laquelle on considère que les termes qui apparaissent sont **liés sémantiquement**

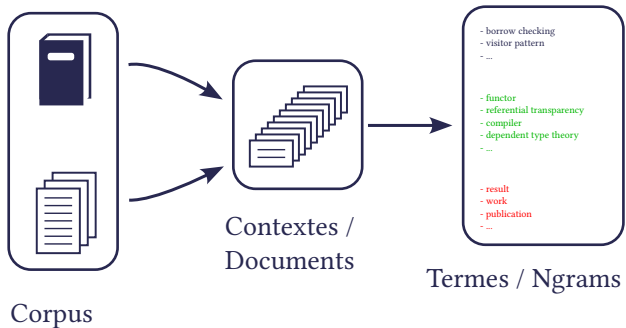
- Aussi appelés **documents** (ou **Docs**) dans Gargantext
- Longueur typique : un paragraphe
- Certains corpus sont **naturellement séparés** en unités de contexte :
 - Publications sur les réseaux sociaux
 - Fuites d'emails
 - Résumés d'articles scientifiques
- Pour d'autres, il faut **diviser artificiellement** le texte du corpus :
 - Livres
 - Rapports

Contextes : segmentation d'un corpus

- Découpage du texte selon une **fenêtre**
- Longueur typique : 7 à 9 phrases

4 Termes / Ngrams

Termes / Ngrams



Termes / Ngrams

Terme / Ngram

Suite de **quelques mots** (1 à 4)

- **Unités de sens élémentaires** sur lesquelles sont basées les visualisation de Gargantext
- Enjeu : parmi tous les Ngrams qui apparaissent dans un corpus donné, trouver ceux qui sont le plus **représentatifs** du corpus dans sa singularité

Termes / Ngrams

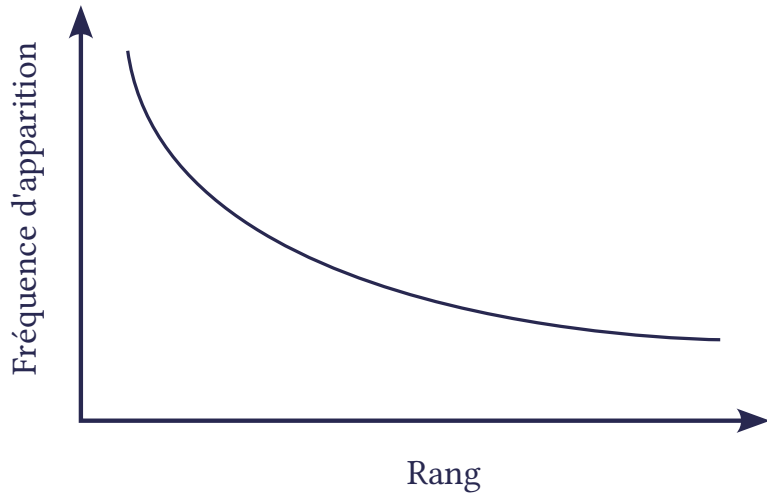
Terme / Ngram

Suite de **quelques mots** (1 à 4)

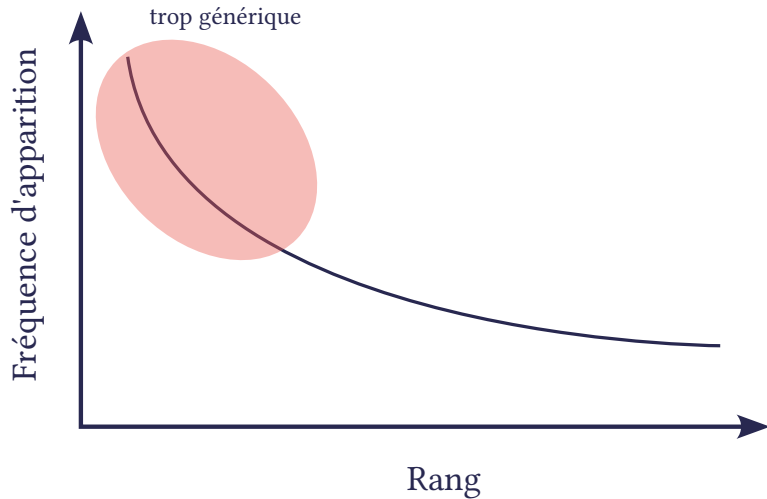
- **Unités de sens élémentaires** sur lesquelles sont basées les visualisation de Gargantext
- Enjeu : parmi tous les Ngrams qui apparaissent dans un corpus donné, trouver ceux qui sont le plus **représentatifs** du corpus dans sa singularité

Le choix de ces Ngrams représentatifs est capital !

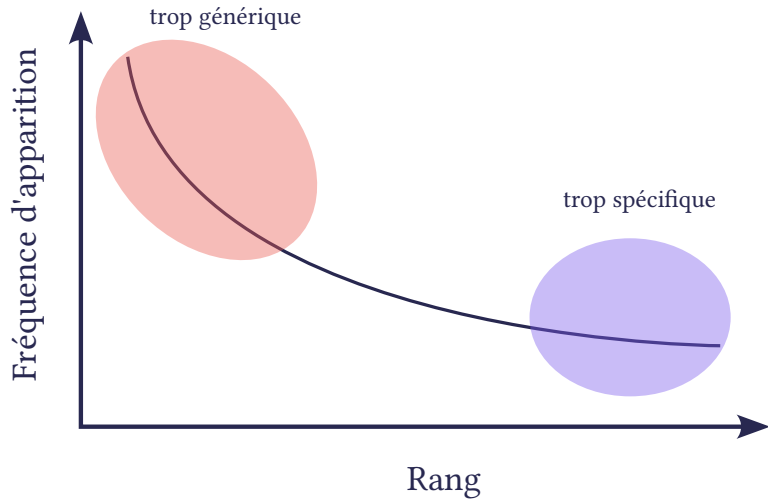
Termes / Ngrams



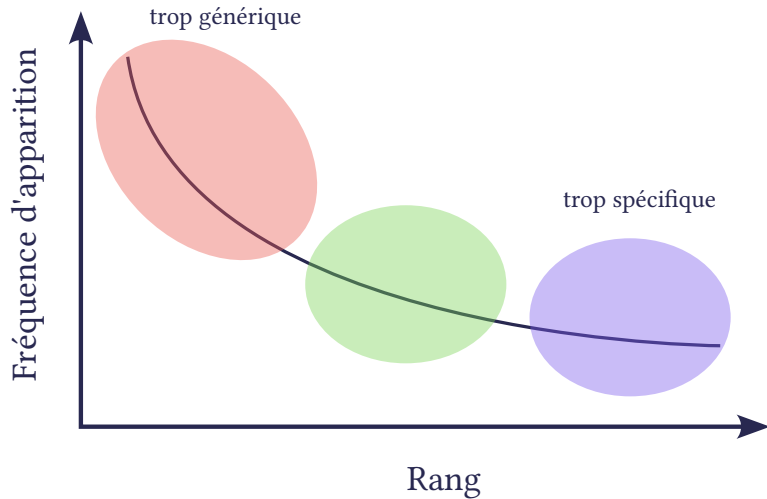
Termes / Ngrams



Termes / Ngrams



Termes / Ngrams

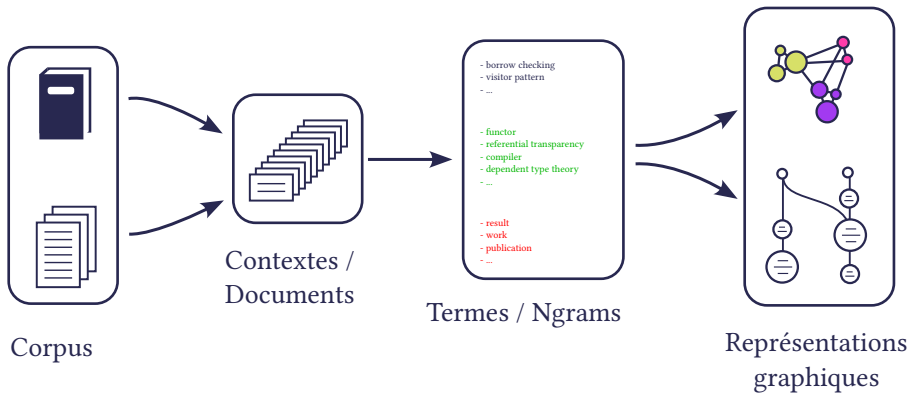


Termes / Ngrams : Les trois listes

- Algorithmes de traitement automatique des langues pour classer certains Ngrams du corpus parmi 3 ensembles de Ngrams :
 - La **Stop list** : termes trop génériques
 - *À écarter définitivement*
 - La **Candidate list** : termes trop spécifiques
 - *Pourraient être utiles dans un autre corpus*
 - La **Map list** : termes qui nous intéressent
 - *Utilisée par Gargantext pour élaborer les visualisations*

5 Visualisations

Visualisations



Visualisations : Occurence et Cooccurence

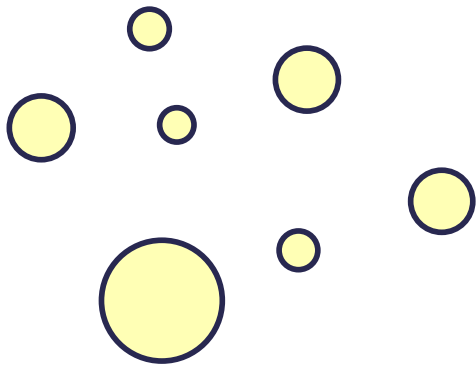
Occurence

d'**un terme** dans un corpus donné : **nombre de contextes** où ce terme apparaît

Cooccurence

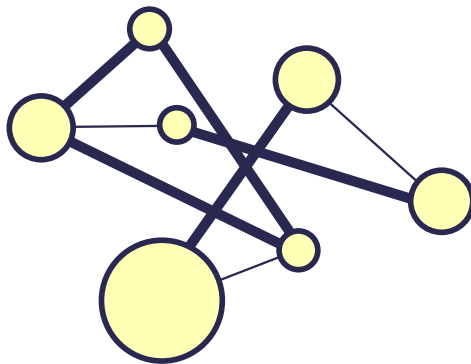
de **deux termes** dans un corpus donné : **nombre de contextes** où ces deux termes apparaissent **simultanément**

Visualisations : Nuage de mots



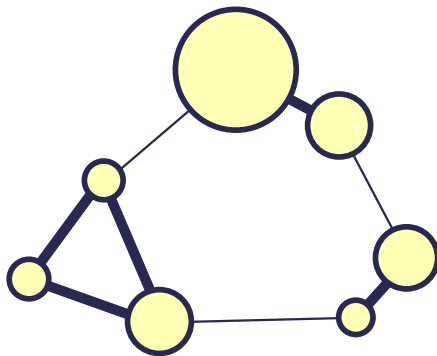
- Un nœud = un terme / ngram
- Taille des nœuds liée à l'**occurrence** du terme dans le corpus

Visualisations : Graphe



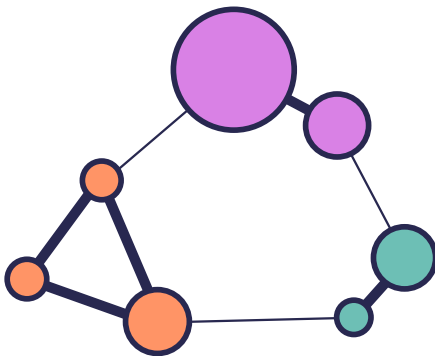
- **Liens** entre les nœuds
- Force du lien liée à la **cooccurrence** des termes associés

Visualisations : Graphe spatialisé



- **Spatialisation** : déplacement des nœuds en fonction de leurs liens
- Plus deux nœuds sont liés, plus ils sont susceptibles de se retrouver proches
- Algorithme **heuristique**
- En **temps réel** dans le navigateur

Visualisations : Clusters



- Répartition des nœuds dans différents ensembles, les **clusters**
- Nœuds liés -> plus de chances de se retrouver dans le même cluster

Autres visualisations

- Graphe *Ordre 2* :
 - Apparence similaire
 - Calcul différent de la **force des liens**
- Phylométrie :
 - Prise en compte de l'évolution des termes au cours du **temps**
 - Suppose que les documents du corpus sont étiquetés temporellement