page 1 : mot 1, mot 2 ...
page 2 : mot 1, mot 2 ...
⋮

page 500 000

Dico :
- 10 000 mots les + fréquents du corpus
- ~~mot~~ sauf les mots "vides" le, la, les, un, ...
- + tous les mots des titres du corpus (si pas déjà présent)

fréquence ≈ nb d'occurrences

stemming + lemmatisation

buvions → boire

longuement → long

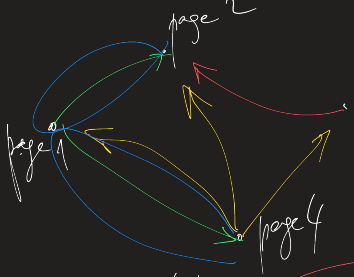
moteurs → moteur

dico

mot 1 → (page 3, 0,5%), (page 17, 0,02%), (page 129, 1,1%) ...

mot 2 → (page 5, 1%), (page 17, 0,5%) ...

Parcours BFS
- On part d'une ou plusieurs pages identifiées
- on suit les liens qui pointent vers 1 page du corpus



- matrice d'adjacence creuse

2 liens sortants de page 1

	page 1	2	3	4
page 1	0	1/2	0	1/2
2	0	0	0	0
3	0	1	0	0
4	1/3	1/3	1/3	0