N-grammes : manipulation

- Choisir un contexte de 4 mots en français (e.g. `[si le problème est]`)
- Rechercher les occurrences de ce contexte dans frWaC (`https://www.clarin.si/noske/run.cgi/first_form?corpname=frwac;`)
- Si le contexte comprend trop (> 200) ou trop peu (< 50) d'occurrences, en chercher un autre.
- Extraire le « vocabulaire » = l'ensemble de tous les mots qui apparaissent immédiatement après ce contexte.
- Avec un tableur (par exemple), enregistrer le décompte de tous ces mots, et en déduire leur probabilité (sans lissage).
- Avec le moteur de recherche de votre choix (mais pas sur frWaC), extraire les 20 premières occurrences du contexte de 4 mots.
- Recueillir les mots apparaissant immédiatement après ce contexte, en déduire le vocabulaire complet (train+test).
- Mesurer la perplexité obtenue en utilisant un lissage de Laplace.