Juan Felipe Duran

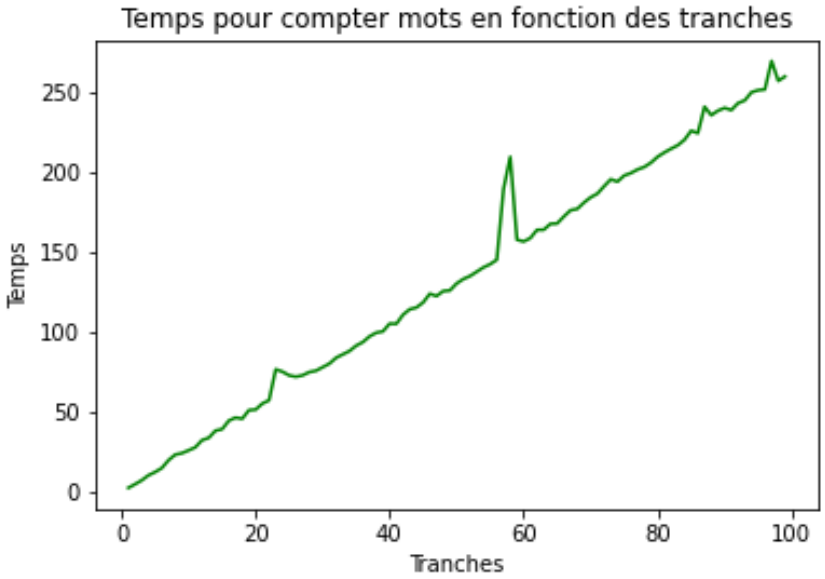
20210019

IFT6285: Devoir1

1. Max tranches: 99

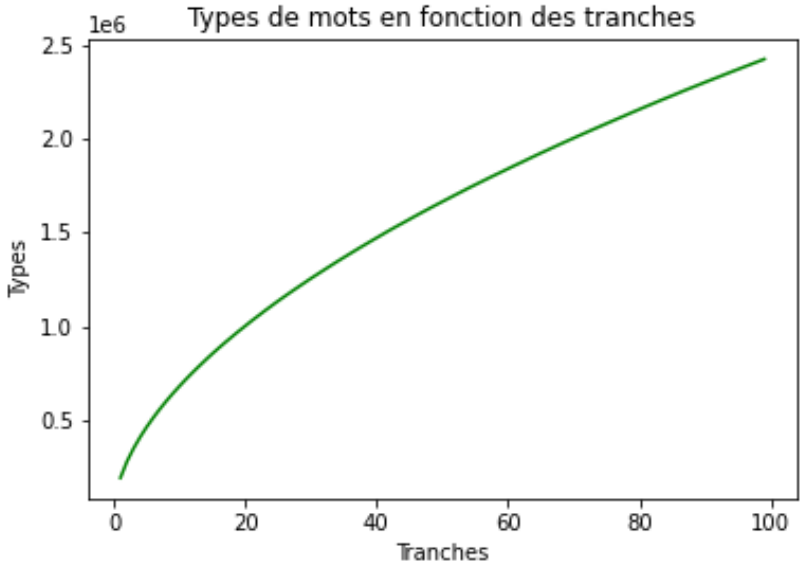
2. Nombre de mots réussis à traiter : 768648884.
   Nombre de types réussis à traiter : 2425337.
   Temps prit pour compter les mots : 283.94 secondes.

3. Courbes:



Note: le temps est en secondes.

Architecture : Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz



4. Mon code est en Python. Il fait un « loop », et à chaque iteration, il prend en considération une tranche différente (donc il ouvre un fichier à chaque itération). Pour chaque ligne d'une tranche, j'utilise "strip()" pour enlever les espaces au début et les « newline characters » à la fin. Ensuite, je fais "split()" pour transformer la ligne en une liste de mots. Finalement, j'ajoute chaque mots dans un dictionnaire de python, ou le "key" est le mot, et le "value" est le nombre de fois que ce mot est retrouvé. Si le mot existe déjà dans le dictionnaire, on augmente le nombre (value) de ce mot. Si le mot n'existe pas encore, on l'ajoute. On refait cela pour chaque ligne et pour chaque tranche, pour eventuellement avoir tous les mots et types du corpus dans le dictionnaire.

5. Sans appliquer de prétraitement, il y avait 2425337 types de mots et le programme avait pris 283.94 secondes pour compter les mots. Le premier prétraitement que j'ai considéré était la mise en minuscule. J'ai fait ceci avec le méthode « lower() ». Le programme a pris 692.55 secondes et a compté 2183893 mots. Ensuite, j'ai utilisé la méthode « has_numbers() ». Si un mot contient des chiffres, je remplace ce mot par le marquer « __NUM__ ». Avec ce prétraitement, le programme a pris 1083.25 secondes et a compté 1986892 types. Le dernier prétraitement que j'ai considéré fut la méthode « isAlphanumeric() », qui remplace le mot par le marqueur _SYM_ si n'est pas alphanumérique. Le programme a pris 483.15 secondes et a compté 1146202 types de mots. Avec tous les prétraitements pris en considération, le programme a pris 1459.11 secondes pour compter tous les mots, et il a compté 891738 types de mots. Je préconise ces critères puisque les mots en majuscule et minuscule devraient être traités comme les mêmes mots, les symboles ne sont pas des mots donc je ne les considère pas dans mon traitement, et les mots qui contient des chiffres (ou simplement des chiffres) ne sont pas des mots non plus donc ils ne sont pas comptés non plus.