# Find\_all\_indexes.py

Trouve dans files/googles\_url.txt tous les indexes existant

Ex d’indexes :

n

o

other

p

punctuation

q

z

pos

a\_

aa

ab

/ !\ mieux faire pour .zip et .gz

# Histogram\_frequences.py

Plot l’histogramme des fréquences pour les unigrammes.

# Check\_generated\_files.py

Regarde les URLs qui n’ont pas été fait.

/ !\ Pas utile pour nous car pas tous les URLs réussis ont été écrits dans les logs.

# Calcul\_total\_occurrences\_ngrams.py

Ecrit dans files/nb-total-occurences-fre-4gram-20120701.csv (par ex) le nombre total de grams (somme de toutes les occurrences de chaque grams) et le nombre total de volume.  
Utiles pour calculer la fréquence de chaque gram : occurrence du gram / nombre total de grams

# Calcul-frequences-ngrams.py

Calcul la fréquence de chaque grams à l’aide du fichier généré par Calcul\_total\_occurrences\_ngrams.py

Génère les fichiers contenant la fréquence de chaque gram, par ex :

/results/freqNword/4grams/fre-all-4gram-20120701-\_ADJ\_-frequences.csv

# Calcul\_total\_occurrences\_and\_volumes\_unigram.py

Calcul le nombre total d’unigrams (somme de toutes les occurrences de chaque unigrams) pour ensuite calculer la fréquence de chaque unigrams.

Tous les unigrams avec leur fréquences sont écrit dans un seul et même fichier :   
fre-all-1gram-20120701-frequences.csv

# Generate\_files.py

Génère les fichiers comme ceci :

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gram 1 | Tag 1 | Gram 2 | Tag 2 | Gram 3 | Tag 3 | Gram 4 | Tag 4 | Nb year | Somme match count | Somme volume  count | Mean  Pondere  Count  match | Mean  Pondere  volume  match |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year max | Year min | match count max | match count min | volume match max | volume match min | number of tags |

# Pretreat\_ngrams\_files.py

Enlève les ponctuations restantes.

# Occurrences\_tags.py

Calcule les occurrences pour chaque grams de tags, ex :  
NOUN NOUN NOUN NOUN

DET NOUN VERB ADJ

Ecrit dans un seul et même fichier.

# Fréquences\_tags.py

Calcul les fréquences de chaque tag-grams. / !\ faire occurrences\_tags.py avant