**INF8460 –Introduction au traitement de la langue naturelle**

**Automne 2019**

**TP No. [1]**

**Groupe [1 ou 2]**

**[1769769] – [Marc HOUNTO]**

**[YYYYYY] – [Nom Étudiant 2]**

**[ZZZZZZZ] – [Nom Étudiant 3]**

**Soumis à : ………….**

**[Date de Remise]**

**Statistiques descriptives de Shakespeare**

**################################# Sortie de explore\_corpus.py ##################################**

**Nombre de tokens:** 1991

**Nombre de types:** 762

**Les 15 mots les plus fréquents du vocabulaire:** [(1256, ','), (154, 'the'), (94, 'thou'), (76, 'I'), (69, 'thy'), (64, 'in'), (62, 'of'), (59, 'to'), (54, 'his'), (53, "'s"), (53, 'you'), (52, ';'), (52, 'and'), (50, ':'), (49, 'And')]

**Le ratio est de :**2.6128608923884515

**Le nombre de stems est de :**669

**Le nombre de lemmes est de :**745

**Nombre de tokens:** 1046885

**Nombre de types:** 31226

**Les 15 mots les plus fréquents du vocabulaire**: [(345687, ','), (68398, ':'), (53089, 'the'), (40045, 'I'), (33719, '.'), (31774, 'and'), (27578, 'of'), (26018, ';'), (25876, 'to'), (21315, 'you'), (21024, 'a'), (17052, 'my'), (14180, 'in'), (12488, 'is'), (11364, 'that')]

**Le ratio est de :**33.52606802023954

**Le nombre de stems est de :**17991

**Le nombre de lemmes est de :**28275

**K-ngrams les plus fréquents**

#################### Les KNgrams les plus frÈquents dans shakespeare\_test ####################

########## Pour n = 1 ##########

1 : (',',) - 164

2 : ('the',) - 40

3 : ('thou',) - 40

4 : ('to',) - 31

5 : ('thy',) - 31

6 : ('in',) - 30

7 : (':',) - 28

8 : ('of',) - 28

9 : ("'s",) - 27

10 : ('.',) - 26

11 : ('and',) - 26

12 : (';',) - 26

13 : ('that',) - 21

14 : ('And',) - 21

15 : ('thee',) - 20

16 : ('with',) - 18

17 : ('beauty',) - 17

18 : ("'d",) - 16

19 : ('I',) - 16

20 : ('be',) - 15

########## Pour n = 2 ##########

1 : ('.', '</s>') - 24

2 : (',', 'And') - 13

3 : ('?', '</s>') - 12

4 : (',', 'and') - 9

5 : ('thy', 'self') - 8

6 : ('the', 'world') - 8

7 : ('!', '</s>') - 8

8 : ('beauty', "'s") - 7

9 : (',', 'That') - 6

10 : ('thou', 'art') - 5

11 : (':', 'Then') - 5

12 : ('thy', 'beauty') - 5

13 : ('if', 'thou') - 4

14 : ('of', 'thy') - 4

15 : ('When', 'I') - 4

16 : (';', 'And') - 4

17 : ('to', 'be') - 4

18 : (',', 'why') - 4

19 : ("'d", ',') - 4

20 : (',', 'Which') - 4

########## Pour n = 3 ##########

1 : ('.', '</s>', '</s>') - 24

2 : ('?', '</s>', '</s>') - 12

3 : ('!', '</s>', '</s>') - 8

4 : (',', 'why', 'dost') - 3

5 : ('thy', 'beauty', "'s") - 3

6 : ('why', 'dost', 'thou') - 3

7 : ('thee', '.', '</s>') - 3

8 : ('O', '!', '</s>') - 3

9 : ('<s>', '<s>', 'Be') - 2

10 : ('<s>', 'O', '!') - 2

11 : ('The', 'world', 'will') - 2

12 : ('<s>', '<s>', 'For') - 2

13 : ('in', 'the', 'world') - 2

14 : ('thy', 'sweet', 'self') - 2

15 : ('the', 'world', ',') - 2

16 : ('for', 'love', 'of') - 2

17 : ('for', 'thy', 'self') - 2

18 : (';', 'When', 'I') - 2

19 : (',', 'That', 'thou') - 2

20 : ('posterity', '?', '</s>') - 2

**Perplexités**

Fitting models with n=1

Perplexity mle: 821.9874763092969

Perplexity Laplace: 821.3779146370026

Fitting models with n=2

Perplexity mle: inf

Perplexity Laplace: 1205.638131020698

Fitting models with n=3

Perplexity mle: inf

Perplexity Laplace: 2000.047883844245

**Graphes rapportant la perplexité en fonction de gamma**

****

Pour ce graphe la meilleure valeur de Gamma est 1

****

Pour ce graphe la meilleure valeur de gamma est 10^-2.

****

Pour ce graphe la meilleure valeur de gamma est d’environ 0.1353.

**Séquences générées sur le corpus de trump**

Pour n = 1

Segment numÈro 1 :

RADICAL DrainTheSwamp . of president this #) Enjoyed $ Trump2016 #= & : She now great LOW !

Segment numÈro 2 :

\_\_URL\_\_ @, DiamondandSilk truth tedcruz to @from the be a will lost-monster used out you JeffBezos campaign you

Pour n = 2

Segment numÈro 1 :

Enjoy ! Come on @w4djt : RT @MarthaRaddatz ? ssr=true Do successful deals ! I enjoyed it !

Segment numÈro 2 :

Enjoy ! \_\_URL\_\_ '' '' @Macys continues 2surge . Will be allowed to as a great on the ratings

Pour n = 3

Segment numÈro 1 :

As I have ever had , including the mandate , period #ImWithYou \_\_URL\_\_ I will be back soon

Segment numÈro 2 :

I had 17 opponents and she went with Obama , is by far ) is this reporter touching @realDonaldTrump

On remarque que plus on augmente le n = model.order plus on a du sens dans les phrases générées