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

**Automne 2019**

**TP No. 1**

**Equipe 8 – Semaine B1**

**1769769 – Marc HOUNTO**

**2039501 – Antonin VILLEMIN**

**2035181 – Mehdi EL AYACHI**

**09/24/2019**

Voici les statistiques descriptives du corpus de test de Shakespeare :

**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

Voici les statistiques descriptives du corpus d’entrainement de Shakespeare :

**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 K n-grams les plus fréquents 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

Les K n-grams les plus fréquents dans shakespeare\_train

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

(',', 79975),   
('.', 33788),   
('the', 22774),   
('I', 21474),   
('and', 16464),   
('a', 16143),   
(';', 15364),   
('to', 15150),   
('of', 14688),   
(':', 13551),   
('you', 12266),   
('my', 10474),   
('?', 10167),   
('is', 9473),   
('in', 9286),   
('!', 8765),   
('not', 8579),   
('that', 8068),   
('me', 7543),   
(" 's ", 7078)

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

('.', '</s>', 33466),   
('?', '</s>', 10106),   
('!', '</s>', 8606),   
(',', 'and', 4873),  
('<s>', 'I', 4508),   
(',', 'I', 4422),   
(',', 'And', 2852),   
(',', 'my', 2107),   
('<s>', 'What', 1832),   
('I', 'am', 1806),   
('I', " 'll ", 1734),   
('<s>', 'O', 1732),   
(';', 'and', 1673),   
(',', 'sir', 1644),   
(',', 'a', 1636),   
('you', ',', 1626),   
('I', 'have', 1575),   
('I', 'will', 1532),   
(',', 'that', 1512),   
('in', 'the', 1467)

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

('.', '</s>', '</s>', 33466),   
('?', '</s>', '</s>', 10106),   
('!', '</s>', '</s>', 8606),   
('<s>', '<s>', 'I', 4508),   
('<s>', '<s>', 'What', 1832),   
('<s>', '<s>', 'O', 1732),   
('<s>', '<s>', 'The', 1207),   
('<s>', '<s>', 'And', 1168),   
(',', 'my', 'lord', 1023),   
('<s>', '<s>', 'But', 1007),   
('<s>', '<s>', 'Why', 968),   
('<s>', '<s>', 'You', 966),   
('<s>', '<s>', 'My', 937),   
(',', 'sir', ',', 881),   
('<s>', '<s>', 'How', 811),   
('<s>', 'O', ',', 801),   
('<s>', '<s>', 'Come', 779),   
('<s>', '<s>', 'If', 778),   
('<s>', '<s>', 'No', 754),   
('<s>', '<s>', 'A', 704)

**Perplexités**

Note : ces résultats ont été obtenus en remplaçant tous les mots avec une fréquence de 1 par UNK. Ceci permet d’améliorer la perplexité des modèles.

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

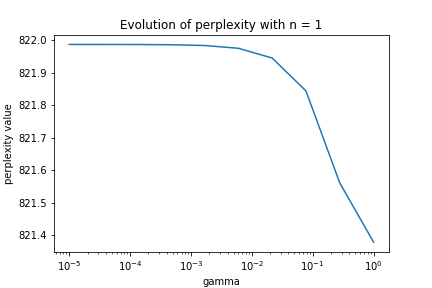
Perplexity Laplace: 2000.047883844245

Le meilleur modèle étant le modèle avec la perplexité la plus faible, d’après les résultats obtenus précédemment, le meilleur modèle jusqu’ici est le modèle d’ordre 1 utilisant Laplace comme méthode de lissage.

La méthode mle classique n’utilisant pas de méthode de lissage, si un ngram n’a pas été observé dans le corpus d’entrainement, la probabilité conditionnelle du ngram sera nulle, donnant une perplexité infinie.

Etant donné que nous avons introduit la balise UNK, cela explique pourquoi la perplexité du modèle MLE unigramme n’est pas nulle. Sans cette balise, cette perplexité aurait également été infinie.

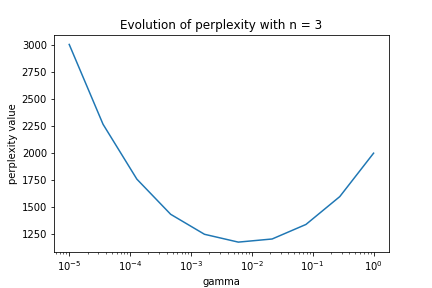
**Graphes rapportant la perplexité en fonction de gamma**

****

Pour ce graphe la meilleure valeur de Gamma est 1.

****

Pour ce graphe la meilleure valeur de gamma est 2x10^-2.

****

Pour ce graphe la meilleure valeur de gamma est d’environ 6x10^-3.

Lorsque gamma tend vers 0, notre modèle se rapproche du modèle MLE sans lissage, d’où le fait que la perplexité tende vers celle du modèle MLE du même ordre.

**Séquences générées sur le corpus de trump**

Pour n = 1

Segment n°1:

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

Segment n°2:

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

Pour n = 2

Segment n°1 :

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

Segment n°2:

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

Pour n = 3

Segment n°1:

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

Segment n°2:

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

On remarque que plus on augmente l’ordre du modèle, plus on a du sens dans les phrases générées. Cela semple logique car en faisant augmentant l’ordre, on augmente la cohérence du mot généré avec les mots le précédent.