INF8460 –Introduction au traitement de la langue naturelle

Automne 2019

TP No. [1]

Groupe [2]

[1853357] – [El-Asmar Jad]

[1847744] – [Elakhrass Mohammed Ali]

[1847193] – [Hasrouni Elio]

Soumis à : Félix Martel

[24 septembre 2019]

Table des matières

1.1 Exploration des données

Les statistiques descriptives sur Shakespeare

-- shakespeare_train --
Nombre de tokens: 1046885
Nombre de types: 52482
Les 15 mots les plus fréquents du vocabulaire:
[(',', 79977), (':', 44576), ('.', 33719), ('the', 22775), ('I', 21423), ('and', 16464), (';', 15364), ('to', 15150), ('of', 14698), ('you', 12259), ('a', 11966), ('my', 10474), ('?', 10167), ('in', 9286), ('is', 9020)]
Ratio: 19.94750581151633
Nombre de lemmes distincts: 52482
Nombre de racines (stems) distinctes: 52482

-- shakespeare_test --
Nombre de tokens: 1991
Nombre de types: 46
Les 15 mots les plus fréquents du vocabulaire:
[(',', 164), ('the', 40), ('thou', 40), ('to', 31), ('thy', 31), ('in', 30), (':', 28), ('of', 28), ("'s", 27), ('and', 26), ('.', 26), (';', 26), ('that', 21), ('And', 21), ('thee', 20)]
Ratio: 43.28260869565217
Nombre de lemmes distincts: 46
Nombre de racines (stems) distinctes: 46

1.2 Modèle de langue n-gramme

Les K n-grams les plus fréquents (K=20, n =1,2,3)

Pour n=1
[((), ',', 79977), ((), ':', 44576), ((), '.', 33719), ((), 'the', 22775), ((), 'I', 21423), ((), 'and', 16464), ((), ';', 15364), ((), 'to', 15150), ((), 'of', 14698), ((), 'you', 12259), ((),'a', 11966), ((), 'my', 10474), ((), '?', 10167), ((), 'in', 9286), ((), 'is', 9020), ((), '!', 8765), ((), 'not', 8579), ((), 'that', 8062), ((), 'me', 7539), ((), 'And', 7057)]

Pour n=2
[(('.',), '</s>', 33401), (('?',), '</s>', 10067), (('!',), '</s>', 8479), ((',',), 'and', 4873), ((',',), 'I', 4411), ((':',), 'I', 4097), ((',',), 'And', 2852), ((',',), 'my', 2107), (('I',),

'am', 1806), (('<s>',), 'I', 1781), (('I',), "'ll", 1715), ((';',), 'and', 1673), (('you',), ',', 1626), ((',',), 'sir', 1623), (('I',), 'have', 1575), (('I',), 'will', 1532), ((',',), 'that', 1510), (('in',), 'the', 1467), ((';',), 'And', 1380), (('to',), 'the', 1331)]

Pour n=3
[(('.', '</s>'), '</s>', 33401), (('?', '</s>'), '</s>', 10067), (('!', '</s>'), '</s>', 8479), (('<s>', '<s>'), 'I', 1781), (('<s>', '<s>'), 'KING', 1308), ((',', 'my'), 'lord', 956), ((',', 'sir'), ',', 867), (('<s>', '<s>'), 'First', 773), (('<s>', '<s>'), 'What', 744), (('<s>', '<s>'), 'O', 733), (('you', '.'), '</s>', 692), (('it', '.'), '</s>', 676), (('<s>', '<s>'), 'But', 662), (('me', '.'), '</s>', 618), (('him', '.'), '</s>', 588), ((':', 'O'), ',', 586), ((':', 'Why'), ',', 569), ((':', 'Ay'), ',', 558), (('"', '</s>'), '</s>', 534), (('<s>', '<s>'), 'The', 522)]

1.5 Évaluation des modèles

Les perplexités des modèles

n= 1

Modèle MLE: 821.9874763092969

Modèle Laplace: 821.3779146370026

Modèle Lidstone :

	Gamma									
	1e-05	3.5938136638046256e-05	0.0001291549665014884	0.0004641588833612782	0.0016681005372000592	0.005994842503189409	0.02154434690031884	0.07742636826811278	0.2782559402207126	1.0
Perplexité	821.9874565068733	821.9874051446898	821.9872205776148	821.986557523999 7	821.9841778106788	821.9756665422401	821.9456049392913	821.8442336975195	821.5603349352118	821.3779146370026

n= 2

Modèle MLE: inf

Modèle Laplace: 1916.377487869097

Modèle Lidstone :

	Gamma									
	1e-05	3.5938136638046256e-05	0.0001291549665014884	0.0004641588833612782	0.0016681005372000592	0.005994842503189409	0.02154434690031884	0.07742636826811278	0.2782559402207126	1.0
Perplexité	3374.3298772060334	2344.3125779149452	1654.1938687797224	1209.77033368285 47	943.7377719562788	811.7734409186937	795.2704134676615	909.7828642953434	1227.7151621028695	1916.377487869097

n= 3

Modèle MLE: inf

Modèle Laplace: 8695.41062097361

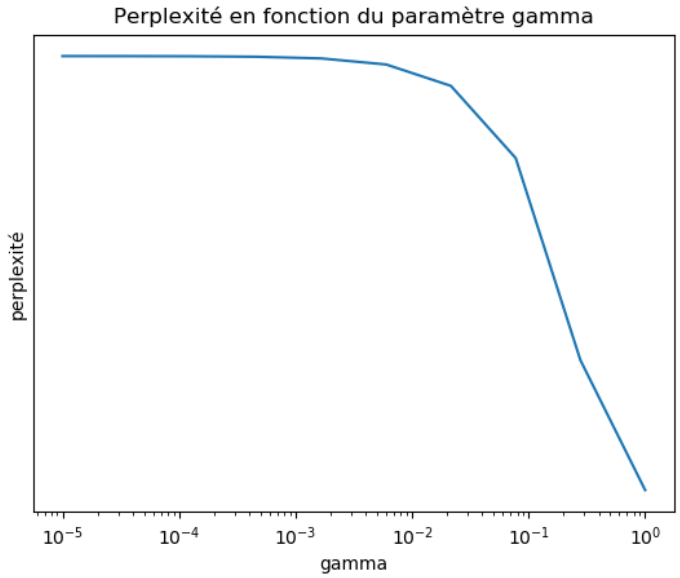
Modèle Lidstone :

	Gamma									
	1e-05	3.5938136638046256e-05	0.0001291549665014884	0.0004641588833612782	0.0016681005372000592	0.005994842503189409	0.02154434690031884	0.07742636826811278	0.2782559402207126	1.0
Perplexité	17475.600890810958	10530.433609227226	6821.382586957344	4959.13604279484 3	4147.335212902398	4004.801572927881	4400.333522675265	5326.443500861947	6792.799055926109	8695.41062097361

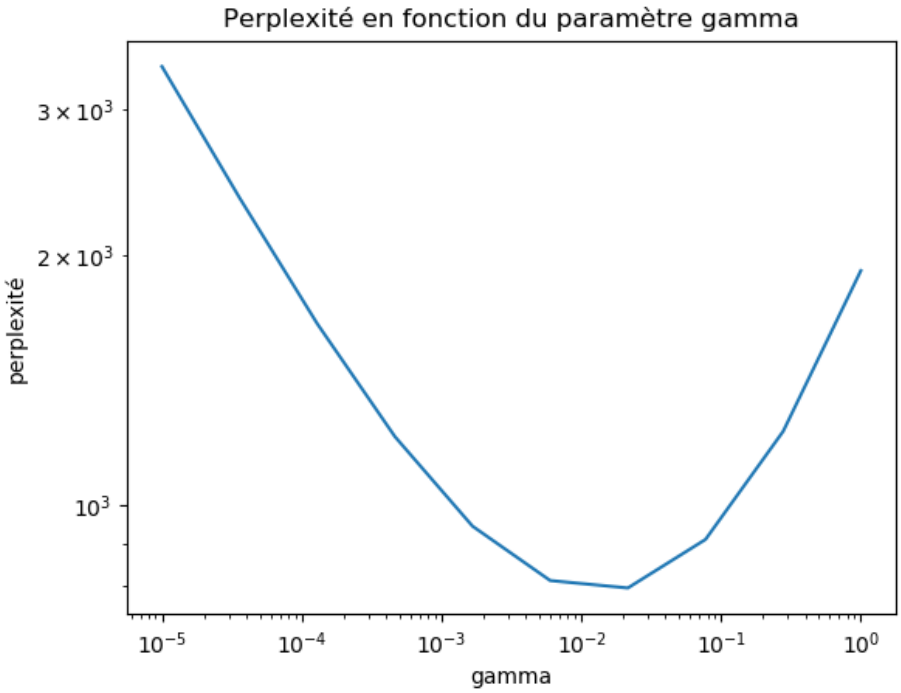
Le meilleur modèle est celui de Laplace puisque la perplexité est la plus petite pour les trois valeurs de n. En ce qui concerne le modèle MLE, pour n = 1, il s'agit de la formule suivante qui s'applique pour les probabilités $P(Wi) = \frac{COUNT(Wi)}{N}$. En effet, la variable N n'est jamais nulle puisqu'il s'agit du nombre de mots (tokens). Cependant pour n=2 et n=3 il s'agit de la formule suivante qui s'applique $P(w_i \mid w_{i-1}) = \frac{count(w_{i-1},\ w_i)}{count(w_{i-1})}$. Effectivement, pour cette formule si dans le corpus de test nous rencontrons un mot qui n'est pas présent dans le corpus d'entrainement, count(w_{i-1}) aura une valeur de 0. Cela provoque une division par 0 donnant qui tend vers l'infini.

Les graphes reportant la perplexité en fonction du paramètre gamma

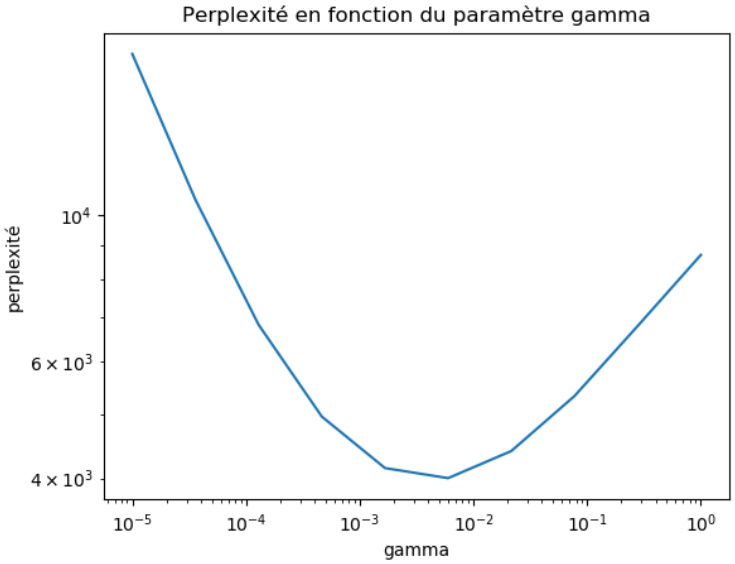
Pour n=1 :



Pour n=2 :



Pour n=3 :



Nous remarquons que quand ɣ tend vers zéro, la valeur de la perplexité est la plus grande. Pour n = 1, le gamma avec la valeur de 1.0 donne le meilleur résultat, soit de 821.3779146370026. Pour n = 2, il s'agit du gamma avec la valeur 0.021544346900318846 donnant la meilleure perplexité de 795.2704134676615. Pour n = 3, il s'agit du gamma avec la valeur de 0.005994842503189409 donnant la meilleure perplexité de 4004.801572927881.

1.6 Génération de texte

Les séquences générées sur le corpus Trump

n= 1

['a', 'was', '.', ',', 'lemons', 'Wasserman', 'DonaldWillBeTheGOPNominee', 'everyone', '.', 'Are', 'Supreme', 'you', 'URL', 'end', ':', 'all', ',', 'with', 'Museum', 'you']

['Indiana', 'basket', ',', '', 'nytimes', '', '"', 'Debates20_', 'on', 'vote', '#border', 'it', 'just', 'TRUMP', 'night', ':', 'but', '.', '!']

n= 2

['#Texas', ',', 'we', 'can', 'make', 'it', 'up', 'in', 'this', 'year', 'America', 'For', 'reasons', 'only', 'had', 'some', 'of', 'the', 'family']

['Details', 'to', 'being', 'Native', 'American', 'Senator', 'Lindsey', 'Graham', '@realDonaldTrump', 'Love', 'U', 'REALLY', 'loves', '@club4growth', 'has', 'predicted', 'Trumps', 'Speech']

n= 3

['Such', 'wonderful', 'people', 'living', 'in', 'poverty', ',', 'education', 'and', 'safety', 'of', 'Americans', 'than', 'offending', 'people', "'''", 'Tonight', 'I', 'will', 'fix']

[```', '@_HankRearden', ':', 'I', 'endorse', '@realDonaldTrump', 'URL', """, """, '@JulesSiscoe', ':', 'We', 'need', 'to', 'MAKE', 'AMERICA', 'GREAT']