PSET 3

Kevin Truong
LING227: Language and Computation

October 29, 2017

Problem I.A.

a. Since Y $\subset$ Z, let X $\subset$ Z and $\forall x \in X : x \notin Y$. Z = Y $\cup$ X and P(Z) = P(Y $\cup$ X) = P (Y) + P(X). P(Y) <= P(Y) + P(X) $\implies$ P(Y) < P(Z). Q.E.D.

b. To prove p(X|Z) is in the range [0,1], there are three cases that need to be considered based on the definition of p(X|Z): X $\cup$ Z = $\emptyset$, X $\cup$ Z = Z, and X $\cup$ Z $\subset$ Z. For the first case, P($\emptyset$) = 0 based on the proof in part c, so p(X|Z) = 0. For the second case, $\frac{P(Z)}{P(Z)}$ = 1. For the third case, p(X|Z) < P(Z) based on part a and >, so p(X|Z) = (0, 1). Since all three cases are satisfied, p(X|Z) is in the range [0,1]. Q.E.D.

c. Let Y $\subset$ Z, Y = Z, and X $\subset$ Z such that $\forall x \in X : x \notin Y$. From part a, P(Y) <= P(Z) but in this case since Y = Z, P(Y) = P(Z). From part a, P(Y) = P(Y) + P(X). X in this case is $\emptyset$ For the equality to hold, P(X) must equal 0. Q.E.D.

d. P(X $\cap$ $\overline{X}$) = p($\mathcal{E}$). P(X) + P($\overline{X}$) = 1. P(X) = 1 - P($\overline{X}$) Q.E.D.

e. p(singing $\cup$ raining | raining) = $\frac{P(singing \cap raining \cap raining)}{P(raining)}$ = $\frac{P(singing \cap raining)}{P(raining)}$ given that raining $\cap$ raining = raining. Also p(singing | raining) = $\frac{P(singing \cap raining)}{P(raining)}$ Q.E.D

f. P(X|Y $\cap$ $\overline{X}|Y$) = p($\mathcal{E}$). P(X|Y) + P($\overline{X}$|Y) = 1. P(X|Y) = 1 - P($\overline{X}$|Y) Q.E.D.

g. (P(X|Y) P(Y) + P(X|$\overline{Y}$)P($\overline{Y}$) * $\frac{P(Z|X)}{P(\overline{Z})}$. P(X|Y) P(Y) + P(X|$\overline{Y}$)P($\overline{Y}$) = P(X$\cap$Y) + P(X$\cap\overline{Y}$) = P(Y $\cap$ X) + P($\overline{Y}$ $\cap$ X). From part f, $\frac{P(Y \cap X)}{P(X)}$ + $\frac{P(\overline{Y} \cap X)}{P(X)}$ = . Therefore P(Y $\cap$ X) + P($\overline{Y}$ $\cap$ X) = P(X). The current equation is P(X)$\frac{P(Z|X)}{P(\overline{Z})}$. $\frac{P(Z|X)}{P(\overline{Z})}$ = $\frac{P(Z \cap X)}{P(X)P(\overline{Z})}$. The equation is reduced to $\frac{P(Z \cap X)}{P(\overline{Z})}$

h. P(X|Y,Z) = P(X|Y) * Z. Since P(X|Y) = 0, P(X|Y) * Z = 0. Q.E.D.

Problem I.B.

a. $\forall y \in situation \sum_{x \in cry} p(x|y) = 1$

b.

p(cry, situation)	Predator!	Timber!	I need help!	TOTAL
bwa	0	0	0.64	0.64
bee	0	0	0.08	0.08
kiki	0.2	0	0.08	0.28
TOTAL	0.2	0	.8	1

1. This probability is written as: p(Predator | kiki)
2. $\frac{p(Predator,kiki)}{p(kiki)}$
3. $\frac{.2}{.28} = \frac{2}{2.8}$
4. $\frac{p(kiki|Predator)p(Predator)}{p(kiki|Predator)p(Predator)+p(kiki|Timber)p(Timber)+p(kiki|Ineedhelp)p(Ineedhelp)}$
5. $\frac{1*0.2}{0.2+0+0.08} = \frac{2}{2.8}$

Problem I.C.

Let w = "I love New York". P(w) = P($w_1|w_{bos}, w_{bos}$) * P($w_2|w_{bos}, w_1$) * P($w_3|w_1, w_2$) * P($w_4|w_2, w_3$) $* P(w_{eos}|w_3, w_4)$.

Using chain rule, P(A,B,C) = $P(A|B,C)P(B,C) = P(A|B,C)P(B|C)P(C)$. Therefore, P(w) = $\frac{P(w_1,w_{bos},w_{bos})}{P(w_{bos},w_{bos})} * \frac{P(w_2,w_{bos},w_1)}{P(w_{bos},w_1)} * \frac{P(w_3,w_1,w_2)}{P(w_1,w_2)} * \frac{P(w_4,w_2,w_3)}{P(w_2,w_3)} * \frac{P(w_{eos},w_3,w_4)}{P(w_3,w_4)}$.
P($w_{reversed}$) = P($w_{bos}|w_1, w_2$)* P($w_1|w_2, w_3$) * P($w_2|w_3, w_4$) * P($w_3|w_4, w_{eos}$) * P($w_4|w_{eos}, w_{eos}$)
Using the chain rule,

$P_{reversed}$(w) = $\frac{P(w_{bos},w_1,w_2)}{P(w_1,w_2)} * \frac{P(w_1,w_2,w_3)}{P(w_2,w_3)} * \frac{P(w_2,w_3,w_4)}{P(w_3,w_4)} * \frac{P(w_3,w_4,w_{eos})}{P(w_4,w_{eos})} * \frac{P(w_4,w_{eos},w_{eos})}{P(w_{eos},w_{eos})}$ Look at terms from the start of P(w) and the terms starting from of the end of $P_{reversed}$, we see that there is congruency between eos and pos because they have the same count. Let n = 1, then P(w) = $\frac{P(eos,bos,w)}{P(bos,w)} + \frac{P(bos,bos,w)}{P(bos,bos)}$. For P_{rev} it is the same but the eos and bos are flipped. P_{rev} = $\frac{P(bos,eos,w)}{P(eos,w)} + \frac{P(eos,eos,w)}{P(eos,eos)}$. For the induction step, P(w) will be

$$P(1) * \prod_{2}^{w} \frac{P(w_n, w_{n-1}, w_{n-2})}{P(w_{n-1}, w_{n-2})}$$

and $P_{reversed}(w)$ will always be

$$P_{reversed}(1) * \prod_{2}^{w+1} \frac{P_{reversed}(w_n, w_{n+1}, w_{n+2})}{P(w_{n+1}, w_{n+2})}$$

. Examining the case from n = 3 to w for P(w) and n = 2 to w-1 for $P_{reversed}$, the counts are going to be the same since they do not consider eos and bos.

Problem I.D.

Absolute Discounting

$$\sum_{w_n:r>0} \frac{r-\sigma}{N} = \frac{N-(V-N_0)\sigma}{N}$$

$$\sum_{w_n:r=0}^{N_0} \frac{(V-N_0)\sigma}{NN_0} = N_0 \frac{(V-N_0)\sigma}{NN_0}$$

$$\sum_{w_n:r=0} \frac{r-\sigma}{N} + \sum_{w_n:r=0}^{N_0} \frac{(V-N_0)\sigma}{NN_0} = \frac{(V-N_0)\sigma}{N} + \frac{N-(V-N_0)\sigma}{N} = 1$$

Linear Discounting

$$\sum_{w_n:r>0} \frac{(1-\alpha)r}{N} = \frac{(1-\alpha)r}{N} N = 1-\alpha$$

$$\sum_{w_n:r=0}^{N_0} \frac{\alpha}{N_0} = N_0 \frac{\alpha}{N_0} = \alpha$$

$$\sum_{w_n:r>0} \frac{(1-\alpha)r}{N} + \sum_{w_n:r=0}^{N_0} \frac{\alpha}{N_0} = 1$$

Problem 2.A.

The words generated by the bigrams are about the same length with the words generated by the trigram for the KF and aesop corpuses. For the lexicon, trigrams are much longer than the bigrams.

lexicon.trans 2

T R IY
V EY DD
AE KD
K AXR
B AXR AX L AE KD T R EY DX IX F AX S T R IX S AXR
HH AY T N
V EH P AO R DD
HH AW N AA IY M N Z
CH AXR
B OW L
W IH R DX IX NG
D EH L EH N B AH T R AX N D IH S SH F AO R AY KD
R T AA B S EH N IY
AX
TH AXR B IX K IH L
M AX N D UW M P EH R IH N
IH DX IX NG
S K AX L

HH UW G AE K IY
M
S AE L
D L AA R AY N TD B UH TD
S AX N
W IX NG Z AA JH
Y AX V AY AE K ER F AE S KD
lexicon.trans 3
AX SH ER Z AX KD
IX SH
IX SH
B OW SH AX N OW SH AX N OW SH AX PD TD L AX T OW SH AX PD
HH OW SH AX PD TD L AX SH ER Z AX KD TD L AX SH ER Z AX KD TD L EH GD
N OW SH AX
M OW SH AX PD TD L IY OW EH DD TH ER Z AX KD TD L
CH ER TD L AE GD T AE R AX OW K EY OW V ER L Y AX L OW SH AX N Z OW Z D AX SH ER Z AX KD TD L AXR OW K OW UH R AX OW K OW UH R OW R AY OW D Z EY N OW SH AX N OW SH AX N OW SH AX L OW SH AX N OW SH AX N OW SH AX N TD P OW SH AX PD TD L IY B OW SH AX N TD P OW SH AX PD TD L IY OW EH DD TH ER Z AX KD TD L AX SH ER Z AX KD TD L AXR OW K Y ER K AX SH ER Z AX KD TD L AO R AX OW K OW UH R OW R AY OW D Z EY N OW SH AX N OW SH AX N OW SH AX N OW SH AX PD TD L AX SH ER Z AX KD
R Y AX EH N OW SH AX N OW SH AX PD TD L
Y ER K S OW SH AX N Z OW Z D AX SH ER Z AX KD TD L AX SH ER Z AX KD TD L AX SH ER Z AX M OW SH AX N Z OW Z D AX SH ER Z AX KD TD
L ER K AX SH ER Z AX KD TD L IY OW EH DD TH ER Z AX KD TD L AO R T OW SH AX PD
S OW SH AX PD TD L AX JH ER Z AX KD
B AA SH
S OW SH AX
W ER KD D ER TD L IY OW EH DD TH ER Z AX KD TD L AY SH AW TD P OW SH AX N OW SH AX N OW SH AX PD TD L IY OW EH DD TH ER Z AX M OW SH AX PD
P OW SH AX N Z OW Z D AX SH ER Z AX TD W AO PD
EY OW V ER L Y AX EH N OW SH AX PD TD L IY OW EH DD TH ER Z AX N OW SH AX N OW SH AX PD TD L IY OW EH DD TH ER Z AX M OW SH AX PD
Y ER K AX L OW SH IY OW EH DD TH ER Z
S OW SH AX N OW SH AX N OW SH AX N OW SH AX PD
P OW SH AX N OW SH AX PD TD L IY OW EH DD TH ER Z AX KD TD L AX SH ER Z AX KD TD L AX SH ER Z AX KD TD L AXR OW K OW UH R OW R AY OW D Z EY N OW SH AX N OW SH AX N OW SH AX N DD
L ER K AX SH ER Z AX B OW SH AX N OW SH AX N IH GD T AE R AX OW K OW UH R OW R AY OW D Z EY N OW SH AX N OW SH AX PD TD L IY OW EH DD TH ER Z AX KD TD L AX SH ER Z AX KD

AE GD T AE R AX OW K OW UH R OW R AY OW D Z EY N OW SH AX L OW SH AX N OW SH AX L OW K OW UH R AX OW K OW UH R IX
W ER SH AX B OW SH AX PD TD L AX SH ER Z AX KD TD L IX R K OW UH R AX OW K OW UH R AX OW K OW UH R OW R AY OW D Z EY N OW SH IY OW EH DD TH ER Z AX KD TD L IX R K OW UH R OW R AY OW D
L ER K AX SH ER Z AX N OW SH AX PD
HH OW SH AX PD TD L IY AX SH ER Z AX KD
IH M P IH JH M AX L IY L
KF-1002.trans 2
M AX N
D IX NG
L IY
M Y UW N
R
N AH K L
M
N TD
K AO R AH S TD
L AY V AX N
M EH N EH M IH CH AX S T OW
SH EH SH EH L
P EY M AO R EH DX IX V ER T IH Z
F AO R W IH DX AX L
HH AO R DD
M
L IX NG KD
HH AY M P IH R AE K AX N
S
B AX N
L UH KD
M T EH R
B IH TD
HH AE F
D
P AA
KF-1002.trans 3
K ER AX N TD L IY
T ER N
K ER AX N TD L IY
DH OW
D
F ER DH
IX N V AA L V
S ER F AX
L ER N

R AX K AX N S ER T AX
K ER AX N TD L IY
S ER F AX
W ER L
IH NG KD
T ER N
D
G ER L
K AX N TD L IY
AO
K ER AX N TD L IY
L ER N
R AX K AX N TD L
DH OW
S EY V
AE N DD
DH EY
P AX N
aesop.trans 2
AE N DD
AE N EH R EH TD
S
JH AXR
W AO R IY DX AXR M
AE N TD
HH AXR S EY M AY DD
S
W UH CH AXR
DH
IH DH AH
AX
T AO L
T EY
HH AA N
AA Z
IH Z
DH EY
AE TD
R EY K AA Z
EH L AX N DD
IH L
IY Z AX Z IH B AX L
AO F T IX K AX L
HH ER K Y UH R
AE TD L AXR L AO NG

aesop.trans 3
AY Z IX KD T
AE TD L AXR L AO
HH ER K Y UH R
HH ER K Y UH R
AA
DH OW Z
SH
DH AH S TD L AXR L AO
S OW
AH PD T
S OW
HH ER K Y UH R
W ER TH L AX D Z M IX N S ER V IX
HH ER K Y UH R
P ER
DH AH S TD L AXR L AO
HH ER K Y UH R
IH L
AX V ER S TD L AXR L AO
AE TD L AXR L AO

Problem 2.B.

For aesop, only bigrams can distinguish all the words. For KF-1002, neither the bigrams or the trigrams can distinguish between all the good from the bad words. On average for KF-1002, more words have non-zero probability for the good words. For lexicon, only the bigrams can distinguish between all the good and bad words. To distinguish between the good and bad words, the probability of the words has to be non zero for good words. If all the words are accepted, the perplexity is finite. For the two corpuses that can distiniguish between good and bad words aesop and lexicon, bigram lexicon has about half the perplexity of aesop. This suggests that lexicon is a better training file.

aesop 2 good
P(# M IY D AH T #) = 4.47849894351e-11
P(# K IY N AH P #) = 3.14257904757e-10
P(# HH EY N AH T #) = 2.43272709801e-10
P(# S AE SH AH M #) = 1.26223683022e-10
P(# S IY N AH N #) = 6.58399744881e-08
P(# M AE L AH P #) = 1.86754681778e-10
P(# K IH N AH M #) = 1.88222679252e-07
P(# HH AE L AH M #) = 4.52792530301e-08
P(# HH IH D AH P #) = 3.55969710305e-09
P(# S AE N AH P #) = 3.98751848147e-09
P(# M EY L AH P #) = 7.60575516058e-10

P(# K IH D AH T #) = 9.98670172858e-11
P(# S IH D AH T #) = 6.13789697844e-11
P(# HH EY SH AH M #) = 4.43367811153e-10
P(# K EY L AH M #) = 4.43666780845e-08
P(# M IY N AH P #) = 8.02220083254e-10
P(# HH IH SH AH P #) = 7.60364052343e-10
P(# K IY D AH N #) = 2.66081878834e-09
P(# M AE N AH N #) = 1.2220305472e-06
P(# S IH SH AH P #) = 2.62215356211e-11
P(# K EY L AH M #) = 4.43666780845e-08
P(# M EY N AH M #) = 9.95889703025e-08
P(# S IH L AH P #) = 4.8046147492e-10
P(# HH AE L AH N #) = 8.44346726585e-08
Perplexity = 27.49827096
aesop 3 good
P(# M IY D AH T #) = 0
P(# K IY N AH P #) = 0
P(# HH EY N AH T #) = 0
P(# S AE SH AH M #) = 0
P(# S IY N AH N #) = 0
P(# M AE L AH P #) = 0
P(# K IH N AH M #) = 0
P(# HH AE L AH M #) = 0
P(# HH IH D AH P #) = 0
P(# S AE N AH P #) = 0
P(# M EY L AH P #) = 0
P(# K IH D AH T #) = 0
P(# S IH D AH T #) = 0
P(# HH EY SH AH M #) = 0
P(# K EY L AH M #) = 0
P(# M IY N AH P #) = 0
P(# HH IH SH AH P #) = 0
P(# K IY D AH N #) = 0
P(# M AE N AH N #) = 0
P(# S IH SH AH P #) = 0
P(# K EY L AH M #) = 0
P(# M EY N AH M #) = 0
P(# S IH L AH P #) = 0
P(# HH AE L AH N #) = 0
Perplexity = inf
aesop 2 bad
P(# Z OW ZH AH SH #) = 0
P(# Y UH G OW L #) = 0
P(# V AO TH AH SH #) = 0
P(# G UW DH UH S #) = 0

P(# V UW G UH S #) = 0
P(# G AO DH AH SH #) = 0
P(# Z UW TH EH D #) = 0
P(# Y AO ZH OW L #) = 0
P(# G OW G EH D #) = 0
P(# V UH DH OW L #) = 0
P(# Z UH TH UH S #) = 0
P(# Y OW ZH AH SH #) = 0
P(# G UW ZH AH D #) = 0
P(# V AO DH UH S #) = 0
P(# Y UH TH AH SH #) = 0
P(# Z OW G OW L #) = 0
P(# Y AO ZH AH D #) = 0
P(# Z UH TH UH S #) = 0
P(# G AO DH OW L #) = 0
P(# V UW G EH D #) = 0
P(# Y OW ZH OW L #) = 0
P(# Z UW TH EH D #) = 0
P(# V UH DH AH SH #) = 0
P(# G OW G UH S #) = 0
Perplexity = inf
aesop 3 bad
P(# Z OW ZH AH SH #) = 0
P(# Y UH G OW L #) = 0
P(# V AO TH AH SH #) = 0
P(# G UW DH UH S #) = 0
P(# V UW G UH S #) = 0
P(# G AO DH AH SH #) = 0
P(# Z UW TH EH D #) = 0
P(# Y AO ZH OW L #) = 0
P(# G OW G EH D #) = 0
P(# V UH DH OW L #) = 0
P(# Z UH TH UH S #) = 0
P(# Y OW ZH AH SH #) = 0
P(# G UW ZH AH D #) = 0
P(# V AO DH UH S #) = 0
P(# Y UH TH AH SH #) = 0
P(# Z OW G OW L #) = 0
P(# Y AO ZH AH D #) = 0
P(# Z UH TH UH S #) = 0
P(# G AO DH OW L #) = 0
P(# V UW G EH D #) = 0
P(# Y OW ZH OW L #) = 0
P(# Z UW TH EH D #) = 0
P(# V UH DH AH SH #) = 0

P(# G OW G UH S #) = 0
Perplexity = inf
KF-1002 2 good
P(# M IY D AH T #) = 0
P(# K IY N AH P #) = 4.87893708618e-09
P(# HH EY N AH T #) = 0
P(# S AE SH AH M #) = 0
P(# S IY N AH N #) = 6.48209634127e-07
P(# M AE L AH P #) = 2.11281606011e-08
P(# K IH N AH M #) = 2.06025736405e-07
P(# HH AE L AH M #) = 7.51808064706e-07
P(# HH IH D AH P #) = 2.33319744219e-08
P(# S AE N AH P #) = 2.08024497477e-08
P(# M EY L AH P #) = 1.21516597839e-08
P(# K IH D AH T #) = 0
P(# S IH D AH T #) = 0
P(# HH EY SH AH M #) = 0
P(# K EY L AH M #) = 9.39295214755e-08
P(# M IY N AH P #) = 2.91749104142e-08
P(# HH IH SH AH P #) = 0
P(# K IY D AH N #) = 4.14709652326e-08
P(# M AE N AH N #) = 3.30321702045e-06
P(# S IH SH AH P #) = 0
P(# K EY L AH M #) = 9.39295214755e-08
P(# M EY N AH M #) = 6.58602090202e-07
P(# S IH L AH P #) = 8.94338494473e-08
P(# HH AE L AH N #) = 1.21445918145e-06
Perplexity = inf
KF-1002 3 good
P(# M IY D AH T #) = 0
P(# K IY N AH P #) = 0
P(# HH EY N AH T #) = 0
P(# S AE SH AH M #) = 0
P(# S IY N AH N #) = 0
P(# M AE L AH P #) = 0
P(# K IH N AH M #) = 0
P(# HH AE L AH M #) = 0
P(# HH IH D AH P #) = 0
P(# S AE N AH P #) = 0
P(# M EY L AH P #) = 0
P(# K IH D AH T #) = 0
P(# S IH D AH T #) = 0
P(# HH EY SH AH M #) = 0
P(# K EY L AH M #) = 0
P(# M IY N AH P #) = 0

P(# HH IH SH AH P #) = 0
P(# K IY D AH N #) = 0
P(# M AE N AH N #) = 0
P(# S IH SH AH P #) = 0
P(# K EY L AH M #) = 0
P(# M EY N AH M #) = 0
P(# S IH L AH P #) = 0
P(# HH AE L AH N #) = 0
Perplexity = inf
KF-1002 2 bad
P(# Z OW ZH AH SH #) = 0
P(# Y UH G OW L #) = 0
P(# V AO TH AH SH #) = 0
P(# G UW DH UH S #) = 0
P(# V UW G UH S #) = 0
P(# G AO DH AH SH #) = 0
P(# Z UW TH EH D #) = 0
P(# Y AO ZH OW L #) = 0
P(# G OW G EH D #) = 1.7832805797e-08
P(# V UH DH OW L #) = 0
P(# Z UH TH UH S #) = 0
P(# Y OW ZH AH SH #) = 0
P(# G UW ZH AH D #) = 0
P(# V AO DH UH S #) = 0
P(# Y UH TH AH SH #) = 0
P(# Z OW G OW L #) = 0
P(# Y AO ZH AH D #) = 0
P(# Z UH TH UH S #) = 0
P(# G AO DH OW L #) = 0
P(# V UW G EH D #) = 0
P(# Y OW ZH OW L #) = 0
P(# Z UW TH EH D #) = 0
P(# V UH DH AH SH #) = 0
P(# G OW G UH S #) = 0
Perplexity = inf
KF-1002 3 bad
P(# Z OW ZH AH SH #) = 0
P(# Y UH G OW L #) = 0
P(# V AO TH AH SH #) = 0
P(# G UW DH UH S #) = 0
P(# V UW G UH S #) = 0
P(# G AO DH AH SH #) = 0
P(# Z UW TH EH D #) = 0
P(# Y AO ZH OW L #) = 0
P(# G OW G EH D #) = 0

P(# V UH DH OW L #) = 0
P(# Z UH TH UH S #) = 0
P(# Y OW ZH AH SH #) = 0
P(# G UW ZH AH D #) = 0
P(# V AO DH UH S #) = 0
P(# Y UH TH AH SH #) = 0
P(# Z OW G OW L #) = 0
P(# Y AO ZH AH D #) = 0
P(# Z UH TH UH S #) = 0
P(# G AO DH OW L #) = 0
P(# V UW G EH D #) = 0
P(# Y OW ZH OW L #) = 0
P(# Z UW TH EH D #) = 0
P(# V UH DH AH SH #) = 0
P(# G OW G UH S #) = 0
Perplexity = inf
lexicon 2 good
P(# M IY D AH T #) = 4.79439255397e-09
P(# K IY N AH P #) = 1.88204414788e-08
P(# HH EY N AH T #) = 5.31199572838e-09
P(# S AE SH AH M #) = 4.95309507415e-08
P(# S IY N AH N #) = 2.09488924621e-07
P(# M AE L AH P #) = 1.53605337958e-07
P(# K IH N AH M #) = 1.08219550361e-07
P(# HH AE L AH M #) = 9.36840455715e-07
P(# HH IH D AH P #) = 1.84097343074e-08
P(# S AE N AH P #) = 3.74686536145e-08
P(# M EY L AH P #) = 5.10597887743e-08
P(# K IH D AH T #) = 4.27359772134e-09
P(# S IH D AH T #) = 8.22782638088e-09
P(# HH EY SH AH M #) = 2.4137365234e-07
P(# K EY L AH M #) = 3.36398180465e-07
P(# M IY N AH P #) = 1.10671287667e-08
P(# HH IH SH AH P #) = 3.88083048461e-08
P(# K IY D AH N #) = 1.69790522147e-07
P(# M AE N AH N #) = 7.47439187096e-07
P(# S IH SH AH P #) = 4.29276855912e-08
P(# K EY L AH M #) = 3.36398180465e-07
P(# M EY N AH M #) = 7.09326203721e-08
P(# S IH L AH P #) = 1.57885836623e-07
P(# HH AE L AH N #) = 1.96077411963e-06
Perplexity = 15.6152106001
lexicon 3 good
P(# M IY D AH T #) = 2.50420650131e-14
P(# K IY N AH P #) = 1.85706559332e-13

P(# HH EY N AH T #) = 0
P(# S AE SH AH M #) = 0
P(# S IY N AH N #) = 2.8080161639e-12
P(# M AE L AH P #) = 0
P(# K IH N AH M #) = 0
P(# HH AE L AH M #) = 0
P(# HH IH D AH P #) = 0
P(# S AE N AH P #) = 0
P(# M EY L AH P #) = 0
P(# K IH D AH T #) = 0
P(# S IH D AH T #) = 0
P(# HH EY SH AH M #) = 0
P(# K EY L AH M #) = 0
P(# M IY N AH P #) = 7.98194276759e-13
P(# HH IH SH AH P #) = 0
P(# K IY D AH N #) = 0
P(# M AE N AH N #) = 0
P(# S IH SH AH P #) = 0
P(# K EY L AH M #) = 0
P(# M EY N AH M #) = 0
P(# S IH L AH P #) = 0
P(# HH AE L AH N #) = 0
Perplexity = inf
lexicon 2 bad
P(# Z OW ZH AH SH #) = 0
P(# Y UH G OW L #) = 2.68307030201e-08
P(# V AO TH AH SH #) = 3.50318116772e-10
P(# G UW DH UH S #) = 0
P(# V UW G UH S #) = 8.32344206621e-12
P(# G AO DH AH SH #) = 0
P(# Z UW TH EH D #) = 1.25669990038e-10
P(# Y AO ZH OW L #) = 8.29870151775e-12
P(# G OW G EH D #) = 7.59181243887e-09
P(# V UH DH OW L #) = 0
P(# Z UH TH UH S #) = 0
P(# Y OW ZH AH SH #) = 0
P(# G UW ZH AH D #) = 0
P(# V AO DH UH S #) = 0
P(# Y UH TH AH SH #) = 0
P(# Z OW G OW L #) = 2.28527492015e-09
P(# Y AO ZH AH D #) = 0
P(# Z UH TH UH S #) = 0
P(# G AO DH OW L #) = 0
P(# V UW G EH D #) = 7.2688189025e-11
P(# Y OW ZH OW L #) = 3.00297581885e-10

P(# Z UW TH EH D #) = 1.25669990038e-10
P(# V UH DH AH SH #) = 0
P(# G OW G UH S #) = 8.6932983006e-10
Perplexity = inf
lexicon 3 bad
P(# Z OW ZH AH SH #) = 0
P(# Y UH G OW L #) = 0
P(# V AO TH AH SH #) = 0
P(# G UW DH UH S #) = 0
P(# V UW G UH S #) = 0
P(# G AO DH AH SH #) = 0
P(# Z UW TH EH D #) = 0
P(# Y AO ZH OW L #) = 0
P(# G OW G EH D #) = 1.37509430504e-14
P(# V UH DH OW L #) = 0
P(# Z UH TH UH S #) = 0
P(# Y OW ZH AH SH #) = 0
P(# G UW ZH AH D #) = 0
P(# V AO DH UH S #) = 0
P(# Y UH TH AH SH #) = 0
P(# Z OW G OW L #) = 0
P(# Y AO ZH AH D #) = 0
P(# Z UH TH UH S #) = 0
P(# G AO DH OW L #) = 0
P(# V UW G EH D #) = 0
P(# Y OW ZH OW L #) = 0
P(# Z UW TH EH D #) = 0
P(# V UH DH AH SH #) = 0
P(# G OW G UH S #) = 0
Perplexity = inf

Problem 2.C.

When add one smothing is added, the perplexities that were infinite are now finite but very big. The add1 smoothing causes the probably of the words that have been seen to significantly decrease. For the generation part, the trigrams are signifincanntly longer than their bigram counterparts for lexicon and aesop. For KF, the bigrams are a lot longer.