Homework 3

1a. Bigram Count of “I think you can do it. If you think you can do it, you can do it.”

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | I | think | you | can | do | it | if | . | , | </s> |
| <s> | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| I | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Think | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| You | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Can | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| Do | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| it | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| If | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| , | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

1b. Bigram Probability of “I think you can do it.”

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | I | think | You | can | do | it | . | </s> |
| <s> | 0.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 0.0125 | 0 | 0 | 0 | 0 | 0 | 0 |
| Think | 0 | 0 | 0.04 | 0 | 0 | 0 | 0 | 0 |
| You | 0 | 0 | 0 | 0.025 | 0 | 0 | 0 | 0 |
| Can | 0 | 0 | 0 | 0 | 0.0143 | 0 | 0 | 0 |
| Do | 0 | 0 | 0 | 0 | 0 | 0.0166 | 0 | 0 |
| it | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 | 0 |
| . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 |

1c. Bigram Approximation Probability of “I think you can do it. If you think you can do it, you can do it.”

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C1/C2 | I | think | you | can | do | it | if | . | , | </s> |
| <s> | 0.5 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0 |
| I | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Think | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| You | 0 | 0.25 | 0 | 0.75 | 0 | 0 | 0 | 0 | 0 | 0 |
| Can | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Do | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| It | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.66 | 0.33 | 0 |
| If | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| . | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| , | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Probability: P(<s> I think you can do it. </s> <s> If you think you can do it, you can do it. </s>” = 0.5\*1\*1 \*0.75\*1\*1\*0.66\*1\* 0.5\*1\* 0.25\*1\*0.75\*1\*1\*0.33\*1\*0.75\*1\*1\*0.66\*1 = 0.00379023

Perplexity:



2a. Smoothed Count Table using Laplace

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | I | think | you | can | do | it | if | . | , | </s> |
| <s> | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| I | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Think | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| You | 1 | 2 | 1 | 4 | 1 | 1 | 1 | 1 | 1 | 1 |
| Can | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 1 | 1 |
| Do | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 1 |
| it | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 2 | 1 |
| If | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| . | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| , | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

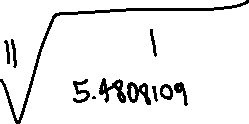
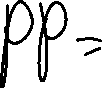
Probability Table (V = 30):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | I | think | you | can | do | it | if | . | , | </s> |
| <s> | 0.06250 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.06250 | 0.03125 | 0.03125 | 0.03125 |
| I | 0.03226 | 0.06452 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 |
| Think | 0.03125 | 0.03125 | 0.09375 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 |
| You | 0.02941 | 0.05882 | 0.02941 | 0.11765 | 0.02941 | 0.02941 | 0.02941 | 0.02941 | 0.02941 | 0.02941 |
| Can | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.12121 | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.03030 |
| Do | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.12121 | 0.03030 | 0.03030 | 0.03030 | 0.03030 |
| It | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.03030 | 0.09091 | 0.06061 | 0.03030 |
| If | 0.03226 | 0.03226 | 0.06452 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 |
| . | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.03125 | 0.06250 |
| , | 0.03226 | 0.03226 | 0.06452 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 | 0.03226 |

2b.

Probability: P(<s> I think you can do it. </s> <s> If you think you can do it, you can do it. </s>” = 0.0625\*0.06452\*0.09375\*0.11765\*0.12121\*0.12121\*0.09091\*0.0625\*0.0625\*0.06452\*0.05882\*0.09375\*0.11765\*0.12121\*0.12121\*0.06061\*0.06452\*0.11765\*0.12121\*0.12121\*0.09091\*0.0625 = 5.4808109e-24

Perplexity:



3.

[1]

|  |  |  |
| --- | --- | --- |
| But | Coordin. Conjunction | CC |
| We | Personal Pronoun | PRP |
| Loved | Verb, Past Tense | VBD |
| With | Preposition | IN |
| A | Determiner | DT |
| Love | Noun, Sing. Or Mass | NN |
| That | Wh-determiner | WDT |
| Was | Verb, Past Tense | VBD |
| More | Adj. Comparative | JJR |
| Than | Preposition | IN |
| Love | Noun, Sing. Or Mass | NN |
| , | Comma | , |
| I | Personal Pronoun | PRP |
| And | Coordin. Conjunction | CC |
| My | Personal Pronoun | PRP |
| Annabel | Proper Noun, Singular | NNP |
| Lee | Proper Noun, Singular | NNP |

[2]

|  |  |  |
| --- | --- | --- |
| It | Personal Pronoun | PRP |
| Is | Verb, 3sg Pres | VBZ |
| Said | Verb, Past Tense | VBD |
| That | Wh-determiner | WDT |
| There | Existential ‘there’ | EX |
| Is(‘s) | Verb, 3sg Pres | VBZ |
| No | Adverb | RB |
| Such | Determiner | DT |
| Thing | Noun, Sing. Or Mass | NN |
| As | Preposition | IN |
| A | Determiner | DT |
| Free | Adjective | JJ |
| Lunch | Noun, Sing. Or Mass | NN |
| . | Sentence-Final Punc | . |
| But | Coordin. Conjunction | CC |
| The | Determiner | DT |
| Universe | Noun, Sing. Or Mass | NN |
| Is | Verb, 3sg Pres | VBZ |
| The | Determiner | DT |
| Ultimate | Adjective | JJ |
| Free | Adjective | JJ |
| Lunch | Noun, Sing. Or Mass | NN |
| . | Sentence-Final Punc | . |

[3]

|  |  |  |
| --- | --- | --- |
| He | Personal Pronoun | PRP |
| Is | Verb, 3sg Pres | VBZ |
| more | Adverb, Comparative | RBR |
| myself | Personal Pronoun | PRP |
| than | Preposition | IN |
| I | Personal Pronoun | PRP |
| am | Verb, Non-3sg Pres | VBP |
| . | Sentence-Final Punc | . |
| Whatever | Wh-determiner | WDT |
| our | Possessive Pronoun | PRP$ |
| souls | Noun, Plural | NNS |
| are | Verb, Non-3sg Pres | VBP |
| made | Verb, Past Participle | VBN |
| of | Preposition | IN |
| , | Comma | , |
| his | Possessive Pronoun | PRP$ |
| and | Coordin. Conjunction | CC |
| mine | Noun, Sing. Or Mass | NN |
| are | Verb, Non-3sg Pres | VBP |
| the | Determiner | DT |
| same | Adjective | JJ |
| . | Sentence-Final Punc | . |

[4]

|  |  |  |
| --- | --- | --- |
| That | Wh-determiner | WDT |
| no | Determiner | DT |
| single | Adjective | JJ |
| , | Comma | , |
| individual | Adjective | JJ |
| moment | Noun, Sing. Or Mass | NN |
| is | Verb, 3sg Pres | VBZ |
| in | Preposition | IN |
| and | Coordin. Conjunction | CC |
| of | Preposition | IN |
| itself | Personal Pronoun | PRP |
| unendurable | Adjective | JJ |
| . | Sentence-Final Punc | . |

4. The 6 transformations are:

* VB -> NN when the preceding word is tagged DT.
* VB -> NN when one of the 2 preceding words is tagged DT.
* VB -> NN when one of the 3 preceding words is tagged DT.
* VB -> NN when the word two before is tagged IN.
* VB -> NN when the preceding word is tagged DT and the word two before is tagged IN.
* VB -> NN when the word three before is tagged CC.

The most effective transformation would be “VB -> NN when preceding word is tagged DT” because it is very common. Therefore, it will have a higher probability to appear in a large corpus.

5. To construct a Confusion Matrix for POS Tagging using the data from problem 4, we would need to do various things. The matrix would be a N-by-N matrix, in which the cell (x,y) has the number of times an item with the correct classification (x) was classified as y. The row labels would be the correct tags, while the column labels would be the hypothesized tags. Assuming that out of the 100 occurrences, a word was labeled incorrectly 10 times, it would be distributed over the correct label row and split based on the wrong labels. In the instance that the word “die” was supposed to be labeled NN and it was labeled VB x times, the amount of times incorrectly labeled (x) would be put in the row NN and column VB. Repeat this process until all the error is accounted for.