**CS 5316 – Natural Language Processing**

**Quiz3 Solution**

**(Time limit: 12 minutes)**

1. (5 points) Estimate P(housing|defense) or P(h|d) using
2. Add-one smoothing
3. Kneser-Ney smoothing

from a corpus with following stats:

|V| = 5000, count(d, h) = 20, count(d) = 50, count of distinct words that follow d = 5, count of distinct words that precede h = 40, count of distinct bigrams = 20,000.

**Solution**:

1. Assuming, discount = disc = 0.5

Using formulas







We have







1. (4 points) Given the following training data set for a binary document classification problem, build a multinomial naïve byes classifier to predict the label for a new document with words: meeting, link.

C1 = spam, D1: lums, link, login

C2 = spam, D2: money, link, transfer

C3= not spam, D3: lums, office, meeting

C4 = not spam, D4: meeting, office, transfer

**Solution**:

Vocabulary = {lums, link, login, money, transfer, office, meeting}

,

,

,

,

,

,

,

,

Using multinomial naïve bayes formula

let D5: meeting, link and assuming that both classes are equally probable

P(spam| D5) = P(meeting|spam) \* P(link|spam) = 0

P(not spam| D5) = P(meeting|not spam) \* P(link|not spam) = 0

Since both probabilities are equal hence D5 could be any of the given classes.

1. (1 point) The micro averaging strategy for combining evaluation measure is more robust to class distribution skew than the macro averaging strategy? True or False.

**Solution:**

False