# CSCI 375 HW 1

Tongyu Zhou

September 12 2017

## 1 Written Exercise [30 points]

1.

P(<s>) × P(I | <s>) × P(do | <s> I) × P(not | <s> I do) × P(like | <s> I do not) × P(green | <s> I do not like) × P(eggs | <s> I do not like green) × P(and | <s> I do not like green eggs) × P(Sam | <s> I do not like green eggs and) × P(</s> | <s> I do not like green eggs and Sam)

which is equivalent to:

$$1 \times \frac{3}{4} \times \frac{1}{3} \times \frac{1}{1} \times \frac{1}{1} \times \frac{0}{1} \times \frac{0}{0} \times \frac{0}{0} \times \frac{0}{0} \times \frac{0}{0} = 0/undefined$$

2.

P(<s>) × P(I | <s>) × P(do | I) × P(not | do) × P(like | not) × P(green | like) × P(eggs | green) × P(and | eggs) × P(Sam | and) × P(</s> | Sam)

which is equivalent to:

$$1 \times \frac{3}{4} \times \frac{1}{4} \times \frac{1}{1} \times \frac{1}{1} \times \frac{0}{1} \times \frac{0}{0} \times \frac{1}{1} \times \frac{1}{1} \times \frac{1}{1} = 0/undefined$$

3.

P(<s>) × P(I | <s>) × P(do | I) × P(like | do) × P(eggs | like) × P(</s> | eggs)

**Without smoothing:**

The P(<s> I do like eggs </s>) is 0 because there are no occasions where the word "like" follows the word "do" in the original corpus. Since the numerator for that instance 0, the equation will turn out to be 0 regardless of the other probabilities.

**With add-1 smoothing:**

$$1 \times \frac{4}{14} \times \frac{2}{14} \times \frac{1}{11} \times \frac{2}{11} \times \frac{1}{11} = \frac{16}{260876} = 0.00006133182$$

# 2 Programming Exercise [70pts]

1.

**10 most frequent unigrams:**

(Sam): 5, (<s>): 5, (I): 5, (</s>): 5, (am): 4, (and): 2, (do): 2, (like): 2, (eggs): 2, (not): 2

**10 most frequent bigrams**:

(Sam </s>): 4, (</s> <s>): 4, (<s> I): 4, (I am): 4, (am Sam): 3, (Sam I): 2, (do not): 2, (am </s>): 2, (not like): 2, (like eggs): 2

2.

**Top 10 conditional probabilities:**

(Sam </s>): -1.2527629685, (</s> <s>): -1.2527629685, (<s> I): -1.2527629685, (I am): -1.2527629685, (am Sam): -1.46633706879, (do not): -1.70474809224, (not like): -1.70474809224, (like eggs): -1.70474809224, (eggs and): -1.70474809224, (and Sam): -1.70474809224

3.

**Top 20 conditional probabilities from moviereview.txt:** <s> and </s> were counted as tokens and no removals of punctuation were made.

(</s> <s>): -1.37886684529, (. </s>): -1.51508105958, (. .): -3.20886012045, (of the): -3.32283443666 (<s> the): -3.37103037886, (, but): -3.65699347802, (<s> a): -3.67945393562, (, and): -3.70655763603, (in the): -3.87865339042, (is a): -4.05347434123, (, the): -4.24904454171 (the film): -4.32245090288, (of a): -4.36762207531, (to the): -4.40454756629, (to be): -4.4538186153, (and the): -4.50241737841, (in a): -4.52825232574, (<s> it's): -4.534767871, (it is): -4.65414707409, (the movie): -4.66641381123

4.

**Trained with bronte jane eyre.txt:**

**1-gram:** hypnotically dramatically disgusted direction aid shame immortals darkens deflated crooned long-cherished 2 perfect simulation tideless wire contend ghetto miller's dispersed single dejected snatch compels cynicism fillers towers exquisitely daytime sneakers deficiencies Cairngorm madame continuação overhearing unexpressed ganesh's jonathan spectacular belabour agitate stand-up unpleasantly pine-forests larry outmoded foundering complexion tirade defence hurricanes doggedness double-crosses block outshined walnut teen

**2-gram:** seek real till two stone dead ; remember you palette and retain it close , its banks , combined result mortifying to my subsequent occurrences with delight , comparatively Vain Poor , I “ wife , irregularities are very sordid the self-abandonment

**3-gram:** can’t this of have and the wits perhaps to-morrow , with two upper I ; but studios I my engulfed cast a the wished . abused worst that the , , excited has me of suddenly under burdens the the high was

**4-gram:** I here must : to the sea said there cold , sight delineated from look business room ? as good , the expressed unfortunately with do wishes

**5-gram:** You weak window he Rochester Miss well tell for him stile the woman