David Schulz

CS 4980 001

3/21/2022

**Homework 2**

1. Given the following bigram counts and probabilities from the Berkeley Restaurant Project corpus, compute the probability of the given sentences.
   1. I want Chinese food
      1. 0.25 \* 0.33 \* 0.0065 \* 0.52 \* 0.68 = 0.000189618
      2. 1.90 \* 10-4
   2. I want to eat Chinese food
      1. 0.25 \* 0.33 \* 0.66 \* 0.28 \* 0.021 \* 0.52 \* 0.68 = 0.0001132106976
      2. 1.13 \* 10-4
   3. I want to eat food
      1. 0.25 \* 0.33 \* 0.66 \* 0.28 \* 0.0027 \* 0.68 = 0.000027991656
      2. 2.80 \* 10-5
2. Now, using add-1 smoothing, recalculate the probabilities.
   1. I want Chinese food
      1. 0.19 \* 0.21 \* 0.0029 \* 0.052 \* 0.4 = 0.000002406768
      2. 2.41 \* 10-6
   2. I want to eat Chinese food
      1. 0.19 \* 0.21 \* 0.26 \* 0.18 \* 0.0078 \* 0.052 \* 0.4 = 0.0000003029539968
      2. 3.03 \* 10-7
   3. I want to eat food
      1. 0.19 \* 0.21 \* 0.26 \* 0.18 \* 0.0014 \* 0.4 = 0.0000010456992
      2. 1.05 \* 10-6
3. Which of the two probabilities you computed in the previous exercise is higher, unsmoothed or smoothed? Explain why.
   1. Unsmoothed because the bigrams that originally appeared at least once happen more often out of the total possible bigrams from the first word. In other words, the numerator of their probabilities stay the same while the denominator is less than the denominator of the smoothed probabilities.
4. Write a program to compute unsmoothed unigrams and bigrams.
   1. Now compare the statistics of the two corpora. What are the differences in the most common unigrams between the two? How about interesting differences in bigrams?
      1. Since the “Never Gonna Give You Up” lyrics use fewer unique words than “All-Star”, as well as the fact that it has more lyric repetition, its unigrams are generally higher values. The repetition also means that some of the bigrams have much higher values as well.