Natural Language Processing (CO3086)
Lab 3 - NLP 242

HO CHI MINH UNIVERSITY OF TECHNOLOGY

Vietnam National University Ho Chi Minh

Problem 1

Write out the equation for trigram probability estimation. Now write out all the non-zero trigram probabilities for the `I am Sam` corpus from

```
< s > I am Sam < /s >
< s > Sam I am < /s >
< s > I do not like green eggs and Sam < /s >
```

Problem 2

Given two tables

Table 1: Bigram probabilities for eight words

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Table 2: Add-one smoothed bigram probabilities for eight of the word

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.52	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Assume the additional Laplace smoothed probabilities $P(i \mid \langle s \rangle) = 0.19$ and $P(\langle /s \rangle \mid \text{food}) = 0.40$. Calculate the probability of the sentence `i want chinese food`. ($\langle s \rangle$ and $\langle /s \rangle$ are not smoothed.)

Problem 3

Which of the two probabilities you computed in the previous problem is higher, unsmoothed or smoothed? Explain why.

Problem 4

We are given the following corpus:

```
< s > I am Sam < /s >
< s > Sam I am < /s >
< s > I am Sam < /s >
< s > I do not like green eggs and Sam < /s >
```

Using a bigram language model with add-one smoothing, what is $P(\text{Sam} \mid \text{am})$? Include $\langle s \rangle$ and $\langle /s \rangle$ in your counts just like any other token.

Problem 5

We are given the following corpus, modified from the one in the chapter:

```
< s > I am Sam < /s >
< s > Sam I am < /s >
< s > I am Sam < /s >
< s > I do not like green eggs and Sam < /s >
```

If we use linear interpolation smoothing between a maximum-likelihood bigram model and a maximum-likelihood unigram model with $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = \frac{1}{2}$, what is $P(\text{Sam} \mid \text{am})$? Include $\langle s \rangle$ and $\langle /s \rangle$ in your counts just like any other token.

Problem 6

You are given a training set of 100 numbers that consists of 91 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity?