

# Tutorial SNLP

January 30, 2020

## 1 Questions

Q1. Let the cost of insertion be 2, deletion be 1 and substitution be 3. Fill in the code snippet each of these cost in place of (A), (B), (C).

```
for i = 1,..., N do
  for j = 1,..., M do
    D(i, j) = min(D(i - 1, j) + (A) , D(i, j-1) + (B), D(i-1, j-1) + (C) )
  end for
end for
```

Find the edit distance between lead and deal.

Q2. While processing a corpus, you encounter a word “insention”. You recognize that this is a spelling error and also the possible candidate words are ‘insertion, inspection, invention, indention, intention’. Assuming that you are using a noisy channel model for spell correction, write down the expression for the correct word in terms of the probabilities to be estimated from a training corpus.

Q3. Corpus: b a b a a c b c a c a c

Vocabulary is  $\{a, b, c\}$

We are considering a bi-gram language model, and a Good-Turing method for probability estimation. Use GT probabilities up to and including  $r = 2$ .

Q4.

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>
```

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.

Q5. From a restaurant corpus, it was found that the number of unique unigrams is 1500 and the total number of bigrams is 8000. Suppose out of 8000, 5000 bigrams occur once in the corpus. Use Good Turing smoothing to estimate the effective count for the bigrams not seen in the corpus. And also calculate the Probability mass that will be allotted to them.

Q6. Given a corpus C2, The Maximum Likelihood Estimation (MLE) for the bigram “ice cream” is 0.4 and the count of occurrence of the word “ice” is 310. The likelihood of “ice cream” after applying add-one smoothing is 0.025, for the same corpus C2. What is the vocabulary size of C and what is the likelihood after applying add-3 smoothing?

Q7. Consider the following corpus of two sentences.  
One fish two fish red fish blue fish black fish blue fish  
Once I caught a fish alive. The red fish turned blue  
Assume you perform case-folding, punctuation removal and stop word removal on the corpus, before attempting any of the following questions.  
(a) Write the count of types and tokens in the corpus.  
(b) Show a table with the bigram counts (as in the text) for this corpus. Given this table, give the  $P(\text{fish}|\text{two})$  and  $P(\text{black}|\text{fish})$ .  
(c) Compute the probability mass that Good-Turing would assign to zero count bigrams, given this corpus.

Q8.

For the following corpus of two documents:

1. how much wood would a wood chuck chuck if a wood chuck would chuck wood
2. a wood chuck would chuck the wood he could chuck if a wood chuck would chuck wood

Which of the following sentences:

1. a wood could chuck;
2. wood would a chuck;

is more probable, according to:

1. An unsmoothed uni-gram language model?
2. A uni-gram language model, with Laplacian (“add-one”) smoothing?
3. An unsmoothed bigram language model?
4. A bi-gram language model, with Laplacian smoothing?
5. An unsmoothed tri-gram language model?
6. A tri-gram language model, with Laplacian smoothing?

Q9.

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 bi-gram model and a maximum-likelihood unigram model with  $\lambda_1 = \frac{1}{2}$  and  $\lambda_2 = \frac{1}{2}$ , what is  $P(\text{Sam}|\text{am})$ ? Include `<s>` and `</s>` in your counts just like any other token.

Q10. What does back-off mean, in the context of smoothing a language model? What does interpolation refer to? Why do we usually use log probabilities when finding the probability of a sentence according to an n-gram language model?

Q11. The backoff probability for a trigram LM is defined as:

$$\begin{aligned} P_{bo}(w_n|w_{n-1}w_{n-2}) &= P'(w_n|w_{n-1}w_{n-2}) \text{ if } c(w_{n-2}w_{n-1}w_n) > 0. \\ &= \lambda(w_{n-1}w_{n-2})P_{bo}(w_n|w_{n-1}), \text{ otherwise.} \end{aligned}$$

Similarly,


$$P_{bo}(w_n|w_{n-1}) = P'(w_n|w_{n-1}) \text{ if } c(w_{n-1}w_n) \text{ if } c(w_{n-1}w_n) > 0.$$

$$= \lambda(w_{n-1})P'(w_n), \text{ otherwise.}$$

$$P'(w_n|w_{n-1}) = P(w_n|w_{n-1}) - \frac{1}{8}$$

$$P'(w_n|w_{n-1}w_{n-2}) = P(w_n|w_{n-1}w_{n-2}) - \frac{1}{8}$$

Suppose your vocabulary contains a, b, c d and you are given the following counts:

<b>b</b>	<b>a</b>	<b>5</b>
<b>b</b>	<b>b</b>	<b>3</b>
<b>b</b>	<b>c</b>	<b>0</b> 
<b>b</b>	<b>d</b>	<b>0</b>

<b>ab</b>	<b>a</b>	<b>4</b>
<b>ab</b>	<b>b</b>	<b>0</b>
<b>ab</b>	<b>c</b>	<b>0</b>
<b>ab</b>	<b>d</b>	<b>0</b>

<b>a</b>	<b>8</b>
<b>b</b>	<b>9</b>
<b>c</b>	<b>8</b>
<b>d</b>	<b>7</b>

Construct the backoff LM  $P_{bo}(w|w_{n-2}w_{n-1})$ , where  $w_{n-1} = b$  and  $w_{n-2} = a$ .