Homework 2

Emmanuel Adebayo

February 2024

1 Problem 1

1. (a) $P(\text{Tom} \mid \text{am}) = \frac{\text{count (am tom)}}{\text{count}(am)} = \frac{3}{4}$

(b) $\frac{\mathrm{c} + 1}{N+V} = \frac{3+1}{4+12} = \frac{4}{16}$

(c) $\lambda P(\text{Tom}) + \lambda P(\text{Tom} \mid \text{am})$
$\lambda = \frac{1}{2}$
$P(\text{Tom}) = \frac{7}{42}$
$P(\text{Tom} \mid \text{am}) = \frac{3}{4}$
$\frac{1}{2}\frac{7}{42} + \frac{1}{2}\frac{3}{4}$
$= \frac{11}{24} = 0.4583333333$

(d) $P(\text{I} \mid \texttt{<s>}) * P(\text{like} \mid \text{I}) * P(\text{Jane} \mid \text{like}) * P(\texttt{</s>} \mid \text{Jane}) = \frac{5}{7} * \frac{2}{6} * \frac{1}{3} * \frac{2}{2} = \frac{5}{63} = 0.07936507937$

(e) $\frac{5+1}{7+12} * \frac{2+1}{6+12} * \frac{1+1}{3+12} * \frac{2+1}{2+12} = \frac{1}{665} = 0.001503759398$

(f) $P(\texttt{</s>}) = \frac{\texttt{count </s>}}{N} = \frac{7}{42}$

$P(\texttt{</s>}|jane) = \frac{\texttt{count (jane </s>)}}{\texttt{count (jane)}} = \frac{2}{2}$

$P(\texttt{</s>}|\texttt{like jane}) = \frac{\texttt{count (like jane </s>)}}{\texttt{count (like jane)}} = \frac{1}{1}$

$P(\texttt{</s>}|\texttt{i like jane}) = \frac{\texttt{count (i like jane </s>)}}{\texttt{count (i like jane)}} = \frac{0}{0}$

$P(\texttt{</s>}|\texttt{<s> i like jane}) = \frac{\texttt{count (<s> i like jane </s>)}}{\texttt{count (<s> i like jane)}} = \frac{0}{0}$

We use $\lambda = \frac{1}{5}$

The probability is then calculated as thus,

$\frac{1}{5}(\frac{7}{42} + \frac{2}{2} + \frac{1}{1} + \frac{0}{0} + \frac{0}{0}) = 0.4333333333$

2 problem 2

1. (a) The total number of non zero unigram for the news data: 12784
   The total number of non zero unigram for the romance data: 7817

(b) The total number of non zero bigram for the news data: 59712
The total number of non zero bigram for the romance data: 36362

(c)

```
10 most common unigrams (in terms of counts) from news data
dataset with their probabilities P (wt) (using MLE).
index - unigram - frequency - probability
1 the 6386 0.06527116253398475
2 <s> 4623 0.04725157914102905
3 </s> 4623 0.04725157914102905
4 of 2861 0.029242216725607638
5 and 2186 0.022343056889960956
6 a 2168 0.02215907929434371
7 to 2144 0.021913775833520718
8 in 2020 0.020646374619268586
9 for 969 0.009904127230728347
10 that 829 0.008473190375927553

10 most common unigrams (in terms of counts) from romance
data dataset with their probabilities P (wt) (using MLE).
index - unigram - frequency - probability
1 <s> 4431 0.06565903534118693
2 </s> 4431 0.06565903534118693
3 the 2988 0.04427650589019782
4 and 1905 0.02822849522116026
5 to 1517 0.022479069422834706
6 a 1383 0.020493442987330517
7 of 1202 0.017811365488627103
8 he 1068 0.015825739053122918
9 was 999 0.014803289619915536
10 i 951 0.014092020448988664
```

In the romance data, the pronouns ”he” and ”she exhibit notable prominence in terms of frequency and prevalence. That is not the case for news data

(d)

```
10 most common bigrams (in terms of counts) from news data
dataset with their probabilities P (wt) (using MLE).
index - bigram- frequency- probability
1 (’</s>’, ’<s>’) 4622 0.99978369024443
2 (’of’, ’the’) 849 0.2967493883257602
3 (’<s>’, ’the’) 780 0.16872160934458144
4 (’in’, ’the’) 589 0.29158415841584157
5 (’to’, ’the’) 277 0.12919776119402984
6 (’on’, ’the’) 253 0.3661360347322721
7 (’for’, ’the’) 220 0.22703818369453044
8 (’at’, ’the’) 196 0.3081761006289308
9 (’<s>’, ’he’) 192 0.04153147306943543
10 (’will’, ’be’) 157 0.40359897172236503

10 most common bigrams (in terms of counts) from romace data
dataset with their probabilities P (wt) (using MLE).
index - bigram- frequency- probability
1 (’</s>’, ’<s>’) 4430 0.9997743173098623
2 (’<s>’, ’i’) 386 0.08711351839313924
3 (’<s>’, ’he’) 372 0.08395396073121192
4 (’in’, ’the’) 273 0.29354838709677417
5 (’<s>’, ’she’) 244 0.05506657639359061
6 (’of’, ’the’) 235 0.19550748752079866
7 (’<s>’, ’the’) 230 0.051907018731663285
8 (’it’, ’was’) 179 0.2496513249651325
9 (’<s>’, ’it’) 154 0.03475513428120063
10 (’<s>’, ’but’) 144 0.03249830737982397
```

In the romance data, the pronouns ”he” and ”she exhibit notable prominence in terms of frequency and prevalence. That is not the case for news data

(e) the probability of ['<s>', 'i', 'loved', 'her', 'when', 'she', 'laughed', '</s>'] in news data is 0

(f) `the probability of ['<s>', 'i', 'loved', 'her', 'when', 'she', 'laughed', '</s>'] in romace data is 1.2443674813741955e-12`

(g)
```
the probability of [’<s>’, ’i’, ’loved’, ’her’, ’when’, ’she’,
’laughed’, ’</s>’] with laplace add-one smoothing in news
data is 3.277727206713981e-27

the probability of [’<s>’, ’i’, ’loved’, ’her’, ’when’, ’she’,
’laughed’, ’</s>’] with laplace add-one smoothing in romance
data is 3.407328911651892e-22
```