```
/Users/kieran/.local/share/virtualenvs/NU-DS5020-NLP-LlRfbsqa/bin/python /Users/kieran/Code/NU/NU-DS5020-NLP/src/main.py
[nltk_data] Downloading package punkt to /Users/kieran/nltk_data...
[nltk_data]   Package punkt is already up-to-date!

        PART 1
        ------

        Load and preprocess the dataset provided:
        - Tokenize the text, keeping only actual words while removing disfluencies such as “uh”
        and “uhm”
        - Add special tokens to indicate the beginning of each sentence



------ Sampled Raw Data - START------

33_1_0001 okay lets see i want to go to a thai restaurant .
[uh] with less than ten dollars per person
33_1_0002 <i> <like> <to> <eat> [uh] i like to eat at lunch time .
so that would be eleven a__m to one p__m
33_1_0003 i dont want to walk for more than five minutes
33_1_0004 tell me more about the [uh] na- nakapan [uh] restaurant on martin luther king
33_1_0005 i like to go to a hamburger restaurant
33_1_0006 lets start again
33_1_0007 i like to get a hamburger at an american restaurant
33_1_0008 id like to eat dinner .
and i dont mind walking [uh] .
for half an hour
33_1_0009 i dont want to spend more than [uh] ten dollars for a hamburger
33_1_0010 <(te)-ll> <me> <more> <about> <the> <two> <barbecue> <restaurants> tell me more about the two barbecue restaurants you listed
33_1_0011 tell me about everett and jones barbecue flints barbecue and the thai barbecue please
33_1_0012 wheres the best place to get soup in berkeley
33_1_0013 wheres the best place to get soup in berkeley for lunch for under ten dollars .

```

```
35 ------ END ------
36
37
38 ------ Sampled Processed Data - START------
39
40 ['</s>', 'okay', 'lets', 'see', 'i', 'want', 'to', 'go', 'to', 'a', 'thai', 'restaurant']
41 ['</s>', 'with', 'less', 'than', 'ten', 'dollars', 'per', 'person', 'i', 'like', 'to', 'eat', 'i', '
   like', 'to', 'eat', 'at', 'lunch', 'time']
42 ['</s>', 'so', 'that', 'would', 'be', 'eleven', 'to', 'one', 'i', 'dont', 'want', 'to', 'walk', 'for
   ', 'more', 'than', 'five', 'minutes', 'tell', 'me', 'more', 'about', 'the', 'nakapan', 'restaurant', '
   on', 'martin', 'luther', 'king', 'i', 'like', 'to', 'go', 'to', 'a', 'hamburger', 'restaurant', 'lets
   ', 'start', 'again', 'i', 'like', 'to', 'get', 'a', 'hamburger', 'at', 'an', 'american', 'restaurant
   ', 'id', 'like', 'to', 'eat', 'dinner']
43 ['</s>', 'and', 'i', 'dont', 'mind', 'walking']
44 ['</s>', 'for', 'half', 'an', 'hour', 'i', 'dont', 'want', 'to', 'spend', 'more', 'than', 'ten', '
   dollars', 'for', 'a', 'hamburger', 'te', 'me', 'more', 'about', 'the', 'two', 'barbecue', 'restaurants
   ', 'tell', 'me', 'more', 'about', 'the', 'two', 'barbecue', 'restaurants', 'you', 'listed', 'tell', 'me
   ', 'about', 'everett', 'and', 'jones', 'barbecue', 'flints', 'barbecue', 'and', 'the', 'thai', '
   barbecue', 'please', 'wheres', 'the', 'best', 'place', 'to', 'get', 'soup', 'in', 'berkeley', 'wheres
   ', 'the', 'best', 'place', 'to', 'get', 'soup', 'in', 'berkeley', 'for', 'lunch', 'for', 'under', 'ten
   ', 'dollars']
45
46 ------ END ------
47
48
49         PART 2
50         ------
51
52         - Count the words
53         - Report the size of the vocabulary
54         - report the number of sentences in the dataset
55
56
57 WORD COUNT - Total number of words in the dataset: 53476
58 VOCAB SIZE - Total number of unique words in the dataset: 1557
59 SENTENCE COUNT - Number of sentences in the dataset: 1055
```

```
60
61
62          PART 3
63          ------
64
65          Read the chapter on N-grams and generate figures 4.1 and 4.2
66          for bigram
67          counts. The figures do not have to be exact.
68
69
70 ------ Bigram Count Table - Figure 4.1 -------
71
72 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
73 |         |   i |   want |   to |   eat |   chinese |   food |   lunch |   spend |
74 +=========+=====+========+======+=======+===========+========+=========+=========+
75 | i       |   8 |    913 |    0 |    13 |         0 |      0 |       0 |       2 |
76 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
77 | want    |   3 |      0 |  677 |     0 |         7 |      6 |       6 |       1 |
78 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
79 | to      |  17 |      0 |    6 |   758 |         4 |      0 |       6 |     233 |
80 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
81 | eat     |   7 |      0 |    3 |     0 |        16 |      2 |      53 |       0 |
82 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
83 | chinese |   5 |      0 |    0 |     0 |         0 |    100 |       1 |       0 |
84 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
85 | food    | 264 |      1 |   19 |     1 |         2 |      4 |       0 |       0 |
86 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
87 | lunch   |  72 |      0 |    2 |     0 |         0 |      2 |       5 |       0 |
88 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
89 | spend   |   2 |      0 |    1 |     0 |         0 |      0 |       0 |       0 |
90 +---------+-----+--------+------+-------+-----------+--------+---------+---------+
91
92 ------ Bigram Probability Table - Figure 4.2 -------
93
94 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
95 |         |      i |   want |     to |    eat |   chinese |   food |   lunch |   spend |
```

```
 96 +=========+========+========+========+========+===========+========+=========+=========+
 97 | i       | 0.0028 | 0.3189 | 0.0000 | 0.0045 |    0.0000 | 0.0000 |  0.0000 |  0.0007 |
 98 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
 99 | want    | 0.0029 | 0.0000 | 0.6497 | 0.0000 |    0.0067 | 0.0058 |  0.0058 |  0.0010 |
100 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
101 | to      | 0.0062 | 0.0000 | 0.0022 | 0.2780 |    0.0015 | 0.0000 |  0.0022 |  0.0854 |
102 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
103 | eat     | 0.0084 | 0.0000 | 0.0036 | 0.0000 |    0.0192 | 0.0024 |  0.0637 |  0.0000 |
104 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
105 | chinese | 0.0259 | 0.0000 | 0.0000 | 0.0000 |    0.0000 | 0.5181 |  0.0052 |  0.0000 |
106 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
107 | food    | 0.2122 | 0.0008 | 0.0153 | 0.0008 |    0.0016 | 0.0032 |  0.0000 |  0.0000 |
108 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
109 | lunch   | 0.1837 | 0.0000 | 0.0051 | 0.0000 |    0.0000 | 0.0051 |  0.0128 |  0.0000 |
110 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
111 | spend   | 0.0065 | 0.0000 | 0.0032 | 0.0000 |    0.0000 | 0.0000 |  0.0000 |  0.0000 |
112 +---------+--------+--------+--------+--------+-----------+--------+---------+---------+
113
114
115         PART 4 & 5
116         ------
117
118         Calculate the joint probability for at least five sentences (with vocabulary in the dataset)
    using bigrams.
119         Repeat step 2 using trigrams. Observe if the estimates have changed.
120
121
122
123 ------ Calculate bigram and trigram probabilities --------
124
125 Sentence: twelve dollars i would like to spend between six to twelve dollars i will travel for ten
    minutes show me german restaurants german restaurants show me german restaurants german food please
    show me other italian restaurants within twenty minutes from the im interested in having dinner on
    friday change the cost to twenty dollars show me more about ristorante venezia show the list of the
    restaurants please list of restaurants please give me more information about caffe giovanni show
    italian restaurants please show italian foods show me caffe giovanni show caffe giovanni i want to
```

```
125 find out about taiwan restaurant lets find out more about some italian restaurants um do you have any
    information about any italian restaurant im looking for
126 Bigram Probability: 0.8518518518518519
127 Trigram Probability: 0.13043478260869565
128
129 Sentence: so that would be eleven to one i dont want to walk for more than five minutes tell me more
    about the nakapan restaurant on martin luther king i like to go to a hamburger restaurant lets start
    again i like to get a hamburger at an american restaurant id like to eat dinner
130 Bigram Probability: 0.021505376344086023
131 Trigram Probability: 1.0
132
133 Sentence: im willing to walk a mile one mile noise what is the maxim cafe tell me about the maxim cafe
     maxim how about the
134 Bigram Probability: 0.2835820895522388
135 Trigram Probability: 1.0
136
137 Sentence: i would not like to go on sunday only i would like to go on any day of the weekend give me a
     list of restaurants in
138 Bigram Probability: 0.25918153200419725
139 Trigram Probability: 0.002699055330634278
140
141 Sentence: soup kitchen heike start over laughter hi
142 Bigram Probability: 0.4166666666666667
143 Trigram Probability: 1.0
144
145
146 Process finished with exit code 0
147
```