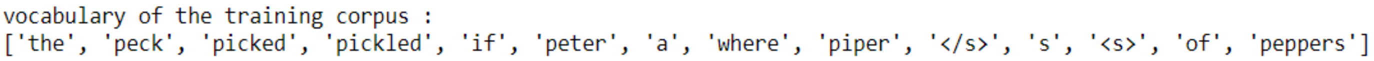
Language modelling

Aman Kumar 17025

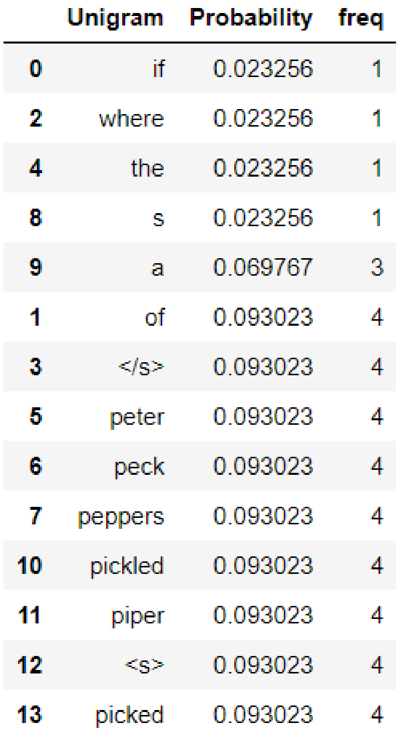
12th October 2020

1 Compute the vocabulary of the training corpus and print it.



```
vocabulary of the training corpus :
['the', 'peck', 'picked', 'pickled', 'if', 'peter', 'a', 'where', 'piper', '</s>', 's', '<s>', 'of', 'peppers']
```

2 In the above training corpus, calculate the probability of each unigram and print the 2 unigrams with the highest probability



	Unigram	Probability	freq
0	if	0.023256	1
2	where	0.023256	1
4	the	0.023256	1
8	s	0.023256	1
9	a	0.069767	3
1	of	0.093023	4
3	</s>	0.093023	4
5	peter	0.093023	4
6	peck	0.093023	4
7	peppers	0.093023	4
10	pickled	0.093023	4
11	piper	0.093023	4
12	<s>	0.093023	4
13	picked	0.093023	4

$$</s>$$

, peter , peck , peppers , pickled , piper ,

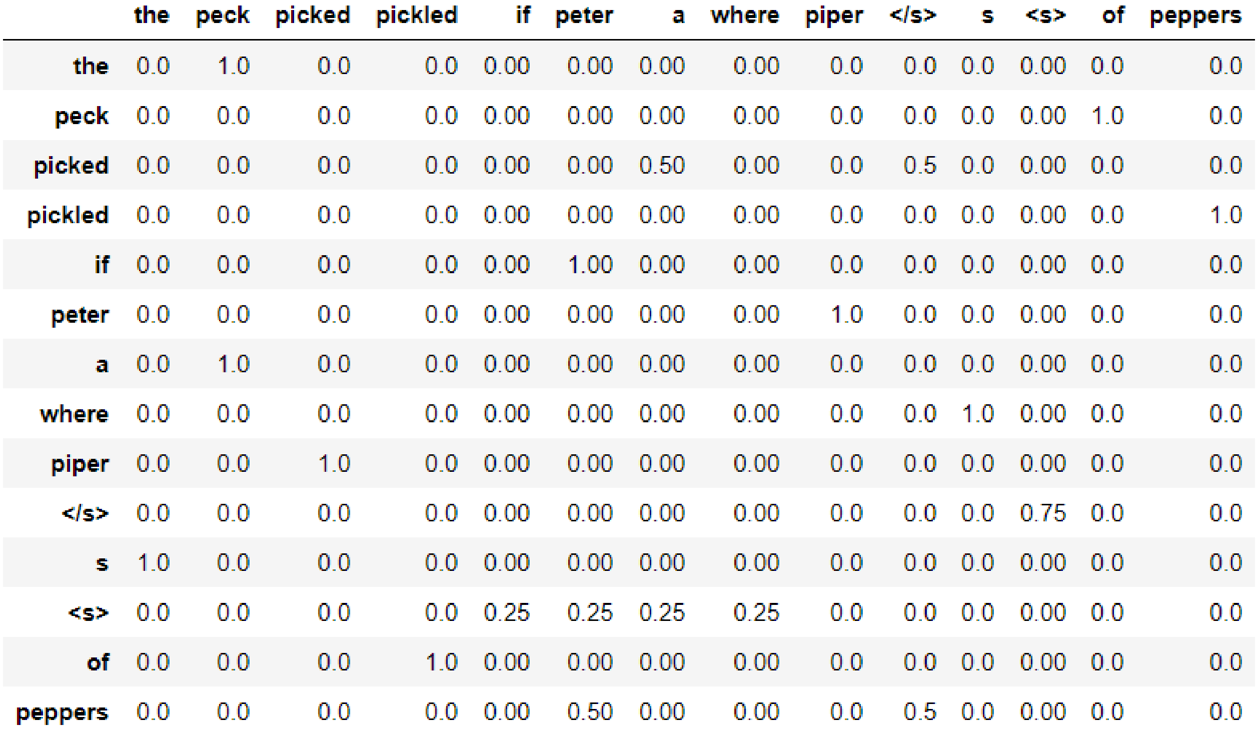
$$<s>$$

, picked all have same highest probability of 0.093023.

note some error due to latex syntax

3 Construct a probability matrix containing the maximum likelihood estimates (MLEs) of all possible bigrams and print it out

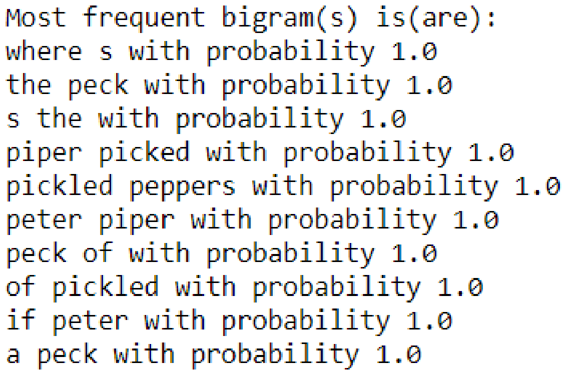
$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$



	the	peck	picked	pickled	if	peter	a	where	piper	</s>	s	<s>	of	peppers
the	0.0	1.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	0.0
peck	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	1.0	0.0
picked	0.0	0.0	0.0	0.0	0.00	0.00	0.50	0.00	0.0	0.5	0.0	0.00	0.0	0.0
pickled	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	1.0
if	0.0	0.0	0.0	0.0	0.00	1.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	0.0
peter	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	1.0	0.0	0.0	0.00	0.0	0.0
a	0.0	1.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	0.0
where	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	1.0	0.00	0.0	0.0
piper	0.0	0.0	1.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	0.0
</s>	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.75	0.0	0.0
s	1.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	0.0
<s>	0.0	0.0	0.0	0.0	0.25	0.25	0.25	0.25	0.0	0.0	0.0	0.00	0.0	0.0
of	0.0	0.0	0.0	1.0	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	0.0
peppers	0.0	0.0	0.0	0.0	0.00	0.50	0.00	0.00	0.0	0.5	0.0	0.00	0.0	0.0

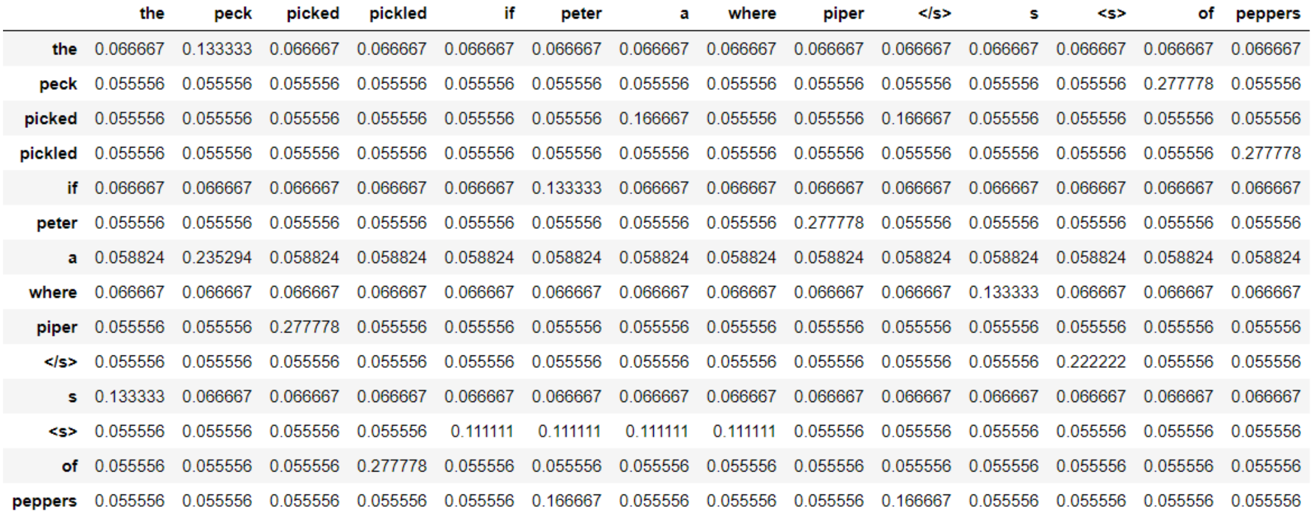
Figure 1: **Maximum Likelihood Estimate matrix**

4 What is the most frequent bigram in this corpus



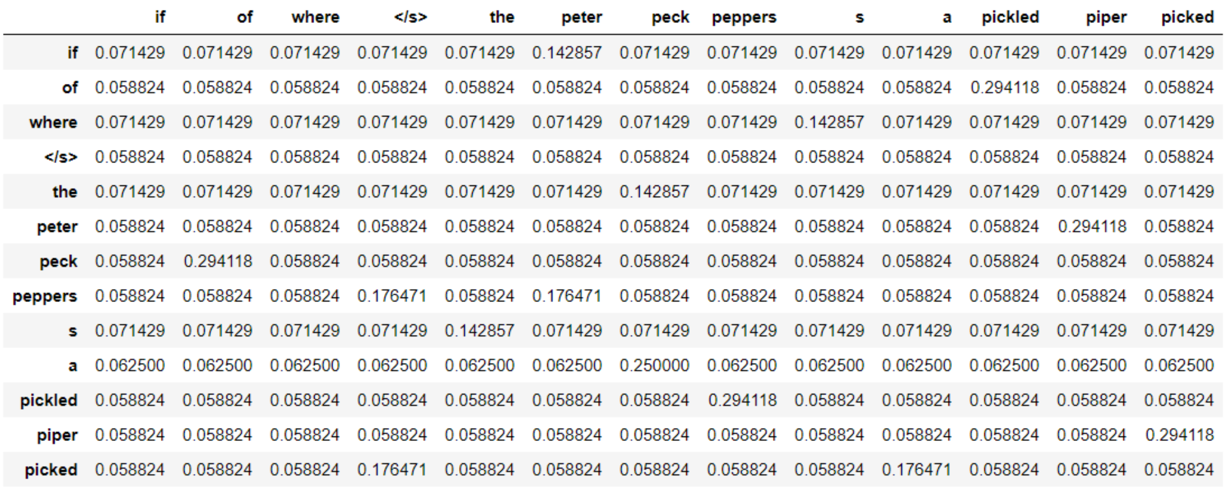
```
Most frequent bigram(s) is(are):
where s with probability 1.0
the peck with probability 1.0
s the with probability 1.0
piper picked with probability 1.0
pickled peppers with probability 1.0
peter piper with probability 1.0
peck of with probability 1.0
of pickled with probability 1.0
if peter with probability 1.0
a peck with probability 1.0
```

5 Construct a probability matrix containing the Laplace smoothed estimates of all possible bigrams and print it out.



	the	peck	picked	pickled	if	peter	a	where	piper	</s>	s	<s>	of	peppers
the	0.066667	0.133333	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667
peck	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.277778	0.055556
picked	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.166667	0.055556	0.055556	0.166667	0.055556	0.055556	0.055556	0.055556
pickled	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.277778
if	0.066667	0.066667	0.066667	0.066667	0.066667	0.133333	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667
peter	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.277778	0.055556	0.055556	0.055556	0.055556	0.055556
a	0.058824	0.235294	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824
where	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.133333	0.066667	0.066667	0.066667
piper	0.055556	0.055556	0.277778	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556
</s>	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.222222	0.055556	0.055556
s	0.133333	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667	0.066667
<s>	0.055556	0.055556	0.055556	0.055556	0.111111	0.111111	0.111111	0.111111	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556
of	0.055556	0.055556	0.055556	0.277778	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556	0.055556
peppers	0.055556	0.055556	0.055556	0.055556	0.055556	0.166667	0.055556	0.055556	0.055556	0.166667	0.055556	0.055556	0.055556	0.055556

.



	if	of	where	</s>	the	peter	peck	peppers	s	a	pickled	piper	picked
if	0.071429	0.071429	0.071429	0.071429	0.071429	0.142857	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429
of	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.294118	0.058824	0.058824
where	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429	0.142857	0.071429	0.071429	0.071429	0.071429
</s>	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824
the	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429	0.142857	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429
peter	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.294118	0.058824
peck	0.058824	0.294118	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824
peppers	0.058824	0.058824	0.058824	0.176471	0.058824	0.176471	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824
s	0.071429	0.071429	0.071429	0.071429	0.142857	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429	0.071429
a	0.062500	0.062500	0.062500	0.062500	0.062500	0.062500	0.250000	0.062500	0.062500	0.062500	0.062500	0.062500	0.062500
pickled	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.294118	0.058824	0.058824	0.058824	0.058824	0.058824
piper	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.058824	0.294118
picked	0.058824	0.058824	0.058824	0.176471	0.058824	0.058824	0.058824	0.058824	0.058824	0.176471	0.058824	0.058824	0.058824

Figure 2: **Laplace Smooothed Estimate matrix**

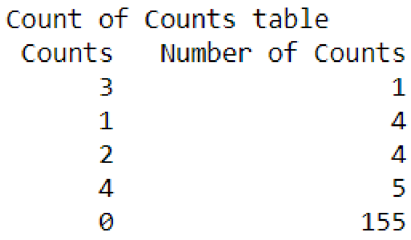
The two matrix corresponding to different vocabulary that is the first cantain all the vovab in part 1 but in 2nd matrix start of sentence tag is not in consider.

6 What is the probability of an unseen bigram obtained after using Laplace smoothing?

After Laplace smoothing techniques, the probabilities of unseen bigram which were previously 0 becomes now non zero. for more detail look at refrences below

see **figure 1** and **figure 2**

7 Construct and show a count-of-counts table for the bigrams.



```
Count of Counts table
 Counts   Number of Counts
      3                  1
      1                  4
      2                  4
      4                  5
      0                155
```

Here ,

- 155 numbers of bigram whose count is 0
- 4 numbers of bigram whose count is 1
- 4 numbers of bigram whose count is 2
- 1 numbers of bigram whose count is 3
- 5 numbers of bigram whose count is 4

8 Briefly explain in your own words the need for smoothing the count-of-counts table

We perform smoothing on the count of counts table in order to account for unseen words to make the count of count zero to some non zero number which maybe close to zero .

9 What is the probability of the above test sentence using the following bigram models you created? a. MLE bigram model b. Laplace smoothed bigram model

MLE of the given sentence: 0.0

Laplace estimate of the given sentence: 4.952083202745067e-05

.

I am attaching two inclass activty to have good idea of whats calculation is going on.

In-class Activity - 6 Aman Kumar

language modelling problem 17025

Training data:

<s> I am Sam </s>
<s> Sam I am </s>
<s> Sam I like </s>
<s> Sam I do like </s>
<s> do I like Sam </s>

Here are calculation of Bigram Probabilities from this corpus.

$P(I \mid \langle s \rangle) = 1/5$
$P(sam \mid \langle s \rangle) = 3/5$
$P(do \mid \langle s \rangle) = 1/5$

$P(am \mid I) = 2/5$
$P(I \mid sam) = 3/5$
$P(I \mid do) = 1/2$

$P(sam \mid am) = 1/2$
$P(like \mid I) = 2/5$
$P(do \mid I) = 1/5$

$P(like \mid do) = 1/2$

$P(\langle /s \rangle \mid sam) = 2/5$
$P(\langle /s \rangle \mid am) = 1/2$
$P(\langle /s \rangle \mid like) = 2/3$

1. **Using a bigram language model based on the above training data, which of the following sentences is better, i.e., gets a higher probability?**

(a) <s> Sam I do I like </s>
(b) <s> Sam I am </s>

(a) $P(\text{Sam} \mid \text{<s>}) \cdot P(\text{I} \mid \text{sam}) \cdot P(\text{do} \mid \text{I}) \cdot P(\text{I} \mid \text{do})$

$\cdot P(\text{like} \mid \text{I}) \cdot P(\text{</s>} \mid \text{like})$

$$= \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{2}{5} \cdot \frac{2}{3} = \frac{9}{125} \cdot \frac{2}{15}$$

(b) $P(\text{sam} \mid \text{<s>}) \cdot P(\text{I} \mid \text{sam}) \cdot P(\text{am} \mid \text{I})\ P(\text{</s>} \mid \text{am})$

$$= \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{2} = \frac{9}{125}$$

Hence according to language modelling (b) is more probable (better) □

2. Consider again the same training data and the same bigram model. Compute the perplexity of the test sentence below:

<s> I do like Sam

Note: $H(W) = -\frac{1}{N} \log P(w_1 w_2 \ldots w_n)$

$$\text{Perplexity}(W) = 2^{H(W)} = \sqrt[N]{\frac{1}{P(w_1, w_2 \ldots w_n)}}$$

for bigram.

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i \mid w_{i-1})}}$$

Now,

for this sequence.

$$\prod_{i=1}^{N} (P_{w_i} \mid w_{i-1}) = \frac{1}{5} \cdot \frac{1}{5} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{150}$$

Perplexity (<s> I do like Sam) $= \sqrt[4]{150} = 3.5$

3. Now use a bigram LM with Laplace smoothing. Give the bigram probabilities estimated by this model:

$$P_{Laplace}(w_i \mid w_{i-1}) = \frac{[C(w_{i-1} w_i) + 1]}{C(w_{i-1}) + V}$$

Note. $|V| = 6$ (size of vocabalary)

(a) $P(do \mid <s>) = \frac{2}{11}$

(b) $P(do \mid sam) = \frac{1}{11}$

(c) $P(I \mid Sam) = 4/11$

(d) $P(I \mid do) = 2/8$

(e) $P(sam \mid <s>) = 4/11$

(f) $P(like \mid I) = 3/11$

(g) $P(sam \mid do) = 1/8$

4. Calculate the probabilities of the following sequences according to this model:

<s> do Sam I like $= \frac{2}{11} \cdot \frac{1}{8} \cdot \frac{4}{11} \cdot \frac{3}{11}$

<s> Sam do I like $= \frac{4}{11} \cdot \frac{1}{11} \cdot \frac{2}{8} \cdot \frac{3}{11}$

Both sequence are Equally probable. □