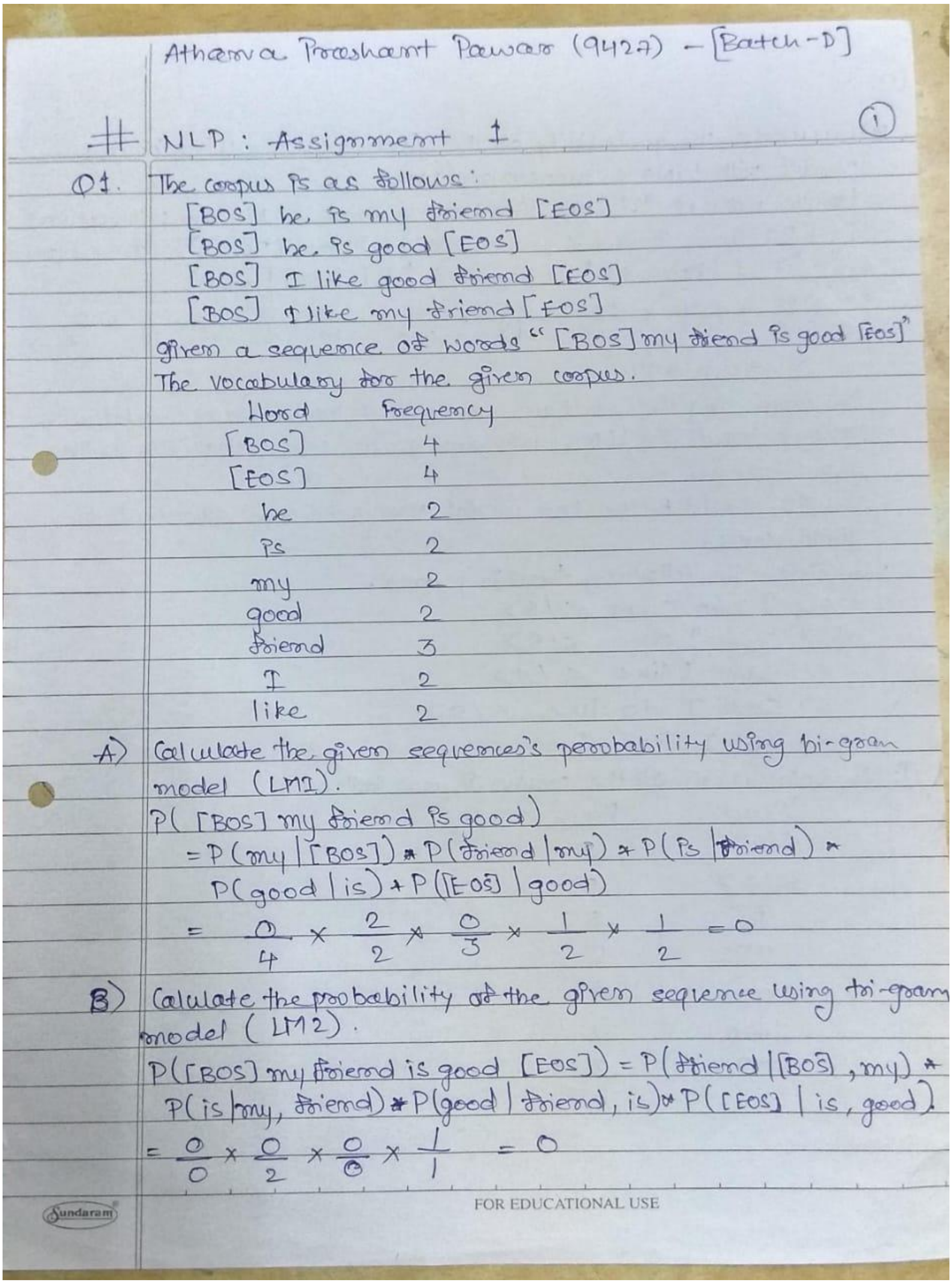
NLP Assignment - 1

Atharva Pawar - 9427



Atharva Prashant Pawar (9427) — [Batch-D]

NLP : Assignment 1

Q1. The corpus is as follows:

[BOS] he is my friend [EOS]
[BOS] he is good [EOS]
[BOS] I like good friend [EOS]
[BOS] I like my friend [EOS]

given a sequence of words "[BOS] my friend is good [EOS]"

The vocabulary for the given corpus.

Word	Frequency
[BOS]	4
[EOS]	4
he	2
is	2
my	2
good	2
friend	3
I	2
like	2

A) Calculate the given sequences's probability using bi-gram model (LM1).

P([BOS] my friend is good)
$= P(my \mid [BOS]) * P(friend \mid my) * P(is \mid friend) * P(good \mid is) * P([EOS] \mid good)$

$= \frac{0}{4} \times \frac{2}{2} \times \frac{0}{3} \times \frac{1}{2} \times \frac{1}{2} = 0$

B) Calulate the probability of the given sequence using tri-gram model (LM2).

$P([BOS]\ my\ friend\ is\ good\ [EOS]) = P(friend \mid [BOS], my) * P(is \mid my, friend) * P(good \mid friend, is) * P([EOS] \mid is, good)$

$= \frac{0}{0} \times \frac{0}{2} \times \frac{0}{0} \times \frac{1}{1} = 0$

(2)

c) Calculate the probability of the given sequence using bi-gram model with Laplace smoothing (LM3)

Unique words : 8...(since [BOS] never comes in bigram calculation)

P([BOS] my friend is good [EOS]) = P(my | [BOS]) * P(friend | my) * P(is | friend) * P(good | is) * P([EOS] | good)

$$= \frac{0+1}{4+8} * \frac{2+1}{2+8} \times \frac{0+1}{3\times8} \times \frac{1+1}{2+8} * \frac{1+1}{4+8} = \frac{1}{12} \times \frac{3}{10} \times \frac{1}{11} \times \frac{2}{10} \times \frac{2}{12}$$

= 0.0000767.

On comparing the result we can see that the bi-gram model with laplace smoothing (LM3) provides a non-zero probability for the given sequence.

∴ LM3 would be the best model among the tree options of the given corpus

* Consider the following 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>

(1) The vocabulary of the corpus is as follows :

Word	Freq.
<s>	5
</s>	5
I	5
am	2
Sam	5
like	3
do	2

Probability matrix for the bigram model:

	<s>	</s>	I	am	sam	like	do
<s>	0	0	0.2	0	0.6	0	0.2
</s>	0	0	0	0	0	0	0
I	0	0	0	0.4	0	0.4	0.2
am	0	0.5	0	0	0.5	0	0
sam	0	0.4	0.6	0	0	0	0
like	0	0.66	0	0	0.33	0	0
do	0	0	0.5	0	0	0.5	0

(2) Predict the most probable word:

a. <s> sam ...
The word with the highest probability after sam ie 'I'
(Probability = 0.6)

b. <s> sam I do ...
The words with highest & equal probability after do
are 'I' & 'like' (Probability = 0.5).

c. <s> Sam I am Sam ...
The word with highest probability is 'I'
(Probability = 0.6)

d. <s> do I like
The most probable word with highest probability afte
like is </s> (Probability = 0.66).

(3) Which of the following sentence is probable:

a. <s> Sam I do I like </s>

P(<s> Sam I do I like </s> = P (sam | <s>) * P(I | Sam
*P(do | I) * P (I | do) * P (like | I) * P (</s> | like)
= 0.6 * 0.6 * 0.2 * 0.5 * 0.4 * 0.66
= 0.0095.

b. <s> Sam I am </s>

$P(\langle S\rangle \text{ Sam I am } \langle /S\rangle) = P(\text{sam} \mid \langle s\rangle) * P(I \mid \text{sam})$

$* P(\text{am} \mid I) * P(\langle /s\rangle \mid \text{am})$

$= 0.6 * 0.6 * 0.4 * 0.5$

$= 0.072.$

c. <s> I do like sam I am </s>

$P(\langle s\rangle \text{ I do like sam I am } \langle /s\rangle)$

$= P(I \mid \langle s\rangle) * P(\text{do} \mid I) * P(\text{like} \mid \text{do}) * P(\text{sam} \mid \text{like})$

$* P(I \mid \text{sam}) * P(\text{am} \mid I) * P(\langle /s\rangle \mid \text{am})$

$= 0.2 * 0.2 * 0.5 * 0.33 * 0.6 * 0.4 * 0.5$

$= 0.00079$

∴ Most Probable sentence is <s> Sam I am </s>

Q3> Use the same corpus as given in Q2. This time we a bi-gram LM with Laplace Smoothing (Unique words:6)

1) a. $P(\text{do} \mid I) = \frac{1+1}{5+6} = \frac{2}{11} = 0.181.$

b. $P(\text{do} \mid \text{sam}) = \frac{0+1}{5+6} = \frac{1}{11} = 0.091$

c. $P(\text{sam} \mid \langle s\rangle) = \frac{3+1}{5+6} = \frac{4}{11} = 0.364$

d. $P(\text{sam} \mid \text{do}) = \frac{0+1}{2+6} = \frac{1}{8} = 0.125$

e. $P(I \mid \text{sam}) = \frac{3+1}{5+6} = \frac{4}{11} = 0.36$

f. $P(I \mid \text{do}) = \frac{1+1}{2+6} = \frac{2}{8} = 0.25$

g. $P(\text{like} \mid I) = \frac{2+1}{5+6} = \cancel{8}\,\frac{3}{11} = 0.273$

2) <s> do sam I like.

P(<s> do sam I like) = P (do|<s>) * P (sam|do) * P(I|sam) * P (like |I)

$= \frac{1+1}{5+6} * \frac{0+1}{2+6} * \frac{3+1}{5+6} * \frac{2+1}{5+6} = 0.00225$

<s> Sam do I like

P(<s> Sam do I like) = P(Sam |<s>) * P (do|sam) * P(I|do) * P(like |I)

$= \frac{3+1}{5+6} * \frac{0+1}{2+6} * \frac{1+1}{2+6} * \frac{2+1}{5+6} = 0.00226$

∴ Most probable sentence is <s> Sam do I like.

Q4) Apply Levensthten minimum edit distance algo. to find the similarity b/w two words 'Honda' & '~~[illegible]~~' hundai'

	#	h	u	n	d	a	i
#	0	1	2	3	4	5	6
h	1	0	1	2	3	4	5
o	2	1	1	2	3	4	5
n	3	2	2	1	2	3	4
d	4	3	3	2	1	2	3
a	5	4	4	3	2	1	2

∴ The minimum edit distance is 2

h	o	n	d	a	x
|	|R	|	|	|	|I
h	u	n	d	a	i

* Operations:

R(o, u) ; I(i)