CSCE5290 Assessment 1 Name: Harshavardhan Reddy Mallannagari
Instructor: Beddhu Murali

1) **What is Bag of words Representation? Represent the string below in bag of word representation.**

   the cat can drink milk in the bowl cat bowl is in window

Ans : Bag of words representation is one of the most common methods to represent text data in machine learning and natural language processing.

This representation has the set of words and the count of these words.

the: 2, cat: 2, can: 1, drink: 1, milk: 1, in: 2, bowl: 2, is: 1, window: 1

the	**2**
cat	**2**
can	**1**
drink	**1**
milk	**1**
in	**2**
bowl	**2**
is	**1**
window	**1**

But when we remove the basic English words such as is,in etc., which doesn't have any influence. After removing them the below is the Bag of words representation

---

cat: 2, can: 1, drink: 1, milk: 1, bowl: 2, window: 1

---

2) **a) Discuss at least one appropriate method suitable for designing a system to identify spam emails**

Ans: One appropriate method for designing a system to identify spam emails is **Naive Bayes** algorithm.

Naive Bayes is a probabilistic algorithm that can be used for text classification task, which includes spam detection. This algorithm works using on Bayes' theorem. This theorem states that the probability can be updated based on new evidence. One of advantage of Naive Bayes algorithm is that it is simple and faster(speed). This algorithm is efficient. Using this algorithm we can work on large datasets. Naive Bayes also works well with sparse data, which is common in text classification tasks.

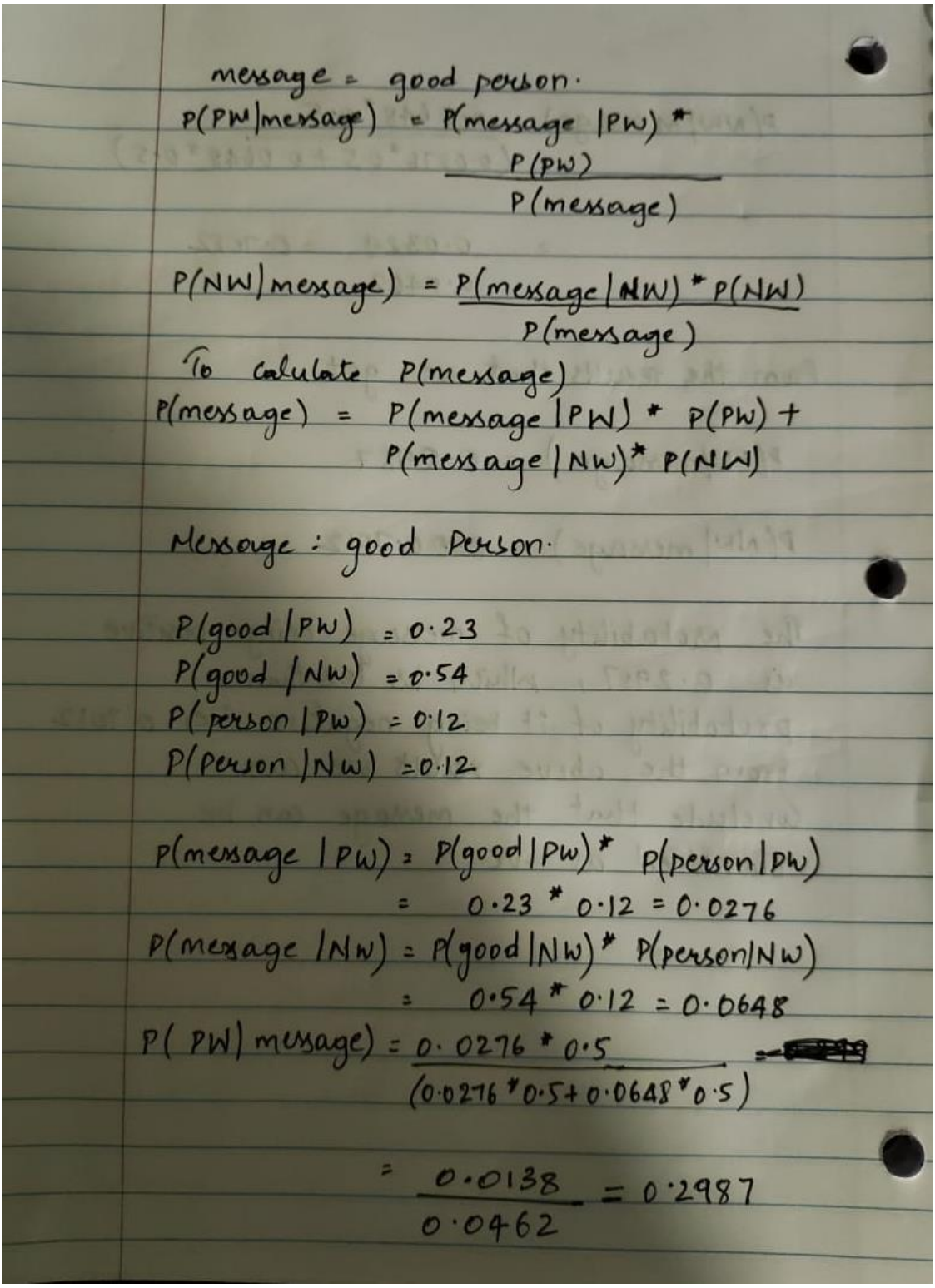
b) This type of problem in machine learning falls under the name **Classification** (Classification/ Regression) problems.

---

3) In a naive Bayes problem let,
P(PW) - probability of positive sentiment = 0.5
P(NW) – probability of negative sentiment = 0.5
Also, the following conditional probabilities are known.

P(you| PW)	**0.18**	**P(you| NW)**	**0.3**
P(good| PW)	**0.23**	**P(good| NW)**	**0.54**
P(are| PW)	**0.47**	**P(are| NW)**	**0.04**
P(person | PW)	**0.12**	**P(person | NW)**	**0.12**

Identify whether the following message would be classified as positive or negative using the above values. Show the details of your conclusion.
Message: **good person**



message = good person.

$$P(PW|message) = \frac{P(message|PW) * P(PW)}{P(message)}$$

$$P(NW|message) = \frac{P(message|NW) * P(NW)}{P(message)}$$

To calculate P(message)

$$P(message) = P(message|PW) * P(PW) + P(message|NW) * P(NW)$$

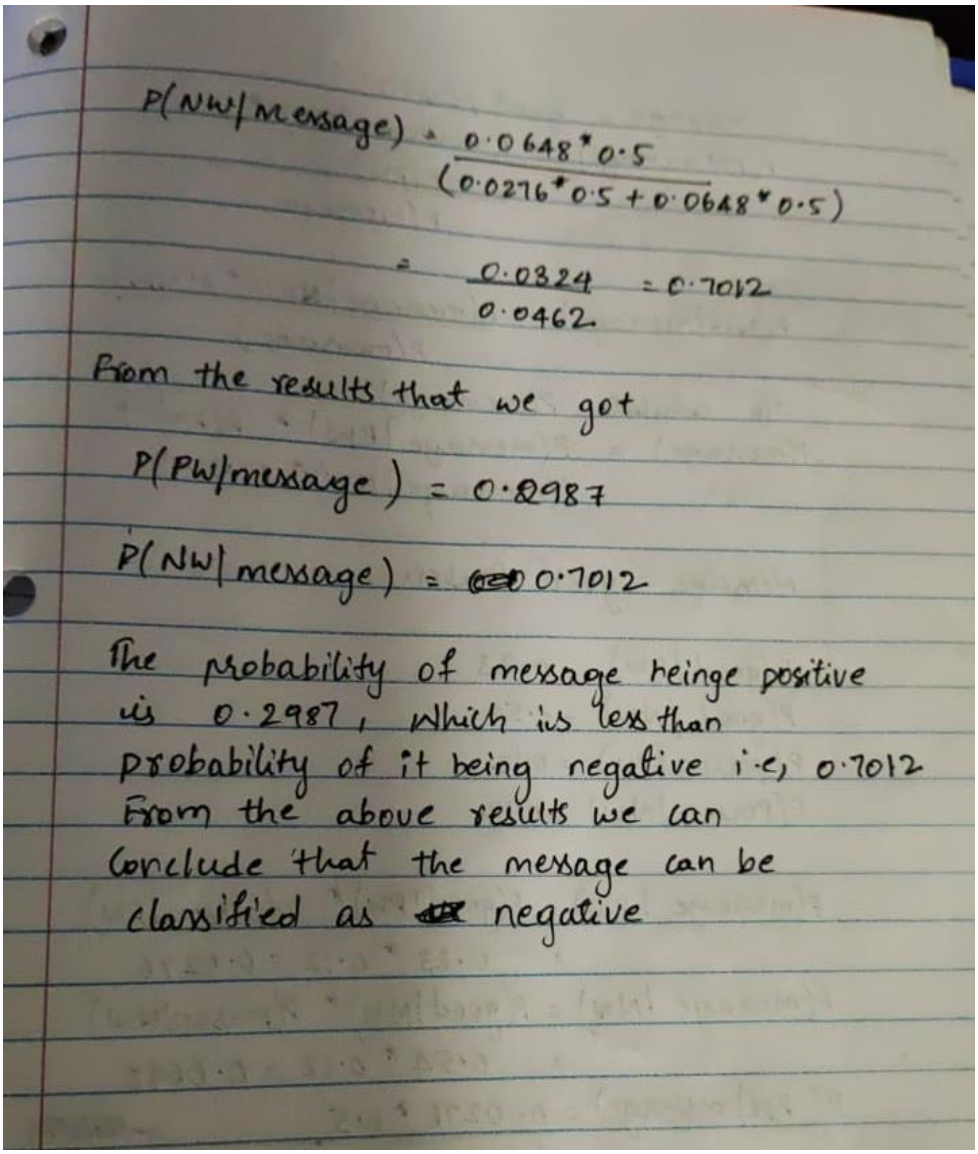
Message : good Person.

$P(good|PW) = 0.23$
$P(good|NW) = 0.54$
$P(person|PW) = 0.12$
$P(Person|NW) = 0.12$

$P(message|PW) = P(good|PW) * P(person|PW)$
$= 0.23 * 0.12 = 0.0276$
$P(message|NW) = P(good|NW) * P(person|NW)$
$= 0.54 * 0.12 = 0.0648$

$$P(PW|message) = \frac{0.0276 * 0.5}{(0.0276 * 0.5 + 0.0648 * 0.5)}$$

$$= \frac{0.0138}{0.0462} = 0.2987$$



---

4) **Suppose one or more words in the test data is not present in the training data. Considering the naïve Bayes model,**
   a) **Can you classify the test data?**

Ans: In the Naïve Bayes model, if one or more words in the test data is not present, it means that the that the probability of that word given each class is zero. This causes a problem when we calculate the probability of each class, it will result in zero. This will make it impossible to compare the probabilities and classify the test data.

When working, it is possible to classify the data even when if one or more words are not present in the training data.

b) Discuss at least one method to handle this situation.

Ans:

Method 1 : Ignore the unknown words and calculate the probabilities using known words. This method is called as "Naive Bayes with add-one smoothing" or "Laplace smoothing", where a small constant value is added to each count to avoid zero probabilities.

Method 2 : The method 2 is known as TF-IDF method.

TF-IDF : Term Frequency-Inverse Document Frequency

This method can also be used to handle unknown words. The TF-IDF weighting technique gives each word a score depending on its frequency in the text and the word's inverse frequency in the corpus.

This will give less weights to unknown words and prevents them from dominating the classification decision.

We can also use a combination of both the techniques.

---

5) What is minimum edit distance? Why is it needed?

Ans: Minimum edit distance is the minimum number of operations that are required to convert one string into other string. There are different operations such as

1. Insertion
2. Deletion
3. Substitution/Replacing

Minimum edit distance is needed in various tasks such as

1. Spell correction
2. Text recognition

It helps in finding best suitable word in the given sentence.

---

6) Calculate the minimum edit distance for the following cases. Transform X into Y. The cost for each operation is 1.

a) X= ab Y= abcd

Inserting 'c' at position 3 of X. ----- Cost =1
Inserting 'd' at position 4 of X. ----- Cost=1
Min cost =2

b) X= IIINTENTION Y = EXECUTION

I I I N T E N **[** T I O N **]**
E X E C U **[** T I O N **]**

Deleting 'I' at position 1 of X. ----- Cost =1
Deleting 'I' at position 2 of X. ----- Cost =1
Replacing 'I' with 'E' at position 3 of X. ----- Cost =1
Replacing 'N' with 'x' at position 4 of X. ----- Cost =1
Replacing 'T' with 'E' at position 5 of X. ----- Cost =1
Replacing 'E' with 'C' at position 6 of X. ----- Cost =1
Replacing 'N' with 'U' at position 7 of X. ----- Cost =1

Min cost =7.

---

7) Using the bigram model, and with the following training corpus:
<s> I Love NLP </s>
<s> NLP l Love </s>
<s> I like NLP and I like python</s>

Find the probability of

a) P(Love | I)

P(Love | I) = count(I Love) / count(I)
Count(I) = 4
Count(I Love) = 2
P(Love | I) = count(I Love) / count(I) = 2 / 4 = 0.5

b) **P(I | NLP)**
P(I | NLP) = count(NLP I) / count(NLP)
Count(NLP I) = 1
Count(NLP) = 3
P(I | NLP) = count(NLP I) / count(NLP) = 1 / 3 = 0.3333

---

8) True or false

a) We will add a special end-of-word symbol for every word in the training corpus while performing BPE.
b) A language model usually needs labels for its training.
c) In regular expression, \d matches any decimal digits 0-9.
d) **In regular expression, $ matches end of line**.

A) True

B) For this statement we can consider 2 cases:

True

When we use machine learning models such as neural network based models(supervised learning) we require labeled data.

False

We do not require labeled data when working on unsupervised models like n-gram language models.

C) True

D) True

---

9) str = " The American Revolutionary War, also known as the Revolutionary War or American War of Independence, secured American independence from Great Britain. Fighting began on April 19, 1775, followed by the Lee Resolution on July 2, 1776, and the Declaration of Independence on July 4, 1776. The paragraph number 2345444."

a. Write a Python regex pattern to find years from str.

pattern = r"\b\d{4}\b"

b. **Write the name of the appropriate method of re (assume Python) that will help you to find all years**

```
import re
match = re.findall(pattern,str)
```

---

10) E**xplain "Precision" and "Recall" and provide their formula's.**

Precision = True Positives / (True Positives +False Positives)

Recall = True Positives / (True Positives + False Negatives)

Precision : Precision measures how many of the items that the model classified as positive are actually positive. Precision gives the accuracy of the model.

Recall : Recall measures how many of the actual positive items the model was able to retrieve. In other words, it measures the completeness of the positive predictions.