1. Answer the following to the best of your ability:(**10points**)

a) Define Corpus:

b) How might you make a corpus for the following problem: I want to be able to learn characteristics of a politician’s language

2. a) Describe briefly 4 difficulties with identifying word boundaries algorithmically?(**8points**)

i)

ii)

iii)

iv)

b) What is the possible differences in the following two implementations of a word identifier(**5points**)

tokens = nltk.word\_tokenize(sentence)

and

tokens = sentence.split(“ “)

c) Why do we use ‘tokens’ instead of ‘word’(**5points**)

3. With the following sentence “The Cat in the Hat”(**12points**)

a) List the Uni-grams

b) List the Bi-grams

c) List the Tri-grams

4. Answer the following about predictive models:(**10points**)

a) What is a backoff model?

b) Give an example of how a backoff may help your model.

5. Why do we need *sent\_tokenize\_list = sent\_tokenize(text)* in NLTK instead of just breaking sentences apart by punctuation?(**5points**)

6. Briefly explain Transformation Based Tagging and how it differs from Ngram tagging for Part-of\_Speech (**8points**)

7) Answer the following:(**12points**)

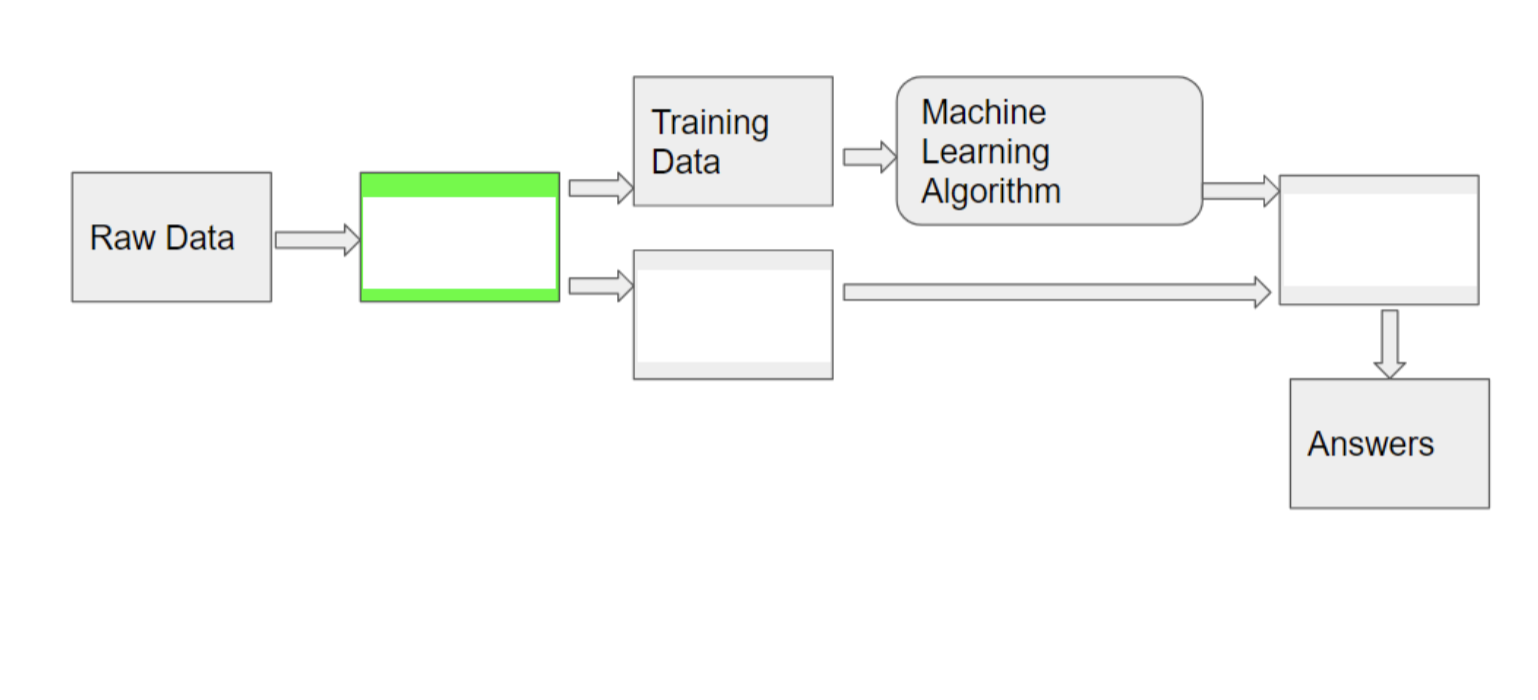
a)What is a False Negative?

b)What is a True Positive?

c)When should Accuracy be used as a metric?

d)What is the difference between Precision and Recall? When would you use them?

8. Fill in the 3 empty boxes for a typical machine learning cycle:(**9points**)



9. What are the two differences when you test on your training data versus testing on your test data? (**4points**)

10. Explain (or draw) k-fold validation when k=5 (**6points**)

11. Show 3 examples where a Named Entity System can get confused by ambiguity (**6points**)

**PART II**

1. Use given script to download 1 Wikipedia page
2. Run NLTK’s NER tool.
3. Examine both the PERSON and LOCATION Entities
4. Calculate Precision, Recall, and F-Measure for both Person or Location (whichever your document has)

pip install pymediawiki

To Harvest WIKI

from mediawiki import MediaWiki  
wikipedia = MediaWiki()  
  
p = wikipedia.page('Marymount University')  
content= p.content  
print(content)  
  
  
  
  
To Tag Named Entities  
1) Word Tokenize a Sentence  
2) POS tag the tokens  
3) print(nltk.ne\_chunk(tagged))