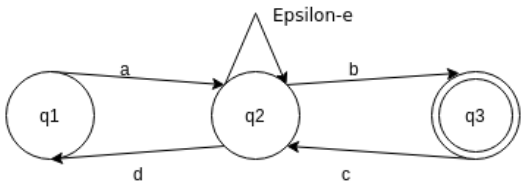
Previous year (2020) NLP questions for reference

Intro, Regex, Morphology, LM,

Quizzes

1. A book catalogue contains details of book in the following format: (3 marks)
   <Name of Book>,<author last name>,<author first name>
   Give a regex that matches all lines in which the author's first name is "Martin". For example, "High School English,Wren,Martin" should match.
   ^[^,]*,[^,]*,[^,]*\b(\Martin+)\b
   Note: This is one regular expression that matches the given pattern. There could also be other correct regular expressions. Therefore, the evaluation is not restricted to only the expression mentioned as the answer.
2. Write the output of the Porter Stemmer for the following words: (1 mark)
   A. Better
   B. Going
   C. Flying
   Better, Go, Fli
3. Consider the sentence: "learn this, learn that, learn all you can but also learn to unlearn."
   A. Give the count of tokens and types for the following sentence.(Specifically mention tokens = x and types = y) (1 mark)
      # Tokens = 16 (0.5 marks)
      # Types = 12 (0.5 marks)
   B. Give the number of bigrams. (1 mark)
      15
   C. Find the probability of 'that' given 'learn', considering a bigram model with Laplace smoothing. (2 marks)
      Ans. 1/7 = 0.1428
      Vocabulary, V = learn, this, that, all, you, can, but, also, to, unlearn
      |V| = 10
      P(that|learn) = (Count(learn that) + 1)/ (Count(learn) + |V|)
      = (1+1)/ (4+10) = 2/14 = 1/7 = 0.1428
4. For the following epsilon-NFA,
   a. How many states will be there in the equivalent minimal DFA (excluding dead state)? (1 mark)
      3
   b. Write regular expression defined by the equivalent minimal DFA (1 mark)



a(da)*b(c(da)*b)*

5. Identify the predicate and assign semantic roles in the following sentence. (2 marks)

   "John showed Bob his Toyota yesterday."

   Predicate: showed (1 mark)

   Semantic roles: Agent - John (0.25 marks)

   Patient/Receiver - Bob (0.25 marks)

   Instrument - Toyota (0.25 marks)

   Time - yesterday (0.25 marks)

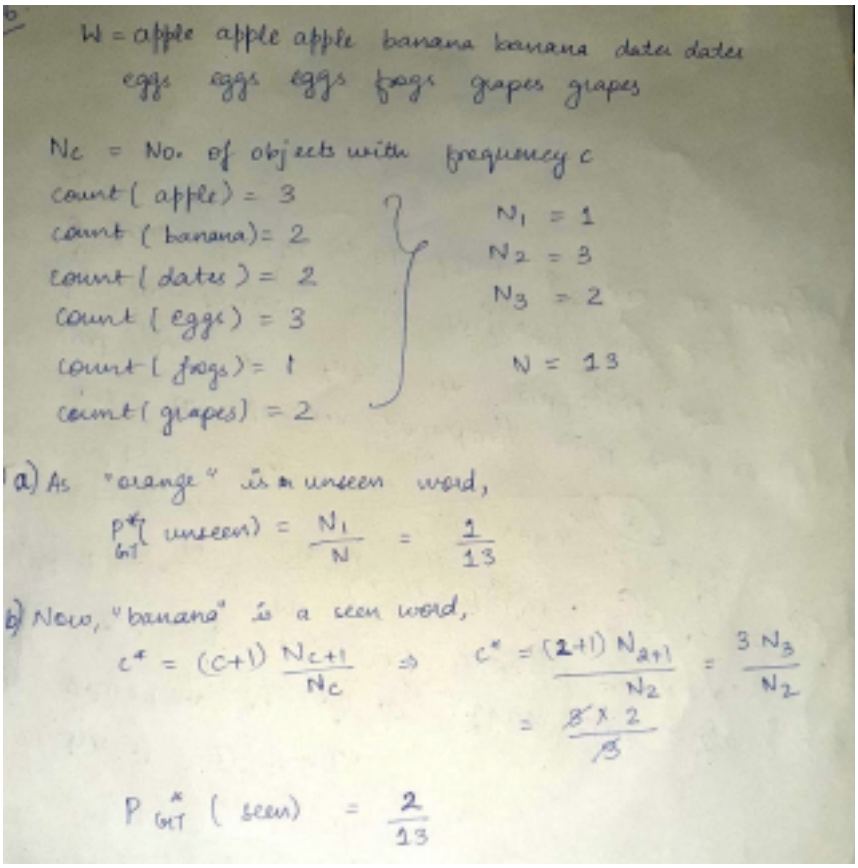
6. Give two examples of nominalization with different affixes (Note: affixes should not be "-ation", "-ness", "-er", and "-ee"). (2 marks)

   Intensity, Movement, …. and many more

7. Suppose we have a sequence of N independent samples: W = apple apple apple banana banana dates dates eggs eggs eggs frogs grapes grapes. Using Good Turing Smoothing, find the updated probability of

   a. drawing an orange. (1 mark)

   b. drawing a banana. (1 mark)



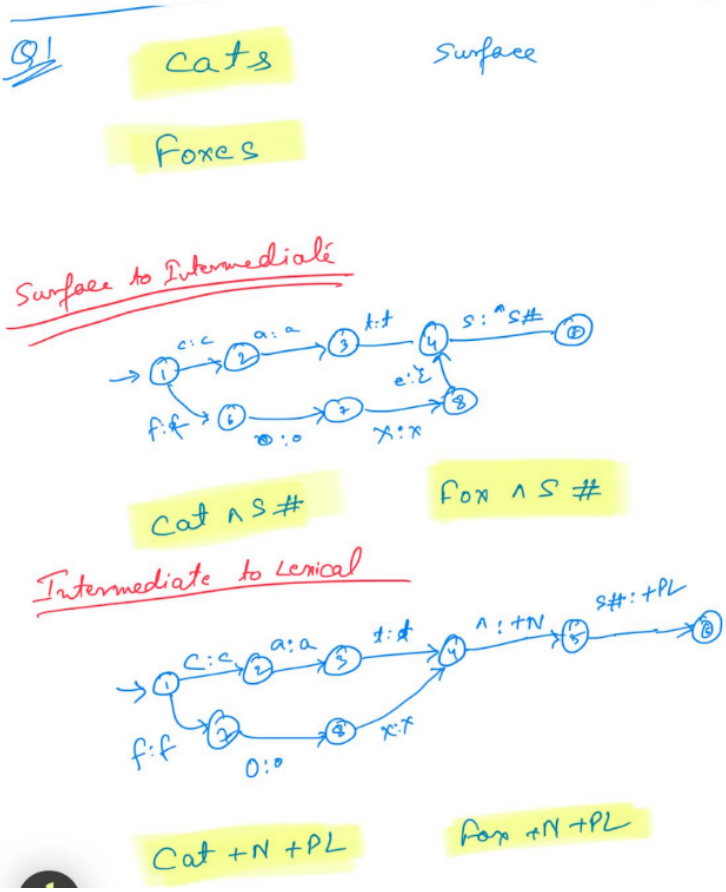
End Sem

8. Construct two-level FSTs (*surface* → *intermediate* → *lexical*) for the morphological analysis of the following words. [20 marks]

   Inputs: cats, foxes

   Outputs: cat + N + PL, fox + N + PL

   [Note: 5 marks each for surface → intermediate for the two words and 5 marks each for intermediate → lexical for the two words]



9. For a trigram language model, find the next probable word (blank) in the sequence considering Laplace smoothing. Show the computation. [4 marks]

   Corpus: [“He is apple of my eye.”, “He hit my eye.”]

   Vocabulary: [He, is, apple, of, my, eye, hit, ate, .]

   Sequence: He ate my _____

   |Vocabulary| = |V| = 9

   For a trigram model, below are the probabilities of the next probable word with Laplace smoothing,

   P(He | ate my) = count(ate my He)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9

P(is | ate my) = count(ate my is)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9
P(apple | ate my) = count(ate my apple)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9
P(of | ate my) = count(ate my of)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9
P(my | ate my) = count(ate my my)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9
P(eye | ate my) = count(ate my eye)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9
P(hit | ate my) = count(ate my hit)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9
P(ate | ate my) = count(ate my ate)+1 / count(ate my)+9 = 0+1 / 0+9 = 1/9
All the probabilities are the same because the word “ate” is not present in the corpus. Hence, the next probable word can be any word present in the vocabulary because the the probability of any word is the same.
One of the heuristics is sorting the vocabulary and then picking the first word from the vocabulary.
Therefore, Next predicted word is “apple”.
Note: The final solution must identify a word with any heuristic.

----------END----------

Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)

CSE556: NLP

Quiz 2