CS 406-01: Unstructured Information Processing

HW 2 (Given Sept. 9, 2019; Due Sept. 16, 2019)

Your answers must be entered in Google Classroom by midnight of the day it is due. If the question required a textual response, you can create a PDF and upload that. The PDF might be generated from MS-WORD, LaTeX, the image of a handwritten response, or using any other mechanism. Numbers in the parentheses indicate points allocated to the question.

---

1. Consider the following toy example (similar to the one from Jurafsky & Martin (2015)). We are given 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>

   Assume that we use a bigram language model based on the above training data. Answer the following questions,

   (a) What is the most probable next word predicted by the model for the following word sequences?

      i. <s>Sam ...
      ii. <s>Sam I do ...
      iii. <s>Sam I am Sam ...
      iv. <s>do I like ...

   (b) Which of the following sentences is more likely, i.e., gets a higher probability with this model?

      i. <s>Sam I do I like </s>
      ii. <s>Sam I am </s>
      iii. <s>I do like Sam I am </s>

   (5 x 7 = 35 points)

2. For this homework, you will do 2 separate/independent exercises. You will download digital versions (just the text portion) of 2 books available from Project Gutenberg and these will serve as the Corpus for each of the 2 independent exercises. The first book is “A Tale of Two Cities” by Charles Dickens and the second book is “War and Peace” by Leo Tolstoy.

   Use each corpus (book) to independently form 1-gram, 2-gram, 3-gram, and 4-gram. Use each of your language model to generate sentences (each sentence is at least 10 words long) beginning with (i) I suppose..., (ii) And Having got...”, (iii) Not two minutes...

   (a) Attach your sentences grouped by the corpus from which they were generated.

   (b) Comment on the quality of the generated sentences with increasing complexity of the language model

   (c) Comment on the quality of the generated sentences from Corpus 1 and from Corpus 2. Explain any difference that you observe

   (200+50+50 points)

- Good hash to choose ?