CS 5/7322
Introduction to Natural Language Processing
Fall 2023

Homework 1

Due date: 10/6 (Fri) 11:59pm

1. (32 points) Consider applying PLSI to the following corpus (each line is a separate document):

   BACCA
   CAABBA
   ACBABAB

   furthermore, assume that there are two topics, and A,B,C are the only types that are available.

   a. Now suppose we initially assigned words to topics as above (black for topic 1, red/underline for topic 2). Calculate the topic-word vectors and document-topic vectors.
   b. Use the vectors generated in part (a) to calculate the topic probability for each word in the corpus
   c. Use the result of (b) to recalculate the topic-word vectors and document-topic vectors.
   d. Calculate whether the vectors in (c) is better for the set of documents.

2. (24 points) Consider the following corpus (each line is a separate sentence):

   ABCCC
   ADBB
   CDADD
   CABB
   DACB

   Suppose we want to build a bigram model based on the corpus above. Assume we have both a begin and end sentence symbol for each sentence.

   Calculate the perplexity of each sentence (separately) for each of the two cases

   a. The base case (no smoothing)
   b. Using Laplace (plus 1) smoothing. Note that the bigram <begin><end> will never occur so it's probability does not need to be smoothed. (Hint: you should start the smoothing process by adding 1 to the frequency of all possible bigrams)

   Also show the probabilities for each bigram (preferably in a 2-d matrix).