600.465 – Natural Language Processing
Assignment 2: Probability and Vector Exercises

February 2016

1. Sample 1:
   log_2probability: -12111.3, word count: 1686, perplexity per word: $2^{12111.3/1686} \approx 145.37$
   Sample 2:
   log_2probability: -7388.84, word count: 978, perplexity per word: $2^{7388.84/978} \approx 188.06$
   Sample 3:
   log_2probability: -7468.29, word count: 985, perplexity per word: $2^{7468.29/985} \approx 191.61$

   When switch to the larger `switchboard` corpus the log_2probabilities go slightly lower while the perplexities go up a lot for they are calculated by taking exponential . This is because typically larger corpus have more words than smaller ones, making the probabilities of words in the sample have lower probabilities to appear.

2. (a) We chose the language ID problem. The lowest error rate we can achieve is 0.933.

   (b) The value of λ we use is 2.7.

   (c) Test result for english:
   ```
   342 looked more like en.1K (92.43%)
   28 looked more like sp.1K (7.57%)
   ```
   Test result for spanish:
   ```
   39 looked more like en.1K (10.57%)
   330 looked more like sp.1K (89.43%)
   ```

   (d)