Homework for *Computational Linguistics* (Week 5)

November 5, 2017

1. Overview
   In this assignment, you will use language modeling to detect what language a document is written in. Specifically, you will write Python3 code to do the following:

   a. build n-gram language models over characters (not words!): read in a text file, collect counts for all character n-grams, estimate probabilities, and write out the models to files.
   b. read in a test document and compute its perplexity according to each of these models, and then decide what language this document is written in.
   c. generate sentences according to the models.

   Use the provided skeleton script called 'ngram.py'. You can use the helper functions if you want. But do not change them. You can write any new functions as you want.

2. Building N-gram Models
   There are two text files under the directory 'data' for training n-gram models: training.de (German) and training.en (English).
   You need to train a bigram model and a trigram model for each language (so you need to train four models in total). Of course, you are very welcomed to try more values for 'n.'

3. Language Detection
   The other text file under 'data', called 'test', is for testing (you need to decide what language this text is written in).
   Use your four language models to calculate the perplexity if the test document. Please see J&M (3rd edition) Chapter 4.4 for the definition of perplexity.

4. Sentence Generation
   Use your English bigram and trigram models to generate five (with each model) sentences (max length = 30).

5. Your observation
   According to the perplexity and sentence generation task, have you observed anything about n-gram language models (e.g., pros/cons of n-gram models, how the choice of 'n' may influence results of both tasks)?

6. Submission
   Things you need to submit for this assignment:

   a. the completed python script 'ngram.py'
   b. four model files
   c. a txt file named 'report.txt', in which there should be: 1, based on each n-gram model, the perplexity of the test text (so four perplexity values in total); 2, based on the two English models, five generated sentences for each (clarify the model under which each sentence is generated; ten sentences in total); 3, your observation.

   Please include all files in a zip file for submission.