LING 570: Hw5
Due date: 11:45pm on Nov 3

All the examples are under **~/dropbox/10-11/570/hw5/examples/.** Also see the slides for Hw5.

Q1 (20 points): write a script, **ngram_count.sh**, that collects unigrams, bigrams, and trigrams.

- The format is: ngram_count.sh training_data ngram_count_file

- The format of the training data: w1 w2 … w_n; that is, one sentence per line (e.g., **examples/training_data_ex**)

- The format of ngram_count_file is: count word1 … word_k (e.g., **examples/ngram_count_ex**). In the file, list unigrams first, bigrams next, and then trigrams. For each n-gram "chunk", sort the lines by frequency.

- You need to "add" BOS and EOS to the input sentence yourself. The BOS string is "<s>" and the EOS string is "</s>". For instance, if the input sentence is
  John call Mary

  After "adding" the markup, the sentence will become
  <s> John call Mary </s>

Q2 (20 points): write a script, **build_lm.sh**, that builds an LM using ngram counts:

- The format is: build_lm.sh ngram_count_file lm_file
- ngram_count_file is the file produced by Q1.
- lm_file follows the modified ARPA format, as discussed in class (e.g., **examples/lm_ex**)
- No smoothing for the probability distributions.

Q3 (30 points): Write a script, **ppl.sh**, that calculates the perplexity of a test data given an LM. For smoothing, use interpolation.

- The format is: ppl.sh lm_file l1 l2 l3 test_data output_file
- lm_file is the file created in Q2.
- Use interpolation to calculate probability: l1, l2, and l3 are lambda_1, lambda_2, and lambda_3 in the interpolation formula, respectively.
- test_data has the same format as the training data (e.g., **examples/test_data_ex**)
- The format of output_file has been discussed in class (e.g., **examples/ppl_ex**)

Q4 (30 points) Use examples/wsj_sec0_19.word as training data, and calculate the perplexity of examples/wsj_sec22.word by running the following commands and fill out the table:

ngram_count.sh examples/wsj_sec0_19.word wsj_sec0_19.ngram_count

build_lm.sh wsj_sec0_19.ngram_count wsj_sec0_19.lm

ppl.sh wsj_sec0_19.lm 0.05 0.15 0.8 examples/wsj_sec22.word ppl_0.05_0.15_0.8

ppl.sh wsj_sec0_19.lm 0.1 0.1 0.8 examples/wsj_sec22.word ppl_0.1_0.1_0.8

…

ppl.sh wsj_sec0_19.lm 1.0 0 0 examples/wsj_sec22.word ppl_1.0_0_0

lambda_1	lambda_2	lambda_3	perplexity
0.05	0.15	0.8	
0.1	0.1	0.8	
0.2	0.3	0.5	
0.2	0.5	0.3	
0.2	0.7	0.1	
0.2	0.8	0	
1.0	0	0	

The submission should include:

- The hw5 note file that includes the table in Q4.
- The source and shell scripts in Q1, Q2, Q3: **ngram_count.sh, build_lm.sh,** and **ppl.sh**, and any scripts called by them.
- The files created in Q4: **wsj_sec0_19.ngram_count, wsj_sec0_19.lm,** and **ppl_*.**