CSCI 8450: Chapter 5b: Assignment 6

Mohan Sai Ambati
Collaborated with Sai Tarun Battula

1.
   (a) The five most frequent nouns in Brown corpus that are more common in their plural form than singular form:
   [('years', 943), ('people', 811), ('men', 736), ('eyes', 391), ('things', 361)]

   (b) Five most frequent tags in decreasing order of their frequency are:
   [('NN', 152470), ('IN', 120557), ('AT', 97959), ('JJ', 64028), ('.', 60638)]
   'NN': noun, common, singular or mass
   'IN': preposition or conjunction, subordinating
   'AT': Articles
   'JJ': adjective or numerical, ordinal
   '.': sentence termination

   (c) I used 4-gram model to find the tags. Three tags that most commonly precedes 'NN' are in all the genres humor, romance and government are ('NN', 'IN', 'AT')
   'NN': noun, common, singular or mass
   'IN': preposition or conjunction, subordinating
   'AT': Articles
   Most common 3 tags that precedes 'NN' is:
   Humor: [(('NN', 'IN', 'AT', 'NN'), 113)]
   Romance: [(('NN', 'IN', 'AT', 'NN'), 284)]
   Government: [(('NN', 'IN', 'AT', 'NN'), 421)]

2.
   (a) Since the model is trained with category news it is obvious that news have more probability of getting correct prediction. Evaluate(category news) is greater than Evaluate(category lore).
   Evaluate on all of the sentences from the Brown corpus with the category lore:
   0.8427274952628764
   Evaluate on all of the sentences from the Brown corpus with the category news:
   0.9826759750979573

   (b) Output of tagger on the 200th sentence of the lore category of the Brown Corpus : [('I', 'PPSS'), ("can't", 'MD*'), ('tell', 'VB'), ('when', 'WRB'), (',', ','), ('but', 'CC'), ("I'm", 'PPSS+BEM'), ('positive', 'JJ'), ('I', 'PPSS'), ('witnessed', 'VBN'), ('this', 'DT'), ('same', 'AP'), ('scene', 'NN'), ('of', 'IN'), ('this', 'DT'), ('particular', 'JJ'), ('gathering', 'NN'), ('at', 'IN'), ('some', 'DTI'), ('time', 'NN'), ('in', 'IN'), ('the', 'AT'), ('past', 'NN'), ("''", "''"), ('!', '.'), ('!', '.')]
   I would tag the sentence in the same manner. I think this tagging is more accurate.

3. Evaluate on all of the sentences from the Brown corpus with the category lore (without back off) : 0.06240310428925013
   UnigramTagger has highest performance ~ 0.79, but by combining all three taggers the tagging process is even more accurate.
   Tagger with back off performs better, because it combines the results of Trigram, Bigram, Unigram and default taggers which performs with more accuracy because it will have more coverage.

4. To solve this, I have taken lemma names for each word class ad formed a dictionary where each word is a key and list of word classes that it occurred is the value. I then counted the total number of words that have more than one word-class.
   More than one tagged words: 7399
   Total words: 147306
   Percentage is: 5.022877547418299