William Vagharfard SID: 860792866

CS173 Class Activity – Apr 15, 2019

Parts of Speech

Source: https://www.nltk.org/

1. Download the Brown corpus using NTLK and answer some basic questions about it.

   Write the absolute path of the Brown corpus files:

   /Users/William/nltk_data/corpora/brown

   How many files are in the Brown corpus:

   There are 500 files in the Brown corpus

   How many sentences are in the Brown corpus:

   There are 57,340 sentences in the Brown corpus

   How are the files packed (i.e., structured):

   The files are broken up by sentences, each word is tagged by it's POS and punctuation.

   What do the filenames mean (HINT this is documented at NLTK):

   The filenames show the different genres of text that the corpus is taken from.
   ca: news, cb: editorial, cc: reviews, cd: religion, ce: hobbies, cf: lore, cg: belles lettres,
   ch: govt, cj: learned, ck: fiction, cl: mystery, cm: sci-fi, cn: adventure, cp: romance, cr: humor

2. Write scripts in python to do the following (HINT try glob & fileinput packages).

   - loop over all of the Brown corpus files

   - extracts each sentence as two vectors: a vector of word tokens and one of POS tokens they should be of the same size, so check that this is true!

   - collect frequency counts:
     - singletons (1-grams) of each of word and POS tokens

     - successive pairs (bi-grams) of words
     - successive pairs (bi-grams) of POS token

     - pairs of POS-word tokens; that is, for every POS, its word distribution (NOT bigrams)

   - how large is each file (both in terms of number of records and storage size):

3. Store these away (e.g., pickle).

   - what data structure did you choose to use?

     Hash table

   - what would it take to also do tri-grams? ... up to 5-grams?

     computationally up to tri-gram is ok, but after 4 or 5-grams it becomes too computationally heavy to make it worth it.