Lab 2: POS Tagging

These solutions available in an html version[1] or a pdf version[2].

Examining the POS Tagset

- Based on your intution guess the most and least frequent tags in a data.

  In English data, determiners (DT), nouns (NN, NNS, NNP etc.), verbs (VB, VBD, VBP etc.), prepositions (IN) might be more frequent. Tags like UH (interjection), LS (list marker) might be less common.

- What is the difference between the tags DT and PDT?

  DT is the POS tag for determiners and PDT is the tag for pre-determiners. As the name says, pre-determiners occur before determiners. For example, in "both the books", "the" is a determiner and "both" is a pre-determiner.

- Can you distinguish singular and plural nouns using this tagset? If so, how?

  We can distinguish singular and plural nouns in this tagset. NN (noun singular) and NNP (proper noun singular) are the tags for singular nouns and NNS (noun plural) and NNPS (proper noun plural) are the tags for plural nouns.

- How many different tags are available for main verbs and what are they?

  There are six different tags for main verbs. VB (base form), VBD (past tense), VBG (gerund/present participle), VBN (past participle), VBP (sing. present, non-3d), VBZ (3rd person sing. present).

Computing the distribution of tags

For implementation, see the answer code[3].

- Using plot_histogram, plot a histogram of the tag distribution with tags on the x-axis and their counts on the y-axis, ordered by descending frequency.

  ```
  plot_histogram(sorted(tag_dist.items(), key=lambda x:  x[1], reverse=True))
  ```

- How many distinct tags are present in the data ?

  There are 45 distinct tags in this data.

- List the tags in the descreasing order of frequency.

  ```
  sorted(tag_dist.items(), key=lambda x:  x[1], reverse=True)
  ```

- What are the 5 most frequent and least frequent tags in this data. How does this compare with your intuition in the previous section ?

  NN, IN, NNP, DT, NNS labels for common noun, preposition, proper noun, determiner, and plural common noun respectively are the 5 most frequent tags. 5 least frequent tags are SYM (Symbols), UH (interjection), FW (foreign words), LS (list marker) and WP$ (possessive wh-pronoun).

[1] http://homepages.inf.ed.ac.uk/sgwater/teaching/lsa2015/labs/lab2-sol.html

[2] http://homepages.inf.ed.ac.uk/sgwater/teaching/lsa2015/labs/lab2-sol.pdf

[3] lab2-sol.py

- What kind of distribution do you see in the histogram, and how does it compare to the histogram of sentence lengths?

  The distribution of tags is very skewed: a few tags are very frequent, but the frequencies drop off rapidly. Although it's hard to tell from this plot, the tag distribution might be similar to the Zipfian distribution of word frequencies (although we would need to do more careful analysis to determine that). This contrasts with the distribution of sentence lengths, which looks approximately normal.

Computing the conditional distribution of tags for each word

For implementation, see the answer code[3].

- How many entries are there in your big CFD?

  There are 11968 entries.

- On average, how many different tags does each word have?

- If you had a larger tagged corpus, would you expect the amount of tag ambiguity to be greater than, less than, or the same as in the current corpus, and why?

  The answer isn't obvious. In a larger corpus, one will observe greater ambiguity for some words, because they may have low-probability tags that haven't occurred in the small corpus we already looked at (in fact, any word that occurred only once in the small corpus, even if highly ambiguous, would appear to be unambiguous in the small corpus, yet with more occurrences would become ambiguous). On the other hand, we will also see more total words in the large corpus, and all of the new low-frequency words are likely to appear with only a single tag (especially if the word only appears once!). So to really know the answer to this question, we would need to actually compute the tag ambiguity for different sizes of corpus. (You could even try it by looking at smaller subsets of this corpus! If you do, let me know what results you get!)

- What is the number of tags and the most frequent tag for the words `the`?

  `word_tag_dist["the"]` will give the results: `{u'CD': 1, u'DT': 4038, u'JJ': 5, u'NNP': 1}`.

- Which word out of the whole corpus has the greatest number of distinct tags?

  For implementation, see the answer code[3]. The words "set", "back", and "hit" have the greatest number of distinct tags, which is 5.

Unigram Tagger

For implementation, see the answer code[3].

- Why is this called a Unigram tagger? How does it differ from an HMM tagger?

  This is called unigram tagger as it uses only the current word as input. HMM taggers are *bigram* taggers: they use the previous tag as context in deciding the next tag.

- Run this simple tagger and tag the sentences a) "book a flight to London" and b) "I bought a new book". Look at the pos tags. Are there any errors in the pos tags? If so, what could be the reason for them?

  In the first sentence "book" is a verb and in the second sentence "book" is a noun. But the unigram tagger assigned the noun tag in both cases because noun is the most frequent tag for "book", and the unigram tagger treats words in isolation. An HMM tagger probably wouldn't make this mistake because it assigns tags in context. Some students suggested that the fact that "book" occurs at the beginning of the sentence provides some (incorrect) evidence that "book"

is a noun, since nouns often occur at the beginning of sentence. That *might* be true (though, I think that noun *phrases* commonly start a sentence, but usually with determiners, less often bare nouns). But, even if true, consider that "book" is also followed by a determiner. Again, this is a possible occurrence in English (e.g., in a ditransitive construction) but not too common, whereas the probability of a determiner following a verb is likely to be much higher. So overall it seems likely that the HMM would decide "book" is a verb. (You might want to try it out and see! Especially if you are doing Q2 of Going Further for your homework assignment. I have not actually checked the HMM tagging of this sentence myself! However, after writing this answer I realized that the tag transition probabilities in question are actually given in slide 19 of Monday's lecture--for a different corpus but let's assume close enough. Was I right in my intuitions?)