Homework 4: ngram tagger

Read Me First

In this homework you will build a bigram part-of-speech tagger and analyze its performance. Before starting,

- read NLTK chapter 5, especially sections 1, 2, 4 and 5 (the sixth won't be needed)
- become familiar with the GACRC teaching cluster's batch system. It's more polite to train models over there, where jobs are distributed across compute nodes, rather than hogging the single shared cpu in our CSCI-provided shared environment.

Test/Train Split

Consider the `news` category within the Brown corpus. Take 90% of those sentences as training data and reserve the last 10% for evaluation.

Q1 Following the steps in NLTK chapter 5 §5.4, build a bigram tagger with two back-off models. The first should be a default tagger that assigns 'NN' by default. Train on the tagged Brown news sentence, then evaluate on the untagged (i.e. held-out) words. Assess performance by examining some mistagged words. In a well-reasoned analytical paragraph, discuss the behavior of the trained tagger on specific cases, including the word "that." Refer to the Brown corpus manual or `nltk.help.brown_tagset()` to get a sense of what these tags mean.

Impact of Tag Set

Q2a Now train another bigram tagger using a different tag set, CLAWS5. There is a collection of news stories from the British National Corpus specially-prepared for this purpose called `bnc-news-wtp.txt` Access it by passing in its name and specifying the `en-claws` tagset as shown below:

```
root=find('corpora/bnc')  # bnc-news-wtp.txt has been placed in that directory
bncnews = TaggedCorpusReader(root,'bnc-news-wtp.txt',tagset='en-claws')
```

Create a test/train split of exactly the same size as the Brown news category and examine performance. Is performance of the exact same model type on BNC-CLAWS higher or lower? Turn in a paragraph discussing why. It may help to compare tag sets in their entirety — form the set of all tags attested anywhere in each of the respective training sets. Analogous to the documention for Brown corpus tags, there is `nltk.help.claws5_tagset()`.

Q2b Same question as 2a but now pass in `tagset = 'universal'` in order to work with the "simplified" tags. Compare BNC-SIMPLIFIED to BNC-CLAWS at a larger scale by creating a new test/train split from the available 508,609 BNC sentences.

Report performance levels in a table. Offer a reason why performance differs, if it does.

Improve Performance

Q3 Choose one of the methods suggested in chapter 5 for increasing performance. Try changing the amount of information that your system relies upon. Again, report your results in a table and comment on which change(s) led to how much improvement. Explain why your chosen improvement helps in an insightful paragraph that accompanies the table.

Q4 EXTRA CREDIT Locate a tagged corpus in a language other than English and repeat the evaluation a 90/10 split, just as you did above with Brown and BNC.