Assignment #1 – Jameson Watts

1. NLP as classification
    a. Genres
        i. Annotations & Resources: Select all 15 top-level categories from DMOZ and generate a random sample of websites within that category. For each of the sampled websites, download the content from the homepage and remove all html tags and other meta-content (i.e. javascript).
        ii. Instance: An instance consists of the raw POS-tagged text with words separated by whitespace
        iii. Labels: Top-level DMOZ categories (Shopping, News, Reference, Sports, etc.)
        iv. 5 features: unigram frequency – because words from different genres have different dictionaries they generally draw from; bi-gram frequency - same; word diversity – reference topics for instance should have more diverse dictionary of words used than Sports; grammar tendencies (e.g. active voice) – because Sports and news should have more causal agents then some document describing art history; function word frequency – different genres have different norms of inference (e.g. the word "thus" vs. "because").
    b. Terrorists
        i. Annotations & Resources: Online news articles from the Wall Street Journal over the past 10 years, cleaned of advertising content, html tags and other meta-content. Might also want the CIA's list of known terrorist names.
        ii. Instance: An instance consists of the raw POS-tagged text from a single article with words separated by whitespace. Important here is that proper nouns are correctly classified.
        iii. Labels: Yes – one of the words in the document identifies a terrorist; No – no terrorists here
        iv. 5 Features: existence of proper noun – for obvious reasons; word found in CIA list – for obvious reasons; proper noun is causal agent or subject in sentence – generally terrorists are either doing something or having something done to them. Unigram frequencies – terrorist activities probably exist in a fairly small dictionary of terms; unigram frequency of words directly before after a proper noun – because person context such as "suspect [proper noun] committed" can be powerful.
    c. Newline
        i. Annotations & Resources: POS-tagged set of OCR documents
        ii. Instance: one line from the annotated OCR document
        iii. Labels: Yes – line ends a sentence; No – line doesn't end a sentence
        iv. 5 Features: last word is a verb – because English sentences don't normally end in a verb. Last word frequency – because there are presumably words, which end sentences more often. Punctuation opened, but not closed – obvious reasons. Last word is a comma or other non-terminal character. No verb since last sentence ending determination.
2. So I don't have the full statistics since my computer is too slow, so I worked with about 40,000 lines from the brown corpus. But I think there are still some interesting conclusions. First, a lot of the tokens just do not exist in the training set. For example, my file of the POS trigrams from the test set had almost no matches in the training set. This is probably a poor feature to use. Another is just the vast number of tokens with a frequency of 1. Seems like it would be almost impossible for a machine to learn differences if there isn't at least some variance in how a feature is used.