What do we do if there are hyphens, capitols in the middle of a word, misspellings?
Do we want to keep track of words at the beginning of sentences? Our test data need to have enough words that have high frequency. What do we do with words like 12.exe. or http://www.weirdo.com and #9 and 1999?