Assignment #2
Karen Mason

Assignment used the Enthought Canopy distribution of Python verson 2.7

The *Natural Language Processing with Python* "(Bird, et al), commonly referred to as NLTK package was used for processing the text files for stemming, stop words and n-gram calculations. The Snowball stemming method was used because according to the NLTK documentation it is a more robust stemmer. The stop word list in NLTK is 127 words including such words as "a", "the", "no". Because the files being processed are patent descriptions, none of the words in the list were determined to have an impact on the meaning. Therefore none were removed from the list. Also none were determined to be necessary to add to the list.

Punctuation in the text files was not removed. The initial reason for this is punctuation can provide meaning or be a pause in the thought being expressed. For example the colon(:), indicates that there are multiple points listed. Including punctuation provides more of the meaning than excluding it. In hindsight, my opinion is that punctuation should be removed. Ease of processing was another consideration.

Citations:
Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Enthought Canopy (Version 1.3.0) [Software]. (2014). Retrieved from http://www.enthought.com