ICS 661 Quiz 10

Calculate the cosine and Jacard similarity between two vectors for disambiguating the meaning of "virus" in two news articles. The two senses of "virus" on WordNet are virus#1 (virology) and virus#3 (a software program capable of reproducing itself and usually capable of causing great harm to files or other programs on the same computer). Use both the WordNet gloss and sample sentences (just as in the Lesk Algorithm) for each sense of virus to create the reference vectors. Use a stop word list such as http://www.lextek.com/manuals/onix/stopwords1.html to filter out overly common words first. Next find two recent news articles using a search engine such as http://news.google.com/, one for each sense of virus. Build test vectors using the words in the sentence before, the sentence after and the sentence containing the word virus from the news article (after deleting stop words). Try to find a sentence that is not the first or last sentence in a paragraph to form the test vectors. Calculate both cosine and Jacard similarity measures for the two test vectors and the two reference vectors.

Note an actual exam would provide the actual word counts rather than ask you to use a search engine and WordNet to get them.