

CS2402
Spring 2009
Lab 4
Lists
Due Friday, March 6, 2009

Instructions

In Natural Language Understanding, we try to enable computers to “understand” languages such as English and Spanish. It has been found that we can enable computers to classify text according to its topic (for example sports, politics, science, etc.) by simply counting the occurrences of meaningful words in the text to be classified and comparing those frequencies with those of texts of known classes.

In this lab you will write a program to count the occurrences of meaningful words in English text. To decide if a word is meaningful (also known as a content word), we will use a list of words that are known to provide little meaning (also known as stop words) that can be found at www.dcs.gla.ac.uk/edom/ir_resources/linguistic_utils/stop_words. We will assume that any word that is not on that list is a meaningful word.

Your task consists of writing two versions of a program to do the following:

1. Prompt the user for the name of a text file to be analyzed.
2. Read the text file selected by the user and store the content words and count the number of occurrences of each of them.
3. Output a list of the 50 content words that were used with the highest frequency in the text, and the number of times each of them was used.

For version one of your program, use unordered lists to store stop words and content words, for version two use ordered lists for both types of words. In both cases, report the total number of string comparisons that were performed by each program and analyze the relative efficiency of both representations. Perform experiments using several input files and observe the relationship between the topics of these texts and the lists of words found by your programs.