Assignment 1

Assigned: Sept. 14
Due: Sept. 28.

Note: The Matlab code for all three problems is *quite* short.

Problem 1: Document Vectors

Write a MATLAB function `DocSimilarity(D,E)` which computes the "similarity" of text documents D and E using the vector model of documents. Specifically, the arguments D and E are each cell arrays of strings, each string being a word of the document, normalized to lower case. The function returns a number between 0 and 1, 0 meaning that the two documents have no two significant words in common, 1 meaning that they have the identical significant words with the same frequency.

A word is considered "significant" if it has at least three letters and is not in the list of stop words provided at SampleCode/GetStopwords.m on the course web site.

A stop word is a very common word that should be ignored.

Your function should execute the following steps.

- Let LargeOdd be any reasonably large odd number that is not very close to a power of 256. 10,000,001 will do fine.
- Load in the cell array of stop words from GetStopwords.m
- Create three sparse vectors $\vec{S}, \vec{D}, \vec{E}$ of size LargeOdd, as follows: For every word W, let i=hash(W,LargeOdd). You can find a hash function at SampleCode/hash.m. Then
  - $\vec{S}[i] = 1$ if W is on the list of stop words.
  - $\vec{D}[i]$ = the number of occurences of W in D, if W is significant.
  - $\vec{E}[i]$ = the number of occurences of W in E, if W is significant.

  (Create $\vec{S}$ first, then use it for a quick test for whether words in the documents are significant.) $\vec{D}$ and $\vec{E}$ are the document vectors (we omit the inverse document frequency).
- Return the quantity $\vec{D} \cdot \vec{E}/|\vec{D}||\vec{E}|$

For instance,

```
>> D = { 'how', 'much', 'wood', 'could', 'a', 'woodchuck', 'chuck',  ...
'if', 'a', 'woodchuck', 'could', 'chuck', 'wood' };
>> E = { 'all', 'the', 'wood', 'that', 'a', 'woodchuck', 'could',  ...
'if', 'a', 'woodchuck', 'could', 'chuck', 'wood' };
>> DocSimilarity(D,E)
ans =
  0.9245
```

Note that the only significant words in these two texts are "chuck", "much", "wood", and "woodchuck".

You don't have to worry about hash collisions, because they are very infrequent, and the technique is completely imprecise in any case.