Automatic Document Comparison

- Problem: Given documents A and B, determine how similar they are (say, on the basis of word usage).
- Applications:
  - Grouping documents in information retrieval systems (Google, Lycos, etc)
  - Automatic essay grading
  - Literary style analysis by computer

Simple Method

- Compute:
  - A – B = set of words in A but not in B.
  - B – A = set of words in B but not in A.
  - $A \cap B$ = set of words in both A and B.
  - $A \cup B$ = set of words in either A or B.
  - Define card(S) = number of elements in S.
- Compute:

  $(card(A-B) + card(B - A)) / (1 + card(A \cap B))$

Cosine Comparison

- Let $A \cup B$ be represented by
  $[w_1, w_2, \ldots, w_n]$
- Represent A by the vector
  $V_A = [a_1, a_2, \ldots, a_n]$
  Where a_i = number of occurrences of w_i in A.
- Let V'_A = normalized version of V_A
  $V'_A = [a'_1, a'_2, \ldots a'_n]$
  Where $a'_i = a_i / || V_A ||$
- Dot product: $V'_A \bullet V'_B = \Sigma a'_i b'_i$
- Cosine "distance": $d_c(A,B) = V'_A \bullet V'_B$