# Document Classification with the Cosine Similarity

Consider the sentences A, B, and C below. First, stem any words having morphological inflections. Next eliminate stopwords in each. Finally determine the minimal reference vocabulary, sort it alphabetically, and encode each sentence as a vector of occurrence counts according to the reference vocabulary.

1. Musk says reserve your 30M ticket to Mars with a 1M deposit.
2. Muir loved the open space of the Mohave.
3. Mind, said Minsky, is a society.

Treat as stopwords: all prepositions, pronouns, articles, numbers or tokens beginning with numerals, and the verb “be”.

1. [1,0,1,0,0,0,0,1,0,1,1,0,0,1]
2. [0,1,0,0,0,1,1,0,1,0,0,0,1,0]
3. [0,0,0,1,1,0,0,0,0,0,1,1,0,0]

Suppose we are trying to decide whether A is more similar to B than to C. Compute the cosine similarity for A-B and for A-C. Which classification is implied by the results?

A o B: 0

A o C: 1

|| A || = sqrt(6)

|| B || = sqrt(5)

|| C || = 2

cos(A, B): 90 degrees

cos(A, C): 78.2 degrees

A is more similar to C.

S. Tanimoto, May, 2018.