Data mining: homework 3

Edo liberty

In this assignment we will construct an algorithm for fining similar text documents in a large text corpus based on trigrams. A trigram is a sequence of three consecutive words. For example the previous sentence contains the trigram "construct an algorithm". We represent a document as a set of trigrams. The distance between two documents A and B is defined to be $d(A,B) = 1 - \frac{A \cap B}{A \cup B}$. Two document are considered duplicates if $d(A,B) \leq 0.1$. Assume no document is over 10,000 words long.

1. Suppose you wanted to represent these documents as vectors on the hamming cube $\{0,1\}^d$ containing 1 for each coordinate for which the corresponding trigram exists and zero else. Also suppose the are 10^6 words in english out of which trigrams can be constructed. What is the dimension d of this space?

2. Given the vector representation of these document $\vec{A}$ and $\vec{B}$ show that

   (a) $d(A,B) \leq ||\vec{A} - \vec{B}||_1 / \max(|A|,|B|)$.

   (b) $d(A,B) \geq ||\vec{A} - \vec{B}||_1 / 2\max(|A|,|B|)$.

3. How many trigrams does a document with 1002 words contain?

4. Considering the similarity between the distance functions and the fact that the Hamming distance in this case is identical to the ℓ_1 norm, you might consider using the algorithm taught in class. Would you recommend that? why? (hint: think about the computation of $g(x)$)

5. Consider a hash function on trigrams $f : t \to [1, \ldots, 10^{20}]$ uniformly over the choice of f. Consider the map $h(A) = \min_{t \in A} f(t)$. Compute the probability that $h(A) = h(B)$ for two documents A and B. (you can neglect the effects of hash collisions)

6. Compute sensitivity coefficients of this function for $r_1 = 0.1$ an $r_2 = 0.2$. Does it suffer from the same problem as the function in question 4?