Programming and Classification

List 3

Marek Klonowski

VII. Edit distance

1. Show that the edit distance $L(w, v)$ is a metric function.
2. Let $\mathcal{W}$ be the set of all words (finite strings) over the English alphabet. Let us consider it with the edit distance as a metric space $(\mathcal{W}, L)$.
   - Descibe the set (a ball) $B('zupa', 1)$.
   - What is the size of $B('zupa', 1)$?
   - Estimate the size of $B('zupa', 3)$.
   - Find $B('zupa', 1) \cap B('kupa', 1)$.
3. Let $H(v, w)$ be the Hamming distance between v and w.
   - Show that $L(v, w) \leq H(v, w)$.
   - Show a pair of words w, v such that $L(v, w) = H(v, w)$.
   - Show a pair of words w, v such that $L(v, w) < H(v, w)$.

VIII. Jaccard similarity

1. Compute the Jaccard similarity between sets $\{a, b, c\}$ and $\{a, b\}$.
2. Compute the Jaccard similarity between **bags** $\{a, b, c\}$ and $\{a, b\}$.
3. Roman chooses randomly two distinct elements from a set of $n \geq 2$ items. In the same way Tadeusz chooses randomly and independently two (possibly the same) elements from the same set. Jet J be the Jaccard similarity between the sets chosen by Roman and Tadeusz. Compute precisely the expected value of J. What can we say if they choose some r out on n elements assuming that n is much greater than r?

IX. Shingling

1. Find the set of all $2-$shingles of a string `abracadabra`.
2. What is the maximal size of a set of all k-shingles in a string of the length n over an alphabet with m symbols?

X. Misc

1. When TF.IDF is equal 0? Consider all cases.
2. Find most stupid existing joke and send it to `klonowski@wp.pl` (Can be both in English or in Polish). How to measure how funny is a joke?
3. Estimate the number of cars produced in mankind history.