Programming and Classification:

4. Similarity + Clustering

Marek Klonowski

Suggested deadline: 15.06.2022

You will need NLTK `https://www.nltk.org/`.

37. Construct a hash function from $\{0,1\}^*$ into $\{0,1,\ldots,m-1\}$.

38. Generate a set S of n random bitstrings of length 100. Find $\min_{x,y\in S}$ sha-1$(x\|y)$, where $x\|y$ denotes concatenation of bitstrings x and y. Estimate, what is the maximal n for this task that can be handled by your computer?

39. (use NLTK). Let S_1, S_2, S_3 be the sets of all words shorter than 8 letters from `text1, text2, text3`, respectively. Compute signatures for S_1, S_2, S_3 represented by 100 minhashes and then estimate Jaccard similarity between each pair of S_1, S_2, S_3.

40. Compare the results from the previous exercise with the exact Jaccard similarity of sets S_1, S_2, S_3. What if random permutation of the characteristic matrix rows were replaced with a random mapping?

41. Using previously defined set S_1 construct sets $S_1^1, S_1^2, \ldots, S_1^{99}$ by removing from S_1 at ranom $1\%, 2\%, \ldots 99\%$ of elements. Then, using the banding technique, try to find similar sets in the set of sets $S_1, S_1^1, S_1^2, \ldots, S_1^{99}$. Try to find reasonable parameters like the number of minhashes, b and r.

42. Banding technique: construct a program that for a given number of minashes n and similarity parameter s suggest parameters b and r such that signatures of two sets are consider "potentialy similar"iff their Jaccard smiliarity is around s.

43. Let $S = \{(1,1),(1,4),(1.1,2),(1,1.1),(1.2,2.2),(5,1)(-1,-1),$
$(-1,-4),(-5.1,-2.1),(-5.2,-0.9),(-5.1,-1.1),(-5,-1)(-6,2),$
$(-6.2,2.1),(-5,3),(-6,3.1),(-6.2,3.2),(-5.5,3.3)\}$
Use *k- menas* algorithm for clustering S. Return pictures of clusters for $k = 2,3,4,5$.

44. (use NLTK). Let S_1, S_2, S_3 be the sets of all words with at most 7 letters from `text1, text2, text3`, respectively. Let $S = S_1 \cup S_2 \cup S_3$. Use *any reasonable* algorithm for clustering S (with edit distance). Return sizes of clusters for $k = 2,3,10$ clusters.