Programming and Classification:

4. Similarity

Marek Klonowski

Suggested deadline: 08.06.2019

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

31. 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?

32. (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.

33. 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?