Programming and Classification:

3. Simple similarity of texts

Marek Klonowski

Suggested deadline: 23.05.2022

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

26. $\star$ For a given bitstring **b** list all bitstrings **b′**, such that the Hamming distance between **b** and **b′** is equal 1.
27. $\star$ Construct a function that returns a Jaccard similarity for two sets. Beware that this function needs to check if at least one of the sets is nonempty.
28. $\star$ Construct a function that computes Jaccard similarity for two strings treated as bags of words.
29. $\star\star$ (use NLTK) List all words in `text1` with edit distance from the word `dog` smaller than 4. Hint: you can safely reject all long words without computations (why?).
30. $\star\star$ (use NLTK) Let `text1` - `text9` be bags of words. Compute similarity between all pairs of texts.
31. $\star\star$ (use NLTK) Let us consider a metric space (S, d), where S is the set of words from `text1` and d is the Hamming distance. Find diameter of (S, d).
32. $\star\star\star$ (use NLTK) Construct a dictionary that assigns each pair of consecutive words in `text1` the Jaccard similarity between them.
33. $\star\star\star$ (use NLTK) Draw a graph with nodes labeled by words in `text2` that appear at least l times. Add edges conecting pairs of words with edit distance smaller than s. Try to minimize l, maximize s and keep the quality of your visualization (`networkx` may be insufficient).
34. $\star\star\star$ (use NLTK). For two words v and w, let *relative edit distance* be the Levensthein distance between v and w divided by the sum of lengths v and w. Find two **different** words in `text2` with minimal relative edit distance.
35. $\star\star\star\star$ For a given bitstring **b** and a natural number n list all bitstrings **b′**, such that the Hamming distance between **b** and **b′** is equal n.
36. $\star\star\star$ Construct a function that for a given string and a natural number k returns a **set** of all its k-shingles.