Programming and Classification Summer 2019

Lecture NUMBER — 20.03, 2019

Dr hab. inż. Marek Klonowski *Scribe: Krzysztof Agieńczuk*

1 Overview

In the last lecture we talked about NLP elements and metrics.

In this lecture we covered:

2 Reminder

Definition 2.1. *Metric Let X be a set. Function $d : X \times X \to [0, \infty)$ is a metric if following conditions are satisfied for and $x, y, z \in X$:*

- $d(x, y) = 0 \iff x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$

Most known metric is euclidean distance, however it can be generalised to Minikowski distance. Another example can be "Jungle river metric".

We will talk a lot about graphs. Graph $\mathcal{G} = (V, E)$ where V is set of vertices and E are edges, being a family of two element subset of V.
Path in a graph is a sequence of vertices such that $\{v_i, v_{i+1}\} \in E$ Some important terms about graphs:

- cycle - path of edges and vertices wherein a vertex is reachable from itself
- connected component - part of graph that is connected
- clique - a fully connected graph
- tree - graph that is connected and doesn't have cycles
- isolated vertex - vertex without any edges
- vertex degree - amount of edges going out of a vertex
- neighbourhood - set of vertices directly connected to a vertex

Sample metric for graphs is the shortest length between two vertices.

Definition 2.2. *Ball* $B(x, r) = \{x' \in X : d(x, x') \leq r\}$

3 Distance between words

3.1 Hamming distance

One of the metrics to determine closeness of two words is Hamming distance. It is defined as number of positions at which the corresponding symbols are different for strings of equal length. Denoted as $d(x, y)$, sometimes $H(x, y)$. This metric has some problem. Eg. strings ”1010101” and ”010101” have a maximal possible distance (H=6) although they ”look” similar.

3.2 Edit distance

Definition 3.1. *Edit distance between two strings is a minimum-weight series of edit operations that transform one into another.*

There are a few edit distances: Levenshtein, LCS, Damerau-Levensthein Levensthein allowes operations:

- deletion
- insertion
- substitution

LCS is only deletion and insertion and Damerau-Levenshtein allows only for substitution

3.3 Wagner-Fisher algorithm

Making computer count edit distance is not very easy. Complexity quickly can go thorough the roof. One of the algorithms is Wagner-Fisher algorithm, basing on dynamic programming.

4 Spell correction

The simplest idea for spell correction is to have a dictionary of words, check if the written word is in this dictionary and if not, suggest some words based on edit distance to the user. This edit distance can be used for example to create graph of words to determine closeness.