Programming and Classification Summer 2019

NLP Elements — 19.03, 2019

Prof. dr hab. Marek Klonowski *Scribe: Krzysztof Agieńczuk*

1 Overview

In the last lecture we continued introduction to python and mentioned basic methods of text analysis.

In this lecture we talked about natural language processing.

2 Quick Reminder

Dictionary is a data structure (in other languages called a map), which contains pairs of value and key. In python they are declared `dict = {}` or by using function `dict()`. We can list elements of the dictionary, but normal `list(dict)` will return a list of keys.

3 Elements of NLP

We come back to categorising documents and once again encounter the problem with using characteristic words for document classification. Better idea for it would be previously mentioned TF.IDF method which stands for Term Frequency times Inverse Document Frequency. However there's no proof confirming it is best method or that it actually is correct. "Best we have is our intuition"[1]

3.1 Using TF.IDF in python

Firstly - we need data. Best way for it is probably NLTK (natural Language Toolkit), which has to be externally installed (by using pip for example).

Secondly, construct a dictionary with frequencies of each word in a document. Probably first preprocess words to make them all lowercase for example. Then we can sort our set. This set contains however a lot of stopwords ("a", "an", "the", etc.). There are prepared lists of stopwords (also in NLTK) used for filtering them out. Better lists give better results (the one in NLTK is pretty bad). Most often it is also needed to remove punctuation marks, as most likely they will be considered as words (depends on your implementation).

[1] prof. Klonowski during the lecture

4 Metrics

Metrics are used to define distance between elements of the set. Eg. Euclidean distance is a metric, but there are also other possibilities.

Definition 1 Let X be a set. Function $d : X \times X \to [0, \infty)$ is a metric if following conditions are satisfied for and $x, y, z \in X$:

1. $d(x, y) = 0 \iff x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Using this definition we can find the distance between eg. two vectors in $\mathbb{R}^4$ or two text bodies (of course, we first have to define function to do it, but this definition describes the function we have to define).

Definition 2 A pair (X, f) where X is a set and f is a function of distance is called a metric space.

Example 1 $x, y \in \mathbb{R}$ $d(x, y) = |x - y|$ is a metric in $\mathbb{R}$.

Example 2 Previously mentioned Euclidean distance is a metric[2].

Example 3 Minikowski distance is used to calculate the distance between 2 vectors $\vec{x}$ and $\vec{y}$ using function

$$d_p(\vec{x}, \vec{y}) = ||x - y||_p \tag{1}$$

for $p \geq 1$

Example 4 Maximum distance is a metric for two vectors $\vec{x}$ and $\vec{y}$ in $\mathbb{R}^n$ and is defined as

$$||\vec{x} - \vec{y}|| = max_{i=1}^{n}\{|x_i - y_i|\} \tag{2}$$

[2] I believe you know the formulas for it. If not - bro, do you even?