Programming and Classification Summer 2019

Lecture Intruduction — 27.02, 2019

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

1 Overview

In this lecture we discussed basics of python programming language and started the topic of Natural Language Processing.

2 Contents

Topics for the lecture include:

- Introduction to Python
- Basics of text analysis
- Finding similar items
- Jaccard simulation
- Locally-sensitive hashing for documents classification
- Random Hyperplanes
- Curse of dimensionality
- Clustering
- K-mean algorithm

Suggested reading for this lecture are "Python for Data Analysis" by Wes McKinney[1] and "Mining massive Datasets" by Jure Leskovec et al[2].

3 Python

The lecture mostly consisted of brief introduction to Python. Some basic facts for you to remember: python is interpreted language (not compiled), it is slower than C++ or Java, but has many libraries for data analysis. Essential include NumPy, pandas, matplotlib, SciPy, sk-learn and others.

Basic rules to remember: spacing is important, everything is an object, there are 5 types of variables. We also talked about strings, loops, conditional expressions, templates, functions, lists and tuples. These notes are not tutorial of python, so for better knowledge I send you to documentation.

4 NLP

We started discussing NLP problems, exactly speaking problem of categorising documents. The question was how do we decide the topic of an article, given set of important words, knowing that they may appear in different contexts or not appear in correct one (ex. word team can appear in article not connected to football, and words yellow card may not appear in one about the sport). The question was up for discussion.

References

[1] McKinney Wes, Python for Data Analysis *O'Reilly*, 2017.

[2] Leskovec Jure, Rajaraman Anand, Ullman Jeffrey D. Mining of Massive Datasets *Stanford Uni*, 2010-2014.