NLP: Final Project: Software License Classificator

Due on 13^{st} of April

Teacher: Leon Derczynski

Author: Evgeny Gryaznov

Project Description

Today there exist three primary categories of software licenses:

I. *Open Source.* Those are the licenses which impose either no or little restrictions on a licensee. Examples: MIT, BSD, Mozilla Public License.

II. *Copyleft-like.* Licenses of this type forbid the licensee from rights escalation. For instance, you cannot make a project closed-source if it's under such license. Typical example: GNU GPL.

III. *Proprietary.* A license for which the software's publisher or another person retains intellectual property rights usually copyright of the source code, but sometimes patent rights.

In every category, there are hundreds of licenses, each represented as a text file. It would be great to have some way of automatically determining the type and the name of a given license.
The goal of the project is to develop a program that can identify the category and the name of a license given its text representation. Technical details are the following:

- Programming Language: Java 8 or Python 3.
- Interface: Command-line.
- Input Format: Text file with a license description.
- Output Format: Two strings – license's type and name.

The cool thing about this project is not only that it perfectly fits into course's syllabus, but also has the potential to be implemented via different methods and skills which we learned in this course.
For example, since the described task is an instance of well-known *classification problem*, we can apply different ML techniques to solve it. Finally, the structure of a license file is typically well-formed, thus a number of rule-based methods can be used.