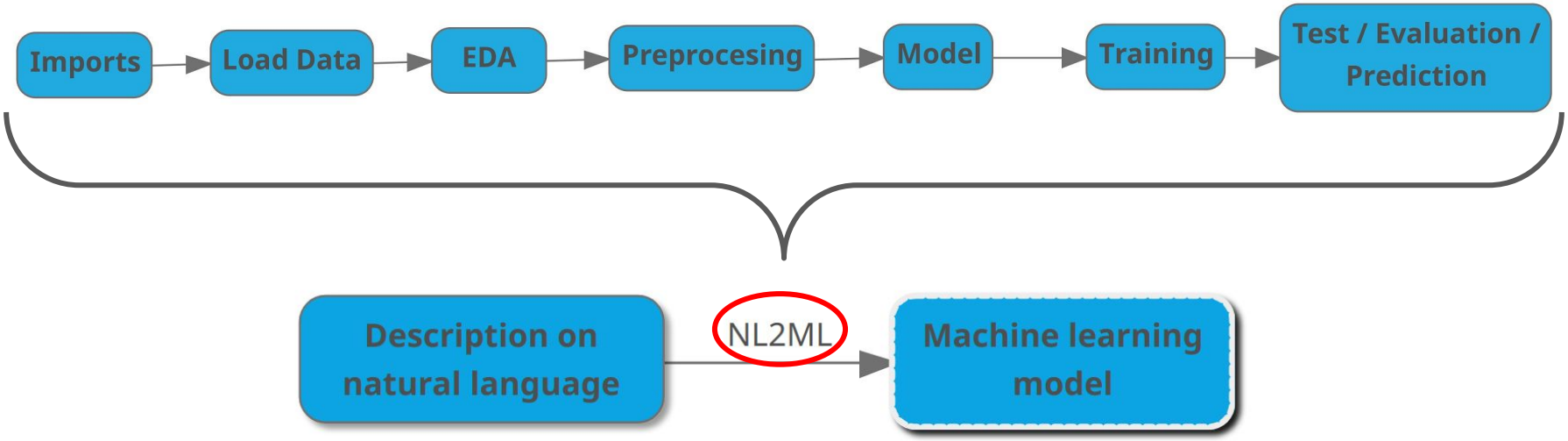
Курсовая работа

Корпус Natural Language to Machine Learning

Гельван Кирилл Павлович, гр. 171
Научный руководитель: Устюжанин Андрей Евгеньевич

Natural Language to Machine Learning



Корпус NL2ML

ML Python code	Description
code_block_1	description_1
...	...
code_block_n	description_n

Цель работы: составить корпус NL2ML

Задачи:

- Построение черновика графа знаний пайплайна машинного обучения
- Сбор первоначальных данных из открытых источников
- Разметка собранных фрагментов
- Разработка модели разбиения фрагментов кода по выбранным классам

Необходимость в большом корпусе данных

Код	Описание
full['Age'] = full.Age.fillna(full.Age.mean()) full['Fare'] = full.Fare.fillna(full.Fare.mean())	Fill missing values of features "Fare" and "Age" with mean value

kaggle GitHub stackoverflow

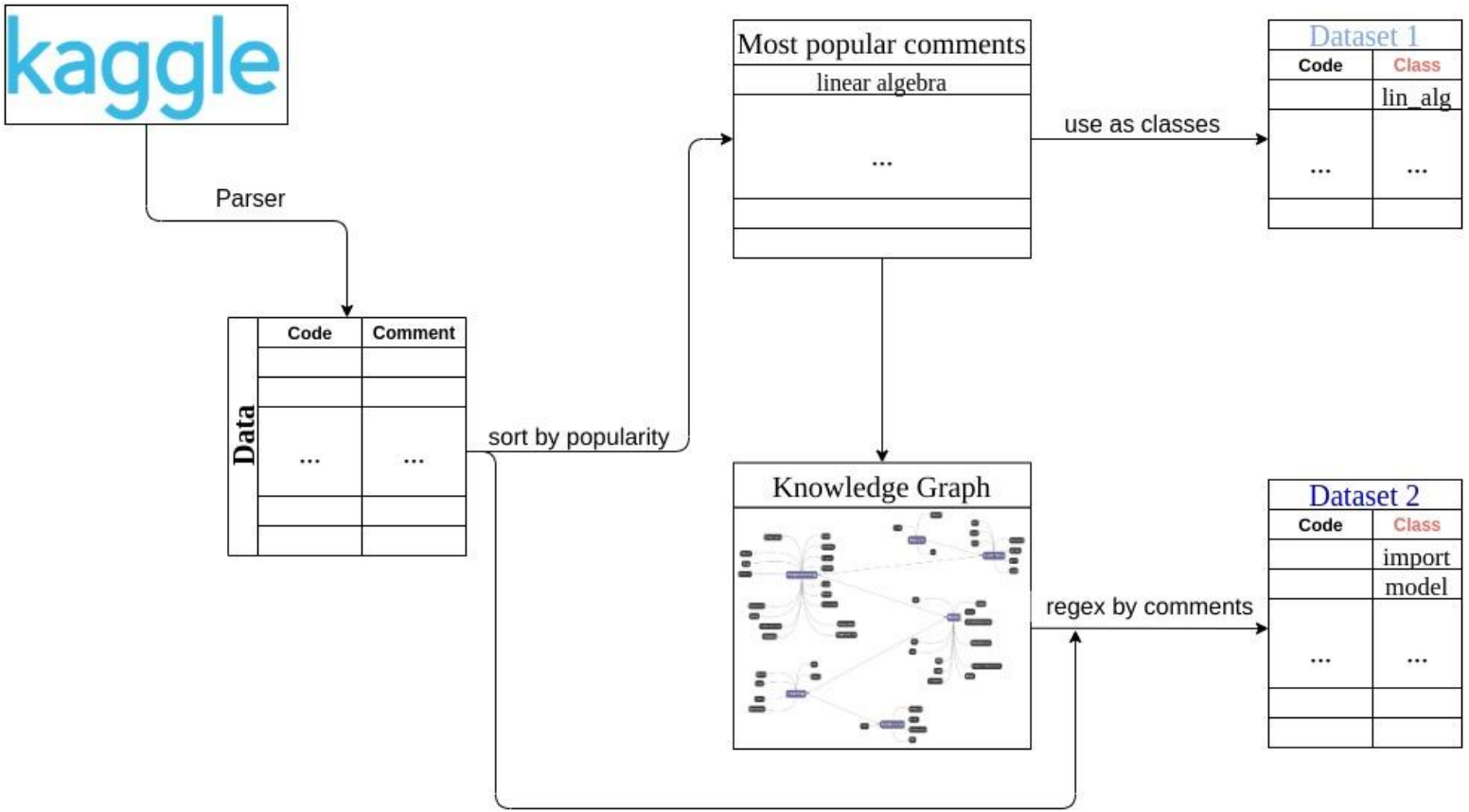
Обзор существующих решений

Датасеты	Подходы к разметке
1. CoNaLa - Python, 600 000 / 2500 2. CodeSearchNet - 300 000 / 750 3. 150k Python Dataset stackoverflow	1. Transformer-based architectures[1)] 2. AST-based models

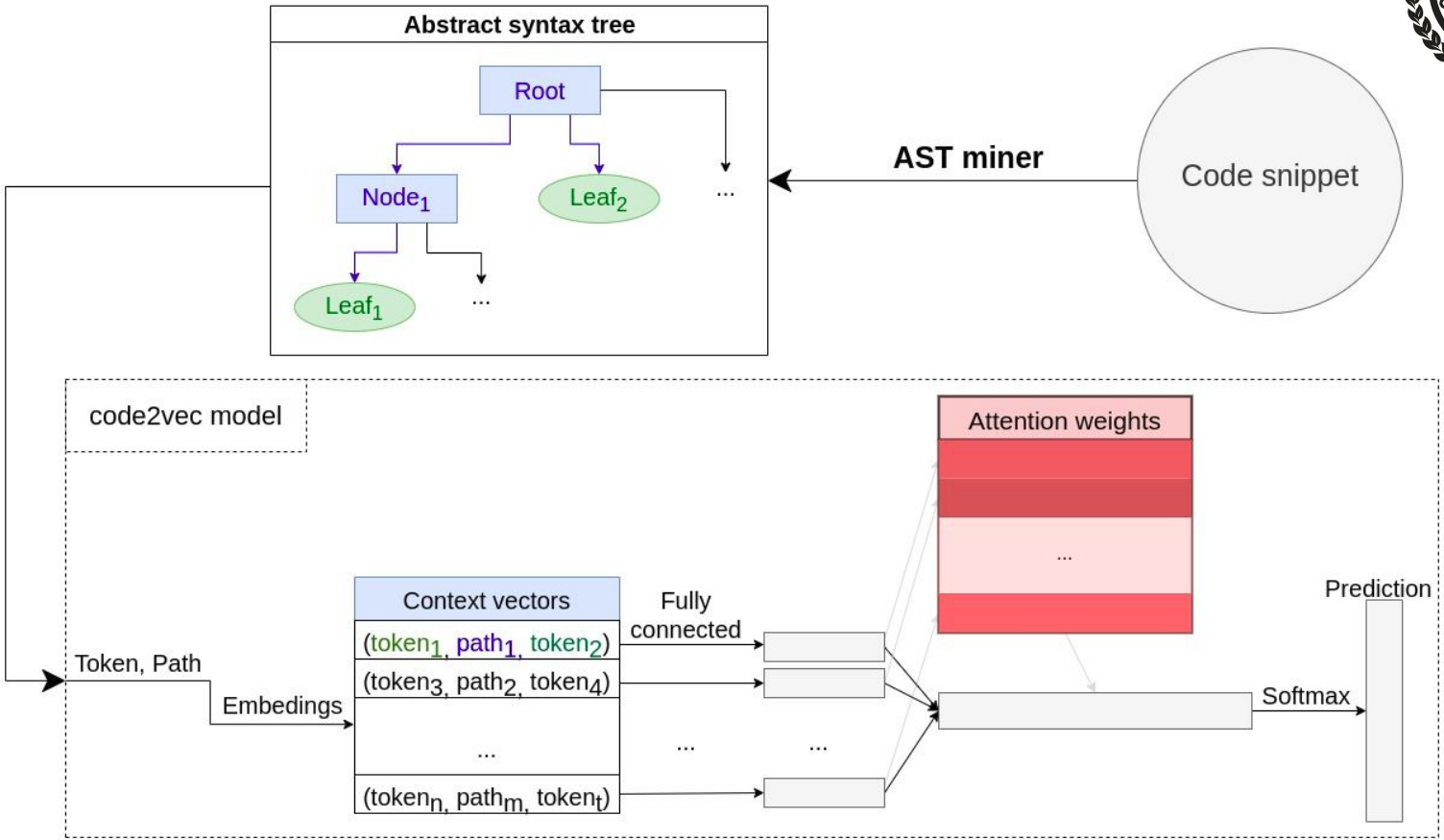
***Корпус NL2ML** - 100 000 / 400*

[1)] Ashish Vaswani et al. Attention Is All You Need. (2017)

Сбор и первоначальная разметка данных



code2vec


Abstract syntax tree
Root
Node1
Leaf2
Leaf1
AST miner
Code snippet
code2vec model
Attention weights
Context vectors
Fully connected
Prediction
Token, Path
Embedings
(token1, path1, token2)
(token3, path2, token4)
(tokenn, pathm, tokent)
Softmax

Сравнение моделей*

	logistic regression	BERT	code2vec
f1 score	**0.7474**	0.5543	0.5863

* Обучение на втором датасете, оценка качества - на экспертно-собранном

Результаты работы

- Построен черновик графа знаний пайплайна машинного обучения
- Собраны данные из открытых источников
- Построены и выявлены лучшие модели для классификации кода

Собран датасет из более чем 100 000 блоков кода с разделением на шесть выверенных классов

GitHub

https://github.com/Kirili4ik/NL2ML