Analiza projektów na platformie GitHub

...

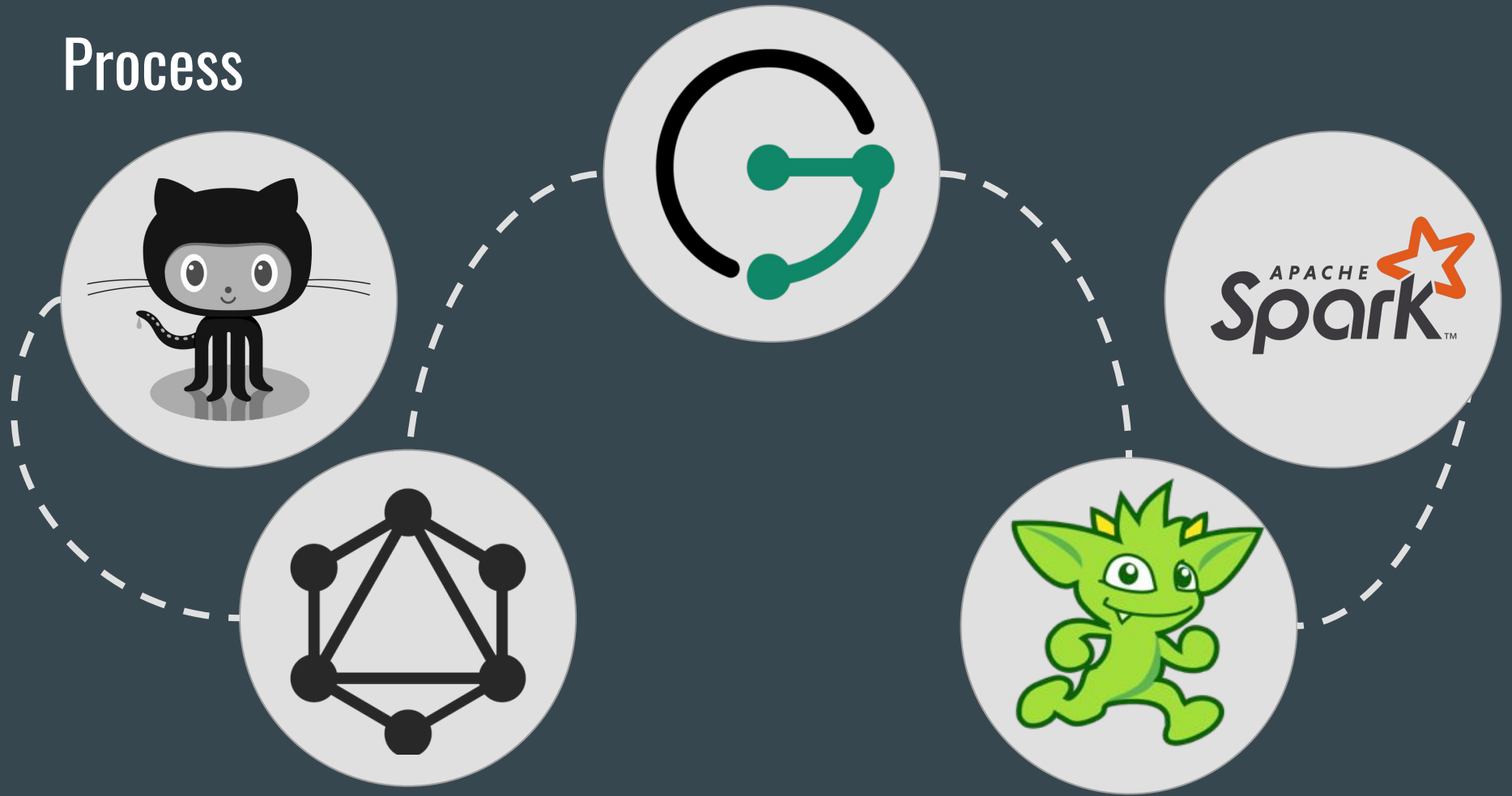
Warszawa 2019

Michał Kośmider, Mateusz Dorobek, Artur Gajowniczek, Stanisław Pawlak

Cel Projektu

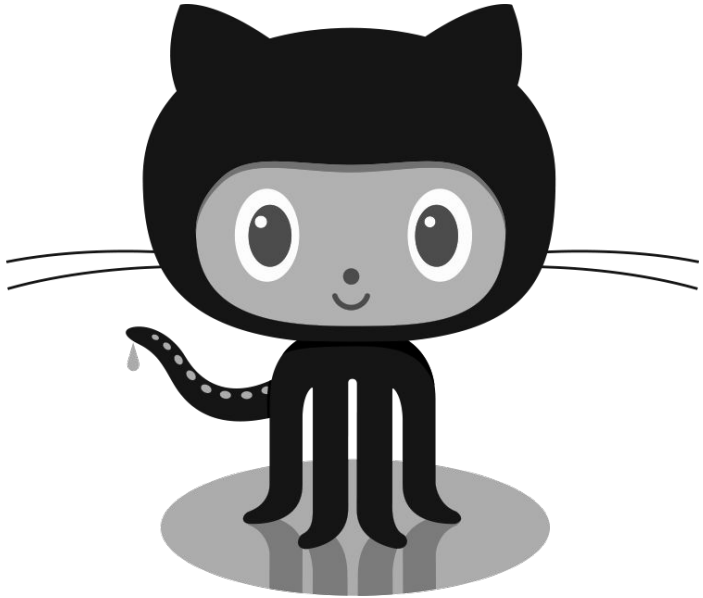
Głównym celem projektu jest wykorzystanie danych udostępnianych przez platformę GitHub do predykcji sukcesu projektów informatycznych na podstawie informacji zawartych w historii repozytorium ich kodu. Wolumen danych jest rzędu setek Megabajtów.

Process



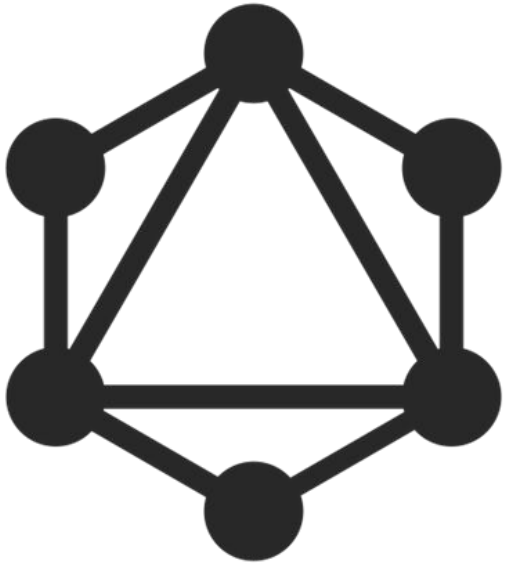
GitHub

Code Repository



GraphQL API v4

Official GitHub API



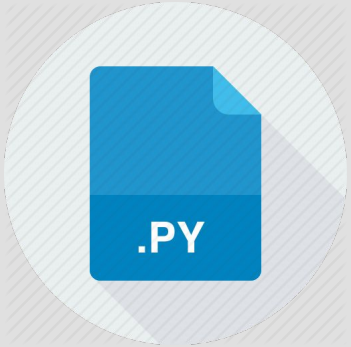
Pobierane dane

Repositories	Forks, Disk Usage, Assigns, Stargazers, Issues, Milestones, Releases
Users	Comments, Followers, Issues, Pull Requests, Repositories, Starred Repositories.

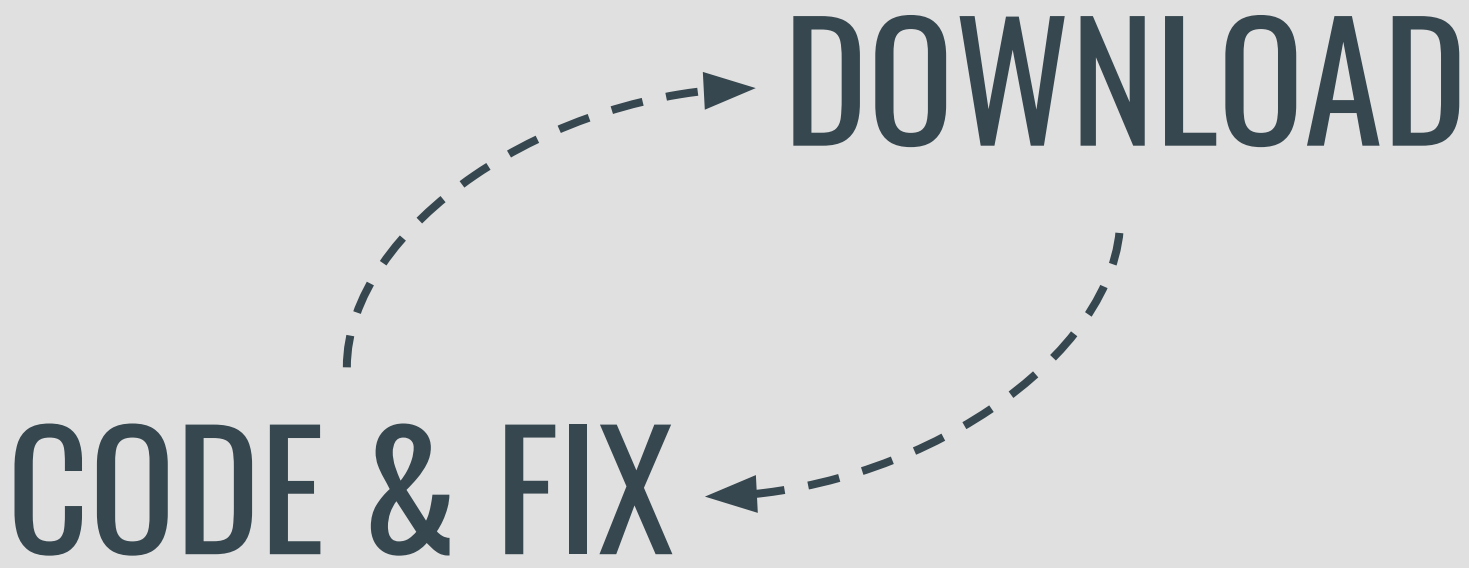
Ładowanie danych

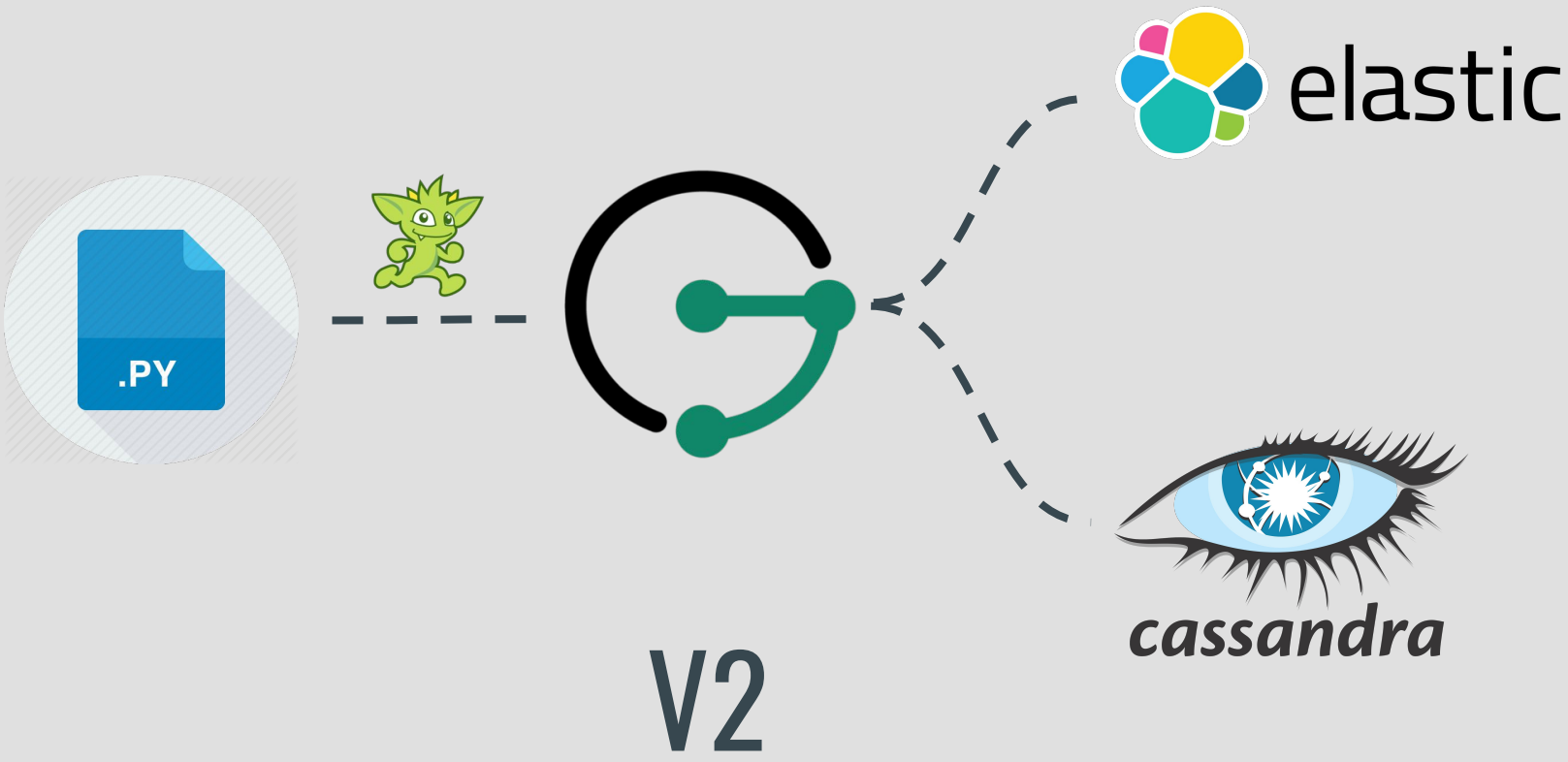
JanusGraph

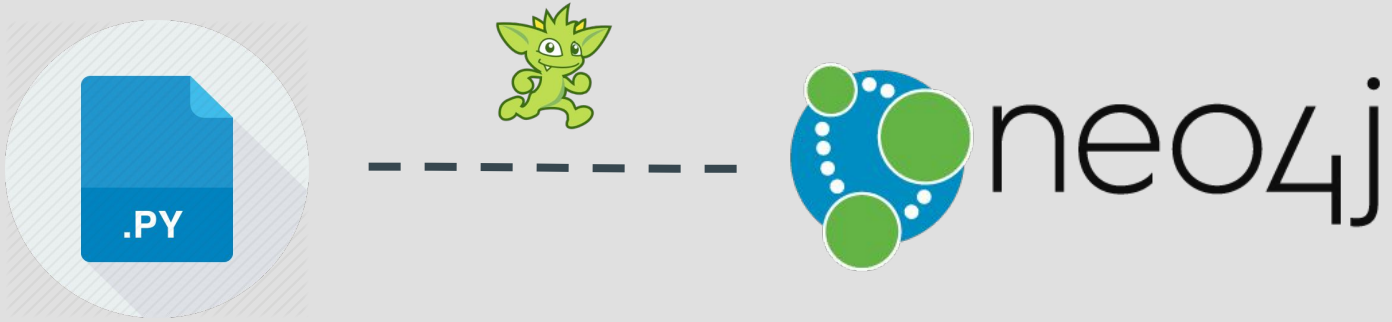
Distributed Graph Database



V1


DOWNLOAD
CODE & FIX


.PY
elastic
cassandra
V2



V3

Analiza i selekcja cech

Gremlin

Graph Traversal Language

Wybrane wybrane cechy ;)

- Wykorzystane technologie (> 192),
- Liczba gwiazdek,
- Liczba niezamkniętych issues,
- Liczba releases,
- Liczba zamkniętych kamieni milowych,
- Liczba kontrybutorów zatrudnionych,
- Liczba followers kontrybutorów.

Uczenie maszynowe

Problem klasyfikacji binarnej.

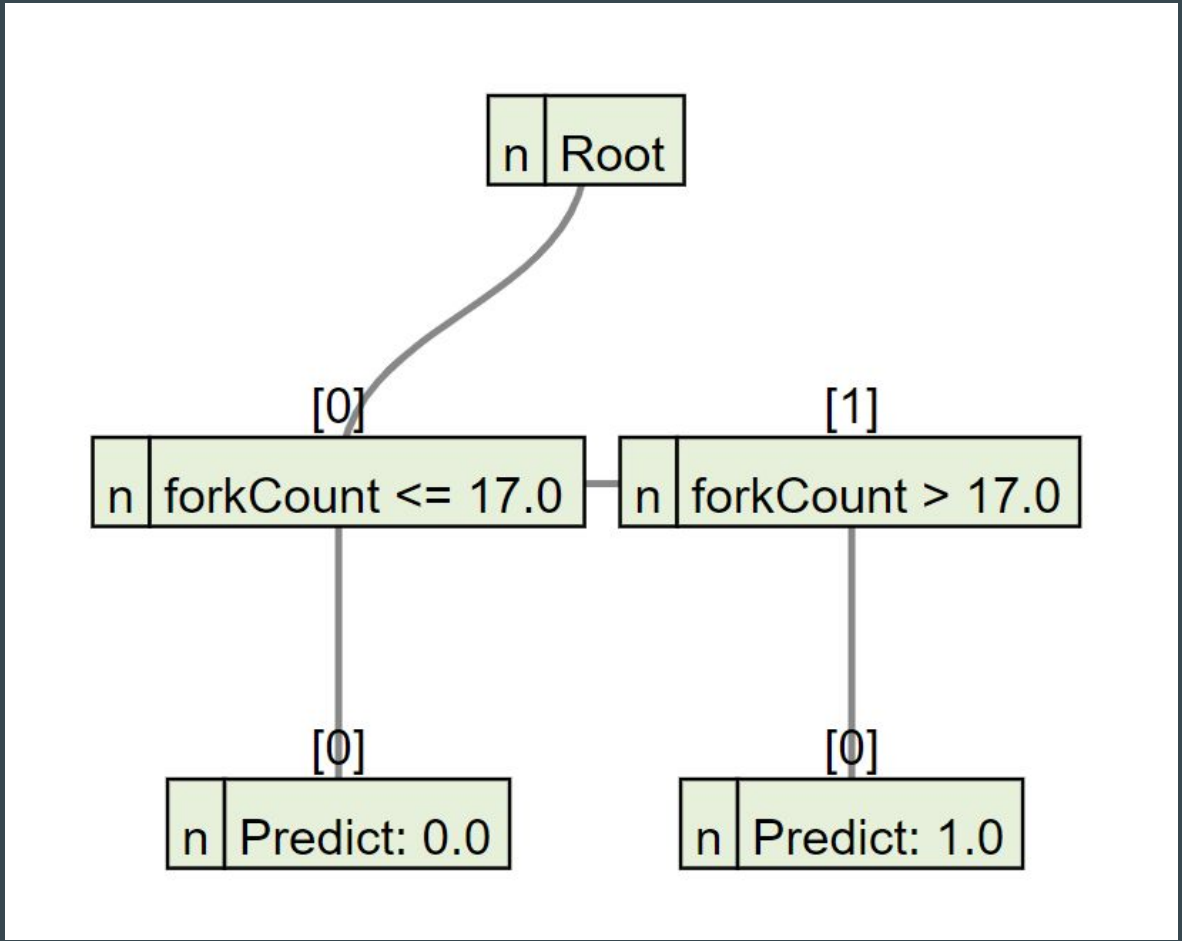
Czy dane repozytorium uzyska commit w ciągu najbliższych 3 miesięcy?

Preprocessing danych

- Usunięcie rekordów bez daty ostatniego commita
- Naprawienie rekordów w których data ostatniego commita była wcześniejsza niż data utworzenia repozytorium (głównie przy forkach)
- Utworzenie binarnej zmiennej objaśnianej (Czy był commit przez ostatnie 3 miesiące?)
- Usunięcie rzadko występujących języków
- Połączenie kolumn opisujących podobne cechy
- Utworzenie zmiennych binarnych dla niektórych zmiennych ciągłych o mocno skośnym rozkładzie

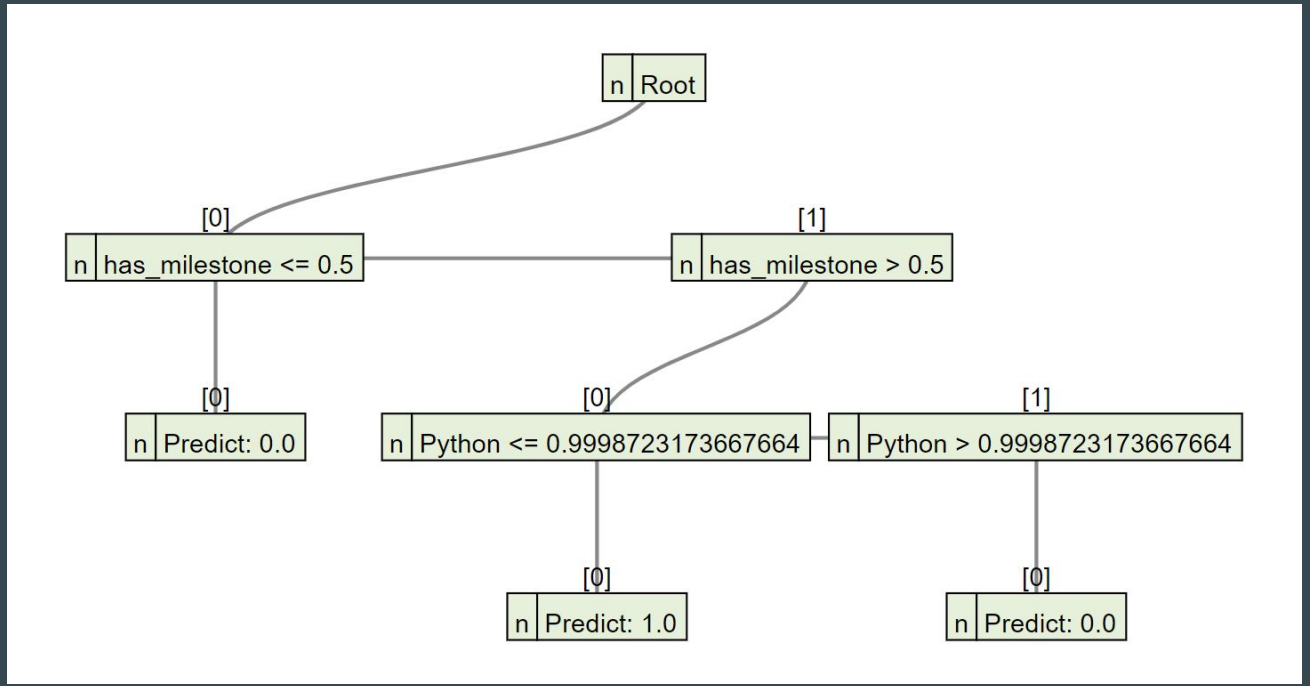
Wyniki #1

Area under PR = 0.24
Area under ROC = 0.79
Accuracy = 0.84
maxDepth = 1



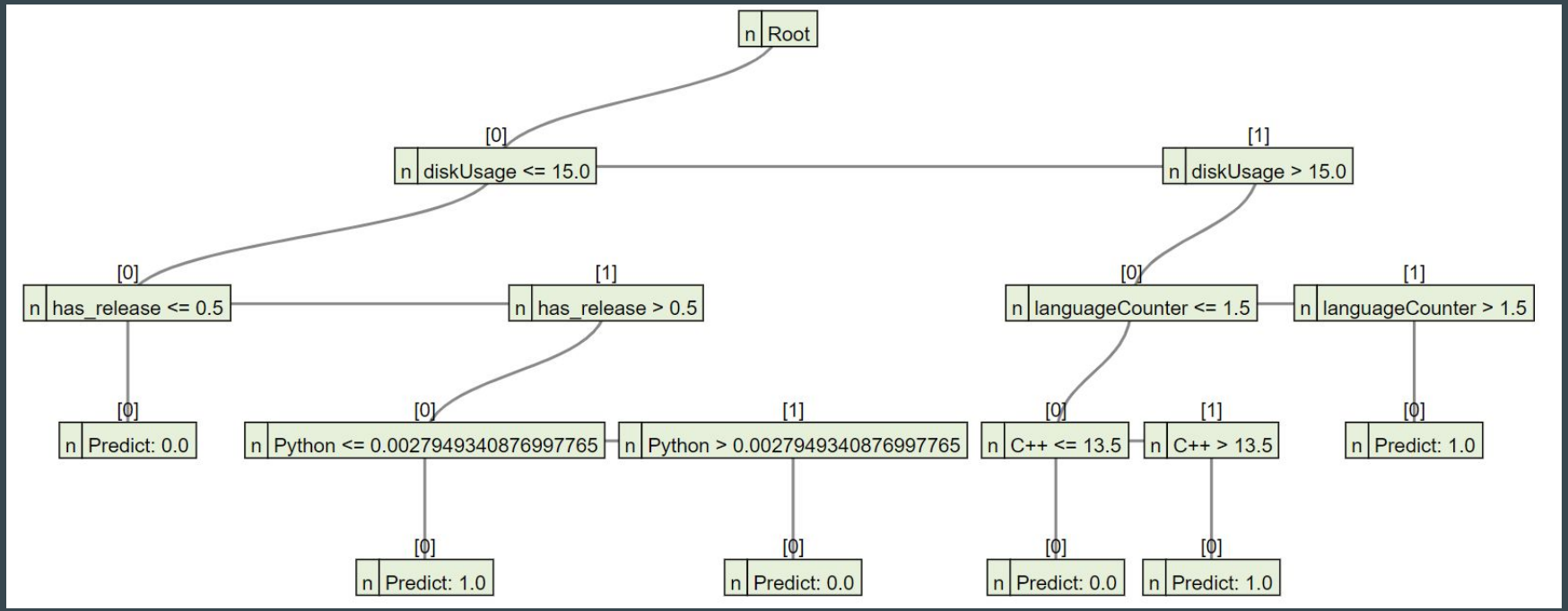
Wyniki #2

Area under PR = 0.44
Area under ROC = 0.85
Accuracy = 0.81
maxDepth = 2



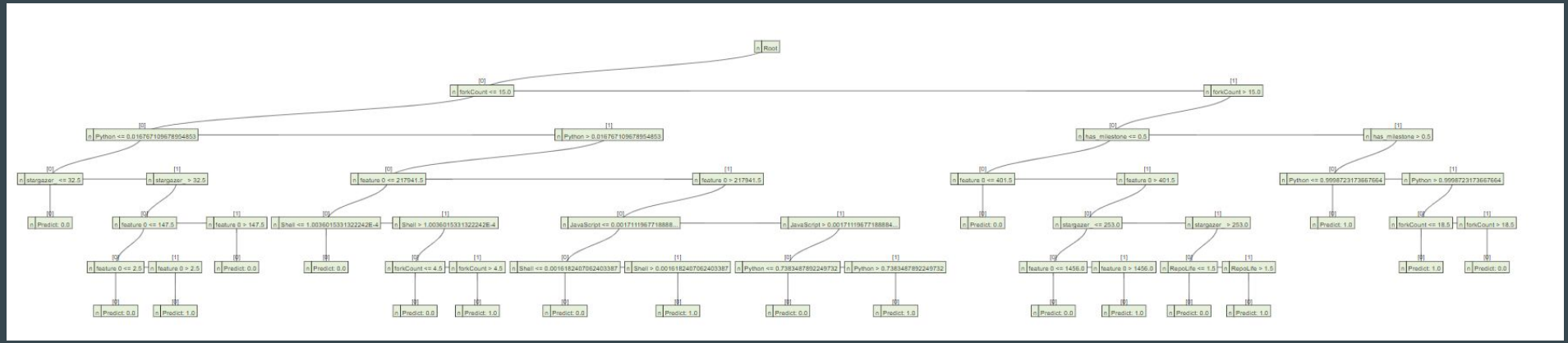
Wyniki #3

Area under PR = 0.30 Area under ROC = 0.76 Accuracy = 0.85 maxDepth = 3



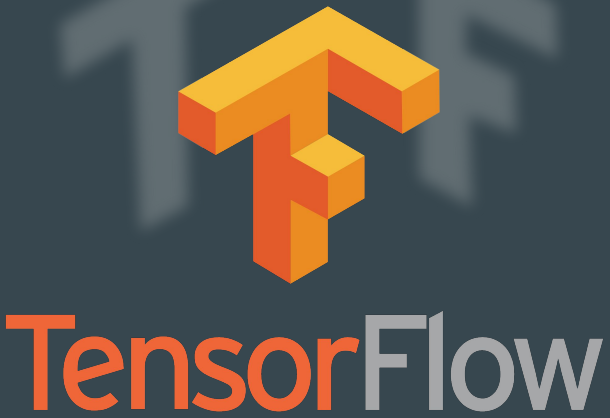
Wyniki #4

Area under PR = 0.32 Area under ROC = 0.85 Accuracy = 0.90 maxDepth = 5



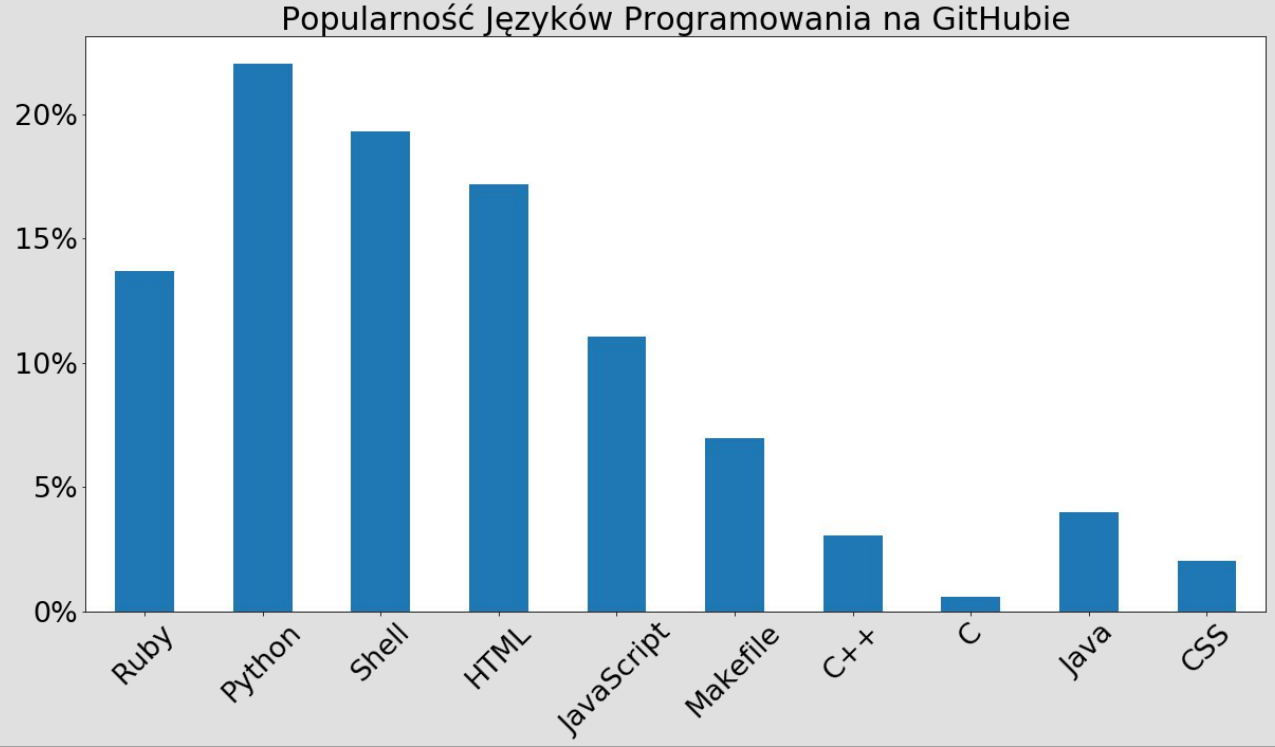
Wnioski

- Duże znaczenie miała ilość forków,
- Jeżeli python był jedynym językiem w projekcie to często oznaczało to, że ten projekt już umarł
- Projekty w których były milestony często miały większą szansę na przetrwanie



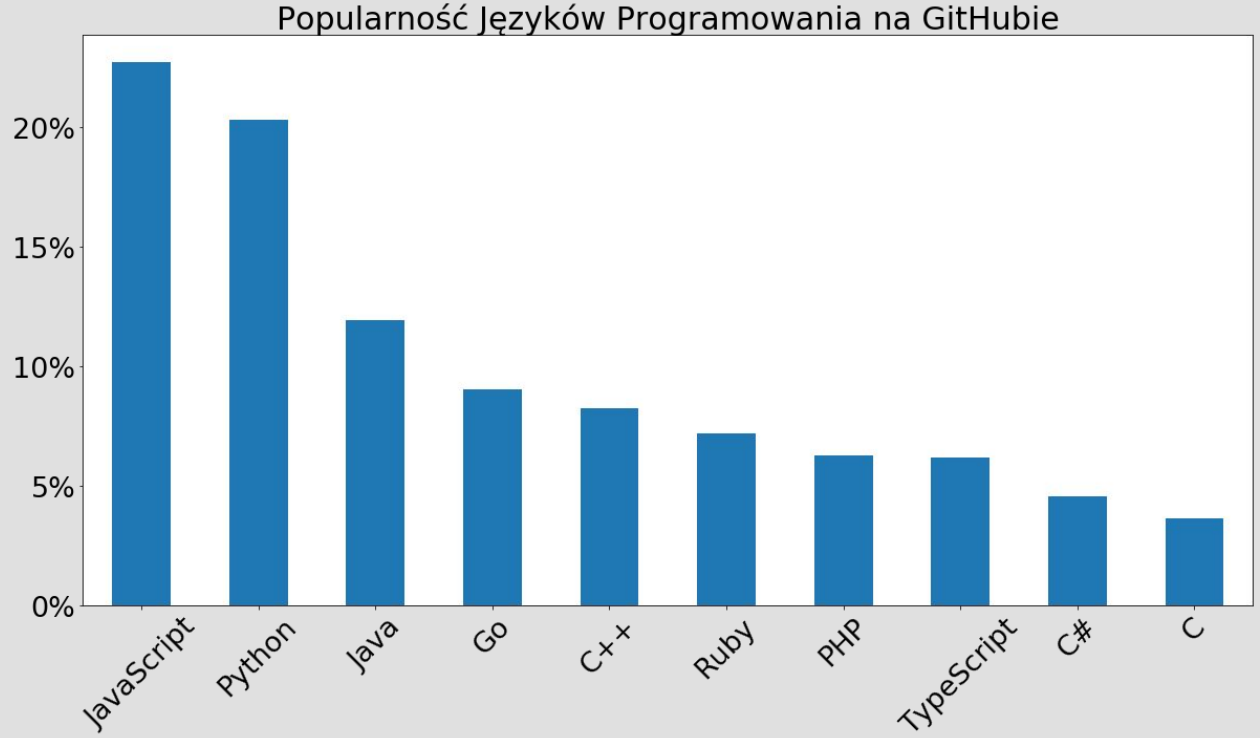
Wnioski

Nasze Wyniki



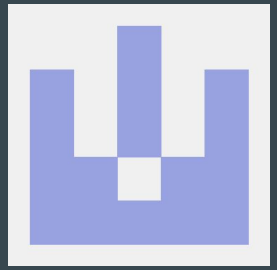
Wnioski

Wyniki z większego zbioru na podstawie Pull Requestów



Dziękujemy za uwagę!

Michał Kośmider



Mateusz Dorobek

Artur Gajowniczek

Stanisław Pawlak