Extrakcia metadát zo zdrojových kódov

Tomáš Kramár, Mária Bieliková

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies

PeWe@FIIT
personalized web group

Kľúčové slová

Značky

Metadáta

Pojmy

Pomenované entity

```
def size
  @size ||= load_size
end

def frequency(token)
  contents[token]
end
```

Štruktúra zdrojových kódov

- Kód
- Zakomentovaný kód
- Komentár ku kódu v prirodzenom jazyku

TF.IDF

- Ako rozdeliť zdrojový kód na kandidátne kľúčové slová?
  - Tokenizácia na alfanumerické reťazce
  - Jazykovo špecifický tokenizer
- Ako zvoliť korpus dokumentov?
  - Samostatný korpus pre každý programovací jazyk a pre komentáre v prirodzenom jazyku
  - Samostatný korpus pre každý projekt
  - Korpus podľa typu zdrojového kódu?

Realizácia

- Projekt PerConIK
- Webová služba (jockey)
- Experiment s doménovými expertami
- ~75% presnosť