Bc. Marek Adamovič

**Parsovanie receptov z wikipédie**

Vyhľadávanie Informácií

Cvičenie: Utorok 19:00 – 20:40

Ing. Igor Stupavský 2022/2023

# Popis projektu

Náš projekt parsuje recepty a ingrediencie z wikipédie. Následne umožňuje používateľovi zadať suroviny (napríklad také, ktoré má doma) a program mu vráti 3 recepty s najlepšou zhodou, kde sa tieto suroviny nachádzajú. Druhá možnosť pre používateľa je vyhľadať všetky recepty pre konkrétnu surovinu.

Program má viacero možných nastavení, vďaka čomu nemusí pri každom spustení prehľadávať celú wikipédiu (čo mu trvá okolo dvoch hodín). Je tu možnosť uloženia/nahratia vyparsovaných receptov z/do programu, vďaka čomu má používateľ výsledok do pár sekúnd.

Ako motiváciu pre náš projekt berieme, že v dnešnej dobe je veľkým problémom vyhadzovanie jedla v domácnosti. To môže byť čiastočne spôsobené tým, že ľudia nevedia, čo môžu navariť zo surovín, ktoré doma majú. Vďaka našej aplikácií vedia veľmi jednoducho zadať suroviny, ktorých sa „potrebujú zbaviť“ a o chvíľu majú návrhy receptov, ktoré s nimi môžu vyskúšať.

# Súčasné riešenia problému

# Moje riešenie

Pre moje riešenie som zvolil skriptovací jazyk python s použitím frameworku Apache Hadoop a Spark pre paralelné spracovanie veľkých dát.

Knižnice a technológie, ktoré som v projekte použil:

* Json -> pre ukladanie / načítanie vyparsovaných dát
* Nltk -> na lematizáciu textu wikipédie
* Time -> meranie času (napríklad pre porovnanie rôznych prístupov)
* Pyspark -> paralelné spracovanie dát
* Docker -> pre ľahšie spúšťanie programu na rôznych strojoch

Projekt sme začali riešiť lineárnym spracovaním a mimo dockeru. Vyhľadávali sme Keď fungovala prvá verzia programu, otestovali sme ju na menších dátach.

# Použité dáta

# Spustenie

# Zdroje