Seminarium magisterskie II
2019/2020
Maciej Wiercioch

Temat pracy magisterskiej:

‘‘Przetwarzanie dużych zbiorów danych z wykorzystaniem PySpark”

Streszczenie:

Tematem pracy magisterskiej jest przetwarzanie dużych zbiorów danych z wykorzystaniem PySpark. Zbiorem danych są reklamy polityczne publikowane w serwisie społecznościowym Facebook. Dane zostały zebrane z kont użytkowników Facebooka, którzy wyrazili zgodę na zainstalowanie rozszerzenia odpowiadającego za selekcjonowanie reklam i oznaczaniu ich jako polityczne. Celem pracy jest odnalezienie zależności między danymi posługując się narzędziami służącymi do analizy dużych zbiorów danych oraz wykorzystaniu algorytmów uczenia maszynowego.

Wstęp:

Do analizy zbioru danych przydatne okazuję się być narzędzie Apache PySpark pozwalające na m.in. uruchomienie obliczeń w klastrze co pozwoli na skorzystanie z większych zasobów pamięciowych. Apache PySpark oferuje szereg funkcjonalności wspomagających przetwarzenie dużych zbiorów danych. PySpark pracuje wydajniej niż inne podobne narzędzia dostępne na rynku. Pozwala na uruchamianie obliczeń lokalnie, w klastrze, w chmurze oraz umożliwia współprace z takimi technologiami jak: Hadoop, Cassandra, Apache HBASE, Mesos czy Kubernetes.
Apache Spark jest podzielony na komponenty, z których każdy odpowiada za innego typu obliczenia lub operacje na danych. Do komponentów Apache PySpark należą: Spark SQL, Spark Streaming, Machine Learning Library i GraphX. Dzięki tym komponentom mamy wyróżnioną część SparkSQL, która odpowiada za komunikację z bazami danych oraz plikami JSON i CSV. Spark Streaming pozwala na strumieniowe przetwarzanie danych. Machine Learning Library dostarcza algorytmów uczenia maszynowego niezbędnych do szczegółowej analizy. Natomiast GraphX pozwala na przedstawianie danych za pomocą grafów. Podstawowowym typem danych w Apache PySpark jest RDD (Resilient Distributed Dataset) na którym wykonuje się transformacje i akcje. Transformacje przekształcają instancję RDD i w inną instancję RDD, natomiast akcje generują wyniki.
W zbiorze reklam politycznych można wyróżnić następujące nazwy danych: id, html, political, not political, title, message, created at, updated at, lang, impressions, political probability, suppressed, advertiser, images. Mając taką liczbę kolumn do przetwarzania będzie można sformułować wnioski o instytucjach umieszczających reklamy, języku używanym w reklamach jak i również informację o tym jaka treść cechuje reklamę polityczną.

Bibliografia:

https://spark.apache.org/