Szövegbányászat és gépi tanulás Apache Sparkkal

Az Apache Spark jelenleg az egyik legnépszerűbb nyílt forráskódú számítógépes klaszter keretrendszer. A saját gépi tanulásos könyvtárának köszönhetően (MLlib) könnyen skálázhatóvá teszi a szövegbányászatban alkalmazott számos előkészítő jellemző kiemelési és gépi tanulási feladatot. Ráadásul mind Pythonból mind R-ből használható.

Az oktatás a “big data” világát bemutató rövid bevezetéssel fog indulni. Ezt követően felállítunk egy Apache Spark klasztert a Google Cloudban. Az MLlib könyvtár eszközeit R-ben fogjuk kezelni a sparklyr csomag segítségével egy RStudio szerveren keresztül. Szó lesz még többek között az Apache Spark klaszter beállításairól, a Hadoop elosztott fájlrendszer használatáról, a megfelelő dokumentáció megtalálásáról, és arról, hogy a PTI-ben mi hogyan használjuk a CAP és POLTEXT projektek keretében a SZTAKI Párhuzamos és Elosztott Rendszerek Kutatólaboratórium munkatársainak segítségével az MTA Cloudban felállított Apache Spark klasztert.

Fontos, hogy ahhoz, hogy az oktatási anyagban szereplő lépéseket a saját gépén is végre tudja mindenki hajtani, kell egy aktivált Google Cloud fiók (az oktatáshoz elég az ingyenes próba: https://console.cloud.google.com/freetrial), vagy egy Ubuntu 18.04 operációs rendszert futtató legalább 4+ maggal/cpu-val rendelkező gép, akár helyben, akár távoli ssh elérésen keresztül (mindkét esetben sudo/root jogosultsággal és internetkapcsolattal).

Kacsuk Zoltán, TK PTI