Data 612 - Research Question #1

Music Recommendation at Scale with Spark - Video Review

Joseph Simone

Overview

The themes I found most interesting during this seminar were:

- The strategy of related artists
- Different strategies used by different compete or equivalent companies, from manual curation to really intensive labeling
- How spotify uses collaborative filtering as their primary strategy: finding relationships to what they are listening to based on the song but not further metadata
- Spotify uses implicit ratings - 1s and 0s for listened to or not - but more plays makes your weight greater in their loss function
- Hadoop at Spotify in 2009 was a couple machines, by 2014 was a 700 Node Data Center in London
- Spark reads the rating matrix from disk ONCE and keep in the rest in cache memory
- Half Gridify Method, ALS implimentation in Spark.
- The run times of Hadoop vs. Spark

Final Thought

One of the main themes of this video was to show how important experimentation is to the furture of a company's Recommendation Systes. There are various ways that will take you to a "suffient" result.

However, the efficiency of achieving that result is very important as well. In addition, where you want theresources to stem from. For example, having enough memory for the fastest methods, while, having to sacrificetime with the continuously shuffling the data.

One of the most important points was the fact that it is not all about how fast a process can run. Efficiency of an algorithm will always lead to the most economic use of the resources at hand.

Today with cloud services, it is much easier to bypass the amount of harware you can run.

RSpark is a great implementation of this and, as mentioned in the video, in Python as well.

Distrubuted Systems is definitely the most efficent way to run real-time data feeds.