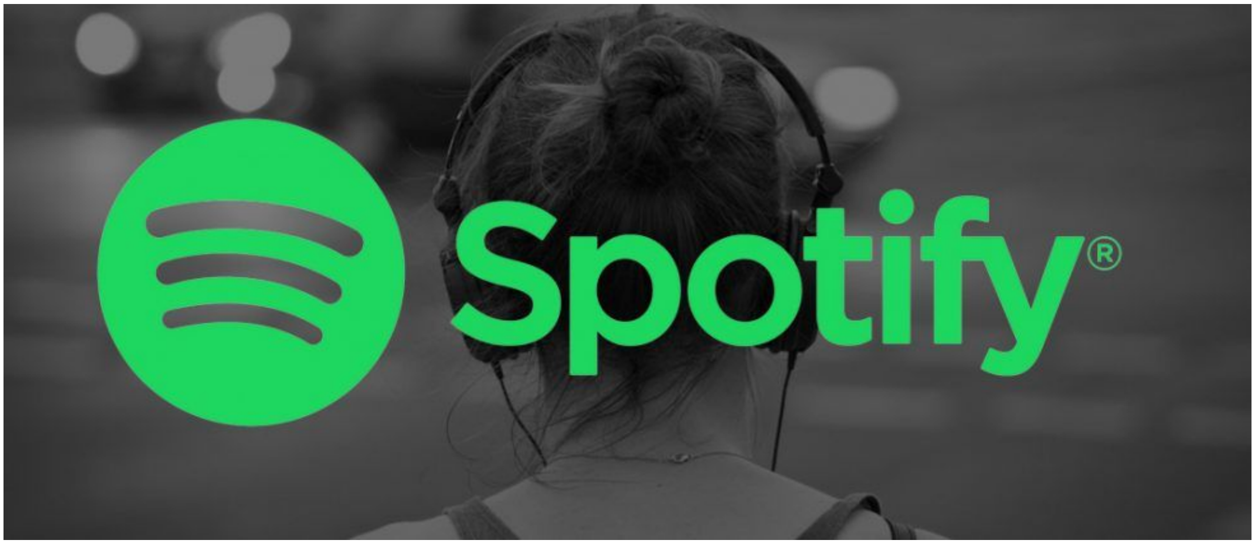
Dan Fanelli
DATA 643 Special Topics: Recommender Systems
Discussion 2

"Summarize what you found to be the most important or interesting points."



In the YouTube video entitled “Music Recommendations at Scale with Spark - Christopher Johnson (Spotify)”, Chris Johnson discussed:

- The strategy of related artists
- The strategy of look at the actual underlying audio content
- Text analysis of music blogs and other textual sites that might refer to the songs or artists
- How spotify uses collaborative filtering as their primary strategy: finding relationships to what they are listening to
- 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 computers, by 2014 was a huge network
- Spark lets them read the rating matrix from disk ONCE and keep in the rest in RAM
- He's talking about RDDs, but this was 2014 - already out of date - Spark has switched to a DataFrame API rather than RDD (I believe)

After years of seeing Hadoop on job posting boards, you become pretty convinced that it must be a very important and powerful technology. Then when you see the charts claiming that Apache Spark is 30x faster than Hadoop, you say “How can this be?”. But that is simply the huge difference between RAM and Disk, the difference in speed between electricity and magnet shillings.

I have a 3 computer Spark network at my home, which was very exciting to me to get set up. My master is my primary windows laptop that I do most my work on, and my 2 slave machines are an old laptop with Linux Mint and my Apple laptop. Windows computers don't seem to be possible to be slaves becase of some SSH problems, but apparently they can be the master. Spark is not just cool in that it can do the job, but also in its admin console. From your master computer, you can see a view of how each of the slaves are doing, how much of their RAM they are eating, etc. Then you can actually click those links and now be on an admin console hosted on the slave machines itself for even more detail. The little spotify network they showed from 2009 made me think of my 3 machine spark network also.

I see Apache Spark somewhat analagous to HTML5 - its definitely the best, but it could take a long time to catch on. I don't see tons of Spark jobs listed just yet, but being a java guy, I think Spark (and JVM based Scala) might be my path into this world of Data Science. If you know java, scala is not that tough to pick up (definitely not easy though). Spark can run and host code written in R and Python (and core java too), but I think there's a bias towards Scala becase that's its native language.

If Spark stays as hot as it is, I think there will be very similar alternatives coming to the market in Python and other core languages. (Is this what Tensorflow is? Gotta look into that...)