Suosittelijajärjestelmän rakentaminen Apache Sparkilla

Jonne Pihlanen

Tampereen Teknillinen Yliopisto

jonne.pihlanen@student.tut.fi

November 12, 2018

Sisältö

Tavoitteet

Elokuvien suosittelua Apache Sparkin avulla. Koska työ tehtiin vain omaksi iloksi, ajatuksena oli hyödyn maksimointi ja tavoitteeksi valittiin: uuden ohjelmointikielen opettelu, pilvipalveluun tutustuminen sekä Spark-ohjelmistokehykseen tutustuminen. Akateeminen osuus hoidettiin etsimällä tutkimuspaperi, johon perustin oman mallini opetuksen.

Suosittelijajärjestelmät

- Suosittelijajärjestelmät ovat joukko tekniikoita ja ohjelmistoja, jotka tarjovat suosituksia mahdollisesti hyödyllisistä tuotteista.
- Tässä työssä keskitytään yhteisösuodattamiseen (collaborative filtering): jos käyttäjät pitivät samankaltaisista tuotteista aikaisemmin, he luultavasti pitävät samoja tuotteita ostaneiden henkilöiden suosituksia merkityksellisinä.
- Muistipohjainen (Käyttäjäpohjainen, Tuotepohjainen)
- Mallipohjainen (ALS)

Alternating Least Squares (ALS)

- Spark ALS yrittää arvata arvostelumatriisin A kahden tekijämatriisin, X ja Y, tulona.
- Perinteinen lähestymistapa on iteratiivinen.
- Jokaisen iteraation aikana toista tekijämatriisia pidetään vakiona ja toinen ratkaistaan käyttäen MSE-algoritmia.
- Juuri ratkaistua tekijämatriisia pidetään vuorostaan vakiona kun ratkaistaan toista tekijämatriisia.
- Löydetty ratkaisu takaa minimaalisen MSE:n, jokaisella askeleella ratkaisu voi joko pienentyä tai pysyä samana, mutta ei kasvaa.
- Algoritmi vaihtelee (**alternates**) näiden kahden askeleen välillä konvergenssiin asti.
- ALS on siis kaksivaiheinen iteratiivinen optimointiprosessi.

Root Mean Square Error (RMSE)

Työn mallin opetusvirheen arviointiin käytettiin RMSE -metriikkaa.

- kenties suosituin ennustettujen arvosteluiden tarkkuuden evaluointiin käytetty metriikka.
- RMSE:n tuntemiseksi tulee tuntea ensin MSE (Mean Square Error).
- MSE on virheiden neliöiden keskiarvo ja se voidaan laskea neliöimällä jokaisen havainnon virhe ja laskemalla virheiden neliöiden keskiarvo.
- RMSE voidaan puolestaan laskea ottamalla neliöjuuri MSE:stä.

Pienimmän RMSE:n omaavan mallin voidaan sanoa sovittuvan parhaiten opetusdataan.

Scala

- Scala on moniparadigmainen ohjelmointikieli, joka tukee sekä olio- että funktionaalista ohjelmointia.
- Muuttumattomat tietorakenteet ja funktiot ensimmäisen luokan kansalaisina.
- Luokat ja oliot, kapselointi, perintä, moniperintä.
- Scala on staattisesti tyypitetty kieli ja sillä kirjoitetut ohjelmat käännetään Scala-kääntäjää käyttäen.
- Scala on JVM-kieli, joten Scala käännetään Java-tavukoodiksi, jota voidaan ajaa missä tahansa Java-virtuaalikoneessa.

Apache Spark

- Yleispätevä analytiikkasovelluskehys
- (RDD), Dataset, DataFrame, MLlib
- Nopeus + käyttäjäystävällisyys
- muistissa tapahtuva prosessointi
- Parhaimmillaan hurjasti nopeampi kuin Hadoop: Ei rajoitettu kahteen steppiin => Spark voi suoriutua yhdellä monitasoisella ajolla kun taas MapReduce tekee aina kaksi ennalta määrättyä toimea (map ja reduce)
- Kirjoitettu Scala-ohjelmointikielellä ja Sparkia voi kirjoittaa Scalalla, mutta lisäksi esimerkiksi Pythonilla, R:llä ja jopa Kotlinilla.

Amazon Web Services (AWS)

Amazonin tarjoama kokoelma pilvilaskentaan (cloud computing) tarkoitettuja tai sitä avustavia palveluita. Työssä käytettiin varsinaisesti kahta AWS:n palvelua: Elastic Map Reducea (EMR) ja Simple Storage Serviceä (S3).

EMR Hallittu klusterialusta esimerkiksi Apache Sparkin ajamiseen

S3 Tietovarasto

Toteutus

- EMR-konfigurointi
- Opetusdata
- Projektin rakenne
- Opetusdatan lataaminen
- Mallin opettaminen
- Opetusvirheen evaluointi
- Ennustaminen

Sisääntulot

Tunniste	Nimi	Arvostelu
112897	The Expendables 3 (2014)	4.0
116887	Exodus: Gods and Kings (2014)	4.0
117529	Jurassic World (2015)	4.0
128520	The Wedding Ringer (2015)	4.5
122882	Mad Max: Fury Road (2015)	4.0
131013	Get Hard (2015)	4.0
132796	San Andreas (2015)	3.0
136305	Sharknado 3: Oh Hell No! (2015)	1.0
136598	Vacation (2015)	4.0
137595	Magic Mike XXL (2015)	1.0
138208	The Walk (2015)	2.0
140523	The Visit (2015)	3.5
146656	Creed (2015)	4.0
148626	The Big Short (2015)	4.5
149532	Marco Polo: One Hundred Eyes (2015)	4.5

Tulokset

- **The War at Home (1979)** Drama
- **Pearl Jam: Immagine in Cornice (2007)** Documentary, Music
- **Octopus (2000)** Adventure, Horror
- **My Brother Tom (2001)** Drama
- **Return to the 36th Chamber (1980)** Action, Comedy
- **Bob Funk (2009)** Comedy
- **Hamoun (1990)** Drama
- **Notebook (2006)** Drama, Musical, Romance
- **Patton Oswalt: My Weakness Is Strong (2009)** Documentary, Comedy
- **Deathstalker II (1987)** Adventure, Comedy, Fantasy

Johtopäätökset

Tulokset eivät olleet kovin lupaavia, enemmän omia arvosteluja ja vieläkin isommalla MovieLens-datasetillä kouluttaminen voisi kuvitella auttavan.

Loppukevennys

Bills

Date: June 2018

Download CSV | Print

Total $1,063.50

- Amazon Web Services, Inc. - Service Charges $1,063.50

Payment Summary and Tax Invoices

- Payment Summary
- Tax Invoices

+ Expand All

Details

AWS Service Charges	**$1,063.50**
Data Transfer	$0.23
Elastic Compute Cloud	$679.03
Elastic MapReduce	$178.32
Key Management Service	$0.00
Simple Queue Service	$0.00
Simple Storage Service	$0.08

Taxes

VAT to be collected $205.84

Usage and recurring charges for this statement period will be charged on your next billing date. Estimated charges shown on this page, or shown on any notifications that we send to you, may differ from your actual charges for this statement period. This is because estimated charges presented on this page do not include usage charges accrued during this statement period after the date you view this page. Similarly, information about estimated charges sent to you in a notification do not include usage charges accrued during this statement period after the date we send you the notification. One-time fees and subscription charges are assessed separately from usage and reoccurring charges, on the date that they occur.

Kiitos! Kysymyksiä?