

MapReduce - Hadoop

Rafaëlle Aygalenq - Sarah Lina Hammoutene
MSc Data Science and Business Analytics

Exercise 5 - Your own MapReduce programs

Question 5.1 - Problem 1: TF-IDF

The goal was to calculate Term Frequency-Inverse Document Frequency (TF-IDF) of a set of documents using MapReduce. We found that the 20 words which have the highest TF-IDF scores in these documents are:

LIST OF 20 WORDS WITH THE HIGHEST TF-IDF SCORES:

Term: could@defoe-robinson-103.txt	Tf Idf: 0.01109284
Term: upon@defoe-robinson-103.txt	Tf Idf: 0.01103342
Term: would@defoe-robinson-103.txt	Tf Idf: 0.00944873
Term: one@defoe-robinson-103.txt	Tf Idf: 0.00829982
Term: one@callwild	Tf Idf: 0.0071866
Term: two@defoe-robinson-103.txt	Tf Idf: 0.00703207
Term: great@defoe-robinson-103.txt	Tf Idf: 0.0068736
Term: made@defoe-robinson-103.txt	Tf Idf: 0.00683398
Term: buck@callwild	Tf Idf: 0.00677233
Term: might@defoe-robinson-103.txt	Tf Idf: 0.00602183
Term: man@callwild	Tf Idf: 0.00587427
Term: found@defoe-robinson-103.txt	Tf Idf: 0.0057247
Term: came@defoe-robinson-103.txt	Tf Idf: 0.00558604
Term: time@defoe-robinson-103.txt	Tf Idf: 0.0055068
Term: much@defoe-robinson-103.txt	Tf Idf: 0.00538795
Term: little@defoe-robinson-103.txt	Tf Idf: 0.00536814
Term: first@defoe-robinson-103.txt	Tf Idf: 0.00528891
Term: shore@defoe-robinson-103.txt	Tf Idf: 0.0052691
Term: back@callwild	Tf Idf: 0.00524934
Term: could@callwild	Tf Idf: 0.00499938
Term: upon@callwild	Tf Idf: 0.00493688

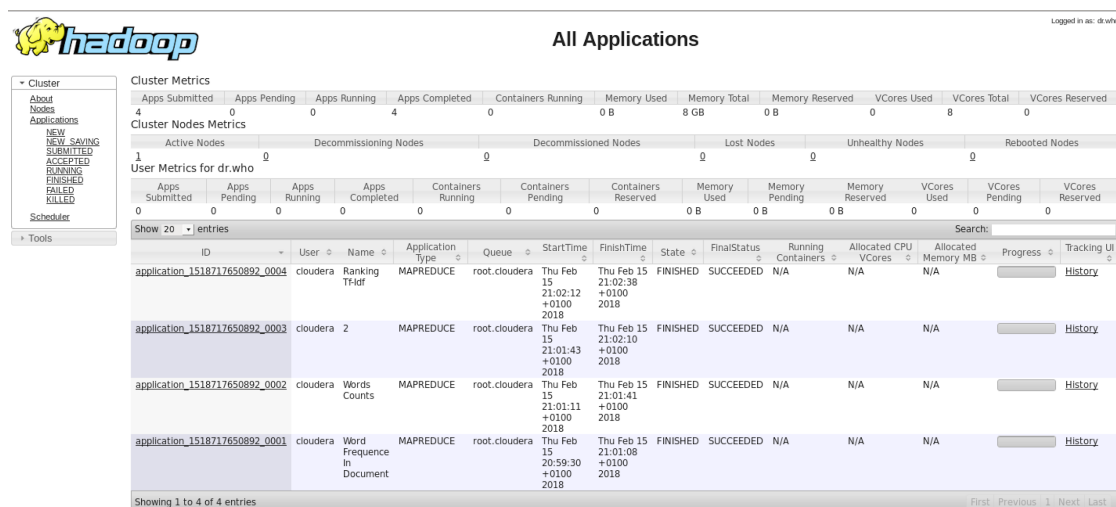


Figure 1 : Screen-shot image of our EMR Job Flows console that shows our program's COMPLETED state as well as the elapsed time

Question 5.2 - Problem 2: PageRank

The goal of this problem was to calculate the PageRank score (with damping factor 0.85) for each user in the Epinions who-trust-whom online social network. We can list the 10 users having the highest PageRank scores in this social network in descending order:

LIST OF 10 USERS HAVING THE HIGHEST PAGE RANK SCORES

Page: 18	Page Rank: 312.935546875
Page: 4415	Page Rank: 144.14703369140625
Page: 737	Page Rank: 133.42269897460938
Page: 790	Page Rank: 116.03601837158203
Page: 1753	Page Rank: 114.63996887207031
Page: 143	Page Rank: 114.12885284423828
Page: 1719	Page Rank: 113.59237670898438
Page: 136	Page Rank: 99.7525405883789
Page: 751	Page Rank: 99.36994171142578
Page: 118	Page Rank: 86.05181121826172



hadoop

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
15	0	0	15	0	0 B	8 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
1	0	0	0	0	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used	VCores Pending	VCores Reserved
0	0	0	0	0	0	0	0 B	0 B	0 B	0	0	0

Show 20 entries

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Progress	Tracking UI
application_1518717650892_0015	cloudera	Job3	MAPREDUCE	root.cloudera	Thu Feb 15 23:06:36 +0100 2018	Thu Feb 15 23:07:07 +0100 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1518717650892_0014	cloudera	Job2	MAPREDUCE	root.cloudera	Thu Feb 15 23:06:02 +0100 2018	Thu Feb 15 23:06:34 +0100 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1518717650892_0013	cloudera	Job2	MAPREDUCE	root.cloudera	Thu Feb 15 23:05:26 +0100 2018	Thu Feb 15 23:06:00 +0100 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1518717650892_0012	cloudera	Job2	MAPREDUCE	root.cloudera	Thu Feb 15 23:04:52 +0100 2018	Thu Feb 15 23:05:23 +0100 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1518717650892_0011	cloudera	Job2	MAPREDUCE	root.cloudera	Thu Feb 15 23:04:15 +0100 2018	Thu Feb 15 23:04:50 +0100 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1518717650892_0010	cloudera	Job2	MAPREDUCE	root.cloudera	Thu Feb 15 23:03:37 +0100 2018	Thu Feb 15 23:04:13 +0100 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History
application_1518717650892_0009	cloudera	Job1	MAPREDUCE	root.cloudera	Thu Feb 15 23:03:03 +0100 2018	Thu Feb 15 23:03:34 +0100 2018	FINISHED	SUCCEEDED	N/A	N/A	N/A	<div></div>	History

Figure 2 : Screen-shot image of our EMR Job Flows console that shows our program's COMPLETED state as well as the elapsed time.