Lecture 4 exercises

Daniel Skjold Toft

Cleanup from lecture 2

- Only necessary if the containers are still running from last time
  - I, personally, removes volumes in order to have a clean slate and thus avoid weird errors
  - We have created no data during these exercises that are necessary to save
- “docker kill namenode datanode1 datanode2 datanode3”
- “docker containe rm namenode datanode1 datanode2 datanode3”
- “docker volume rm namenode datanode1 datanode2 datanode3”

Future cleanups

- “docker-compose down” where the docker-compose.yml is located
  - Stops and removes containers
- “docker volume rm namenode datanode1 datanode2 datanode3”
  - Remove volumes

Docker-compose a Spark cluster

- “docker-compose up –d”
  - Leave out the –d if you’re uncertain of your cleanup step
- HDFS cluster with a spark-master and two spark-workers
  - Including network and volumes

Download the book!

- “docker exec –ti namenode bash”
- “apt update”
- “apt install wget”
- “wget -O alice.txt https://www.gutenberg.org/files/11/11-0.txt”
- “hdfs dfs –mkdir /txt” – Create a directory in the HDFS
- “hdfs dfs -put alice.txt /txt/”

Run a local spark job

- Does not require a spark-cluster, but only uses local resources
- Navigate to “BDDST21\Lecture4\localPyspark” and “run”
  - See the result in the console log
- Examine the folder content

Run a clustered job!

- Navigate to “BDDST21\Lecture4\clusterPyspark” and “run”
- See the result in the HDFS
  - “hdfs dfs –ls /”
  - “hdfs dfs –ls /txt-out/”
  - “hdfs dfs –cat /txt-out/part-00000”

Sentiment exercise!

- Use the
  - alice.txt book
  - The two word lists in “sentimentExercise”
  - Example.py with the loaded lists
- First, count the positive and negative words
  - Print the total score to the console

Sentiment exercise! Part 2

- Second, split the sentences in the book by ".", and find positive and negative sentences.
  - Ex. A positive word gives the sentence +1, and negative -1 (called sentiment score)
  - Find the score of all the sentences in the book!
  - Simply print it to the console
- Thirdly, think about storing this new sentiment score in the HDFS for later retrieval. Raw text, JSON, Avro or Parquet?
- Optionally, load the word files in from HDFS, instead of "in-memory"