GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
INSTITUTE OF COMPUTER SCIENCE
Software Engineering for Distributed Systems
http://www.swe.informatik.uni-goettingen.de

Data Science and Big Data Analytics
WS 2016/2017
Dr. Steffen Herbold

Exercise · Due at 2017-01-31

Data Science Challenge

Your Task

Your final task in the data science exercise is a small data science project in which you compete against each other. We give you all the same data stored in a MongoDB. Your task is to analyze it and find something interesting. The data we are giving you is about the development of Apache Zookeeper and contains information about the complete project history, including software metrics, information from the issue tracking system of the project, as well as its mailing list. The data was collected using the SmartSHARK platform, which is under development in our research group[1]. A full description of the data structure can be found online[2].

You have two options to access the data.

- Use a running instance of MongoDB running at the GWDG. The host of the DB is 141.5.113.177:27017, the name of the database is `smartshark_test`. To use this endpoint, you need to use the credentials we provide for each group.
- Download a dump of the database and load it into a local MongoDB. The dump can be downloaded from our website[3]. To install MongoDB and load the dump you can find lots of tutorials using Google.

The `mongolite` package can be used to load the data into R. Please find some data loading examples below:

```
if(!require(mongolite)) install.packages("mongolite")
library(mongolite)

# URL for connecting to the remote MongoDB hosted at the GWDG
# You must insert your DB credentials here
MONGO_URL = "mongodb://username:password@141.5.113.177:27017/smartshark_test"

# URL for connecting to a local MongoDB.
MONGO_URL = "mongodb://localhost:27017/smartshark_test"

# create connection to people collection
con_people = mongo(collection="people", url=MONGO_URL)
```

[1] `https://github.com/smartshark/`
[2] `http://smartshark2.informatik.uni-goettingen.de/documentation/`
[3] `https://user.informatik.uni-goettingen.de/~sherbol/smartshark_test20170112.gz`

```
# fetch all data from the people collection into a data frame
people = con_people$find()

# perform query on DB to fetch only people with username "zookeeper-user"
people = con_people$find('{"username" : "zookeeper-user"}')

# determine the latest commit of a project
con_commit = mongo(collection="commit", url=MONGO_URL)
commits = con_commit$find(fields='{"_id":1, "committer_date":1}')
print(paste("latest commit:", max(commits$committer_date)))
latest_commit_id = commits[which.max(commits$committer_date),1]

# find all code entity information in latest commit
con_codeentitystate = mongo(collection="code_entity_state", url=MONGO_URL)
query_str = paste('{"commit_id" : {"$oid" : "', latest_commit_id,'"}}', sep="")
code_entities = con_codeentitystate$find(query_str)
```

Due to the variety of information in the data, you have lots of possibilities for the analysis of the data. We have numerical data (e.g., software metrics), we have textual data (e.g., diffs, comments), and we have data over time (e.g., issue events, commits). Basically anything we did in the lecture and exercise could be applied in some way to this data. If you already have knowledge about additional data mining techniques, you may as well apply those. Which parts of the data you use, how you prepare it, and how you analyze it is completly up to you. It is also okay to use something other than R. You are on your own.

A note of caution: the overall DB is roughly 10 GB large. If you try to load everything you may have trouble. Especially the `code_entity_state` collection is very large, the others are much smaller.

Presentation and Voting

Each group must give a five minute presentation, with one minute for short questions. Within this presentation, you should briefly describe which data you analyzed, how you treated it, which models you used and your key findings.

Afterwards, everybody in attendance will vote to determine the best project. Each group votes for the best project (3 points), second best (2 points), and third best (1 point). The one with the most points wins.

Submission of Presentation and Project Title

In order to allow the presentation session to run smoothly, each group must do the following.

- Send the title for the project which will be printed on the voting sheets latest on Sunday, Jan 29th.
- Send the presentation latest on Jan 31st, 10:00 o'clock to Steffen Herbold via Email (herbold@cs.uni-goettingen.de). The presentations must be either PDF, PPT, PPTX, or ODP. Other formats are not allowed.

Advertisement

You liked working with the data? You want to dive deeper into the analysis or work with Apache Spark? Just contact Steffen for a project work or BSc/MSc thesis topic.