Stockholm Software Craftsmanship Community   
Scriptastic Session 

[Background - Reddit](#h.j70er6t0i9bc)

[System design](#h.zseit2z8ank)

[Link aggregator](#h.qyzjig2b19ie)

[Database analyser](#h.erwiclfzv44q)

[User front end](#h.jeg761uoizcz)

[Recommended tools](#h.2vo38pfe0v4e)

[Accessing the amazon instance](#h.khxfdtnb4gj3)

## Background - Reddit

<https://www.reddit.com/> is a social link aggregation site where users can post content, comment and upvote/downvote content. Content can be links to videos, homepages, images, or just plain text. There are subreddits which specialize in specific content - each one accessible through [https://www.reddit.com/r/${SUBREDDIT](https://www.reddit.com/r/SUBREDDIT)} where SUBREDDIT is the name of the subreddit in question.

Each users history can be seen by accessing [https://www.reddit.com/user/${USER](https://www.reddit.com/user/$%7BUSER)} where ${USER} is the name of the user.

The task - we will construct a Reddit stalking system in which we can analyze the reddit data from a user and his/her friends.



## System design

The Reddit analysis system is to be located on one machine, but consist of several small programs which cooperate in order to create the grand system, i.e, we are constructing our Reddit analysis system by composing it of microservices.



The full system details are not given - there are minor parts you will have to figure out yourself by talking between the groups.

### Link aggregator

Input to this is a file located in ${DEFINED\_BY\_LINK\_AGGREGATOR\_GROUP} which is read and wiped every time the service is started. The contents of the file is a list of reddit usernames, one username per line. The service will not start unless there is at least one (1) name in the file.

The output of the service is a database file which contains one record per line, each record has the format

**${TYPE}, ${DATE}, ${SUBREDDIT}, ${OP}**

where

**${TYPE}** is either ‘S’ for submitted or ‘C’ for commented

**${DATE}** is a YYYY-MM-DD-HH-MM-SS formatted string of the event

**${SUBREDDIT}** is the name of the subreddit, excluding <https://www.reddit.com/r/>

**${OP}** is the name of the original poster, i.e., ${USER} is type is ‘S’ or commenting

on his/her own post, or the name of another user if type is ‘C’ and commenting

on another user’s post.

the output file is located in ${DEFINED\_BY\_LINK\_AGGREGATOR\_GROUP} and has the name ${DEFINED\_BY\_LINK\_AGGREGATOR\_GROUP}

### 

### 

### Database analyser

Input to this is a file located in ${DEFINED\_BY\_DATABASE\_ANALYSER\_GROUP} which is read and wiped every time the service is started. The contents of the file is a list paths to Link aggregator databases, one path per line. The service will not start unless there is at least one (1) path in the file.

The database analyser will perform the following analysis

For each user

Which subreddits are the most common for this user to operate in?

What portion of the day is this user most active?

What is the distribution of actions for this user between ‘S’ and ‘C’?

For the whole dataset

Is there any ${OP} which is prevalent in the dataset?

Is there any portion of the day that seem particularly active?

Is there any subreddit that is particularly prevalent?

The output of the service is a file located in ${DEFINED\_BY\_DATABASE\_ANALYSER\_GROUP} which contains one record per line, each record is either of the type

**${USER}**,**${SUBREDDITS}**,**${DAY\_PORTION}**,**${ACTION\_DISTRIBUTION}**

**${OP\_LIST},${DAY\_PORTION},${PREVALENT\_SUBREDDIT}**

**NOTE: the second record is the last record in the output of the service.**

where

**${USER}** is the name of the user the analysis data belongs to

**${SUBREDDITS}** is a bracket enclosed list ‘[‘ ${SUBREDDIT} ‘]’ of subreddits, in *descending* order of activity, in which this user is active.

**${DAY\_PORTION}** is a HH-MM:HH-MM formatted string indicating the span during which the user is mostly active. the HH-MM before the ‘:’ indicate the start, and the HH-MM after the ‘:’ indicate the end of the time span.

**${ACTION\_DISTRIBUTION}** is a bracket enclosed and ‘:’ (colon) separated pair ‘[‘‘${PERCENT}:${PERCENT}’]’’ where the first ${PERCENT} is the percentage ‘S’ actions for this user, and the second ${PERCENT} is the percentage of ‘C’ actions for this user.

**${OP\_LIST}** is a bracket enclosed - ‘:’ (colon) separated list ‘[‘ ${OP} ‘]’ of ${OP} in descending order, based on their prevalence in the full dataset.

**${PREVALENT\_SUBREDDIT}** is a bracket enclosed and ‘:’ (colon) separated list ‘[‘ ${SUBREDDIT} ‘]’ of ${SUBREDDIT} in descending order, based on their prevalence in the full dataset.

**The database analysis report is also stored in the report storage located in ${DEFINED\_BY\_DATABASE\_ANALYSIS\_GROUP} so that the result can be revisited.**

### User front end

The user front end is a command line based program which allows easy usage of the Reddit analysis system. It makes it possible to

1. Enter usernames for analysis
2. See a progress report while the system is running, so that the user knows what is happening, e.g.,
   1. Aggregating
   2. Producing DB
   3. Analysing
   4. Producing report
3. Displays the report
4. Based on the result from the report - immediately enter new names from the report into the next round of analysis
5. Easily browse and view past analysis reports which are stored in the Database Analysis Report Storage
6. Create backups of the Link Aggregator databases and the Database Analysis Report storage

The user front end also checks whether the entered Reddit usernames exist.

The front end produces data suitable for the Link Aggregator to consume, and consumes data produced by the Database Analyser.

## Recommended tools

This exercise is a barebone **bash** scripting session - so the idea is to stay as close as possible to the basic tools and stay out of Perl, Python etc. You may however, use available command line tools. If you don’t know where to get started - I recommend the following for this session.

Bash in itself goes without saying - you can find several targeted examples of how to utilize bash snippets in several places, but I will add my own example repository here for your to check out if you feel unsure.<https://github.com/Gianfrancoalongi/bash_by_example>

Working with files

find, tar, awk, sed, grep

Downloading stuff

wget, curl, w3m

## Accessing the amazon instance

There are two amazon instances, both are running Ubuntu Trusty 14.04 amd64 server and are the intended work platforms during this exercise.  
  
Instance1

IP 52.32.22.65

DNS ec2-52-32-22-65.us-west-2.compute.amazonaws.com

instance2

IP 52.32.88.11

DNS ec2-52-32-88-11.us-west-2.compute.amazonaws.com

You need to save the key to your machine

<http://files.meetup.com/3554542/SSCC_2015_10_29_scriptastic_session.pem>

chmod it

chmod 400 /path/my-key-pair.pem

ssh as the user ubuntu

ssh -i PATH\_TO\_PEM\_FILE ubuntu@DNS\_TO\_INSTANCE