Functional Specification

Detecting brigading on social media with machine learning

Name: Adam Pegman
Student number: 14351741

Functional Specification Contents

0. Table of contents

1. Introduction

1.1 ***Overview***

The purpose of the system is to monitor a subreddit (user created forum on reddit.com) and take actions such as creating posts, removing posts and banning users automatically. The system would assist volunteer moderators keep unsuitable posts off the subreddit. The system will achieve this by using the reddit api to retrieve posts and comments, and to automatically remove posts and users.

1.2 ***Business Context***

Reddit is a social media and news aggregator site with over 250 million users. Users on reddit can create, moderate and subscribe to ‘subreddits,’ which are forms on which links and text can be posted, each with a thread of comments. Each subreddit has it’s own topic, culture and rules. Use of language also varies considerably, with some subreddits using their own unique terminology. [1][2][3]

The moderators of each subreddit remove dozens or hundreds of comments from their subreddits every day, and must keep track of frequent offenders so they can be banned from the subreddit. This task is made more difficult by the fact that creating a new account takes less than a minute, so bans can be easily avoided.

There is an existing bot called automoderator that moderators can use to remove posts automatically, but it uses regular expressions to detect unsuitable comments. This makes the system difficult to use for many moderators, and limits the scope of what can be detected. The system should partially automate the moderation of a community in an easier to use and more versatile way, saving time and resources. [5]

1.3 ***Glossary***

Reddit is a social media and news aggregator site.

A subreddit is a user-created form on reddit where other users can post links and comments.

Web components are widgets made using Javascript, HTML and CSS that can be imported into web pages

Sass is a language that can be compiled to CSS.

Springboot is an MVC Java web framework with dependency injection.

Project Lombok is a Java library that allows automatic generation of getters, setters and constructors.

Vavr is a functional programming library for Java.

Sqlite is a lightweight file-based SQL database.

A REST API is an api where calls are made using HTTP methods and data is exchanged in JSON or XML which aims to be stateless, cacheable and layerable.

Anomaly detection is a process used to detect if a row of data is an outlier from a dataset in a supervised or unsupervised manner.

A Naive Bayesian classifier is a classifier that uses posterior probability to estimate the probability a row of data is in a given class.

A Decision tree classifier is a classifier that uses a tree of tests to determine the category of a row of data.

2. General Description

2.1 Product / System Functions

The system will be configured via a web interface, so as to be easy to use and configure. This web interface should accept credentials to use when contacting the Reddit API, the subreddits to act on, uploaded “samples” of posts, and rules on how to act on new posts. These rules will consist of a trigger and an action.

The triggers will use a variety of machine learning techniques to identify posts similar to uploaded samples, whilst the actions will be implemented with REST calls to the reddit api. The machine learning techniques employed should include Bayesian classifier or Bayesian spam filter, a k-nearest classifier, and an anomaly detection algorithm.

The actions performed by the system should include flagging a comment for manual review, making a comment as NSFW (Not safe for work), removing a comment, and banning a user.

2.2 User Characteristics and Objectives

The users of the system would be the moderators and owners of subreddits. At present most are users of the existing automoderator bot, which is configured using regular expressions in a yaml config file. These config files are passed around between moderators and are not well understood by many of the users.

These users are normally heavy internet users who should be able to navigate the web interface. These users are very familiar with the demographics that use their subreddit and the terms used, so they should be able to upload samples of the type of comments they want to flag and create rules in the system, allowing them to moderate their subreddits more effectively and without copy-pasting configs.

2.3 Operational Scenarios

User adds subreddit to monitor:
A user should be able to navigate to the web interface, click the add button under the list of subreddits, and enter the url of the subreddit. The system should then save this setting and begin retrieving comments.

User uploads sample:
A user should be able to navigate to the web interface, Click the “add group” button under the samples section, name a group of samples, and then upload the samples via csv or reddit link. These samples and groups should be stored in the database, along with statistics.

User attaches system to reddit account:
The user should be able to navigate to the settings page, hit the “Link to reddit account” button, and log into reddit in order to supply the bot with credentials. These credentials should be stored in the database.

User adds a rule for a subreddit:
The user should be able to click the name of a previously added subreddit, and click the add rule button, select a trigger and action, and then hit the submit button. The rule should be saved to the database and new comments from the subreddit should have the rule applied.

User selects a classification trigger in a rule:
The user can select the classification trigger from a dropdown, and choose the group of samples to classify within. The user should then supply an action for at least one sample, which will be run on comments that match that sample. The rule should be saved in the database when the user clicks the apply button.

User selects an anomaly trigger in a rule:
The user can select the anomaly trigger from a dropdown when creating a rule. They must then pick the sample to compare against from the list of available samples and an action. New comments will trigger the action if the comment is an outlier from the sample.The rule should be saved in the database when the user clicks the apply button.

User selects a regex trigger in a rule:
The user can select the regex trigger from the dropdown and supply a regular expression and an action. New comments will trigger the action if there is a match for the regex in the comment. The rule should be saved in the database when the user clicks the apply button.

User selects a flag action in a rule:
When a new comment is posted and a rule matches with this action, the system should make a Reddit API call to report the post to the moderation team. The post should then show up in the moderator’s queue on Reddit.

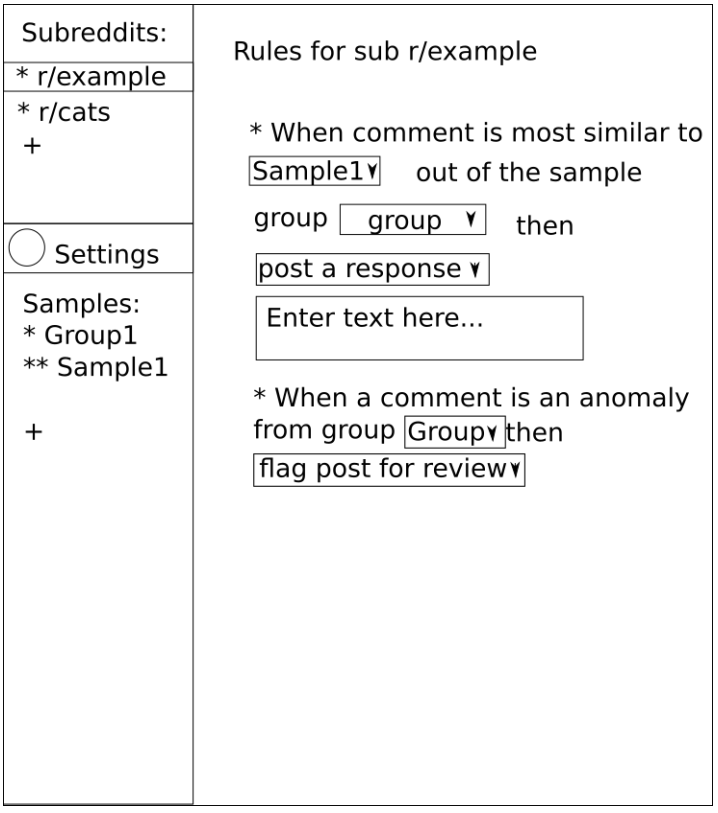
User selects a reply action in a rule:
This action must be configured with a message. When a new comment is posted and a rule matches with this action, then a Reddit API call is made to post the message as a response to the comment.

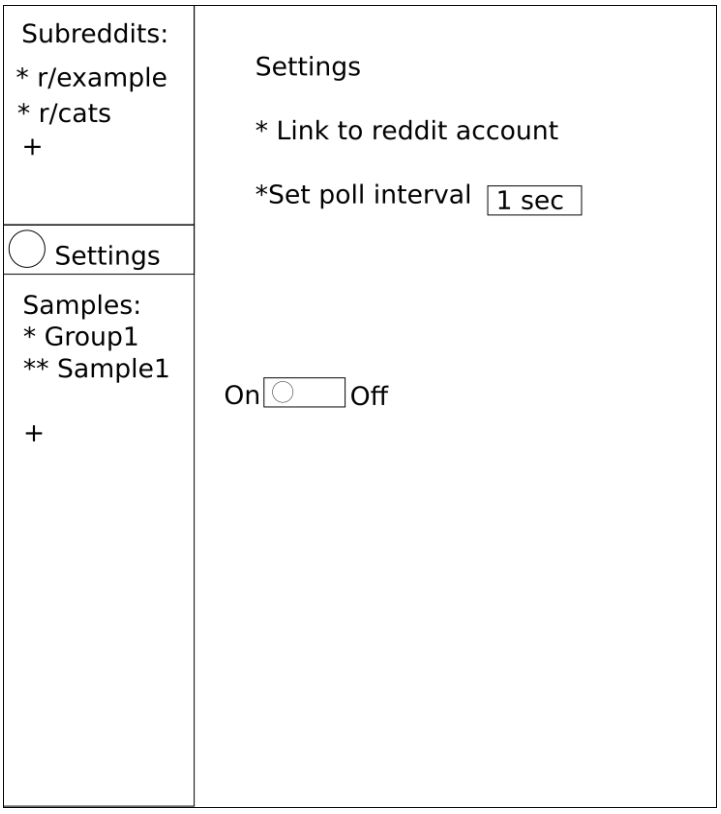
User selects a removal action in a rule:
When a new comment is posted and a rule matches with this action, a Reddit API call is made to remove the comment from the subreddit.

User selects a ban action in a rule:
When a new comment is posted and a rule matches with this action, a Reddit API call is made to ban the user who posted the comment from the subreddit.

A new comment is made to a monitored subreddit:
The comment should be retrieved either by polling new threads on the subreddit for new comments or subscribing to the thread via a websocket to the Reddit Live API.
Any rules that match the comment should have their actions executed.

UI Mockups:





2.4 Constraints

Time constraints:
The project must be completed and documented over the 12 weeks before the due date. So as to be delivered on time.

Speed requirements:
The system must run on an the lab computers, and must process each comment fast enough to make good use of the API request limit.

Resource constraints
The system should make good use of the API limit of 60 requests per minute. It should not make unnecessary requests, and should try to request only unseen comments.

Protocols
The system must use the Reddit api to interact with reddit, including OAuth for authenticating. [4]

3. Functional Requirements

Description	The system should store settings and rules in a database.
Criticality	The system would need hardcoded rules and credentials without a database or file to store them in.
Technical Issues	Formatting the settings as a DB table.
Dependencies	None

Description	The system should be configurable using a webapp.
Criticality	Without this the system would have to be configured via file or hardcoded values.
Technical Issues	Implementing a Webapp and Rest API. Linking them to a database for persistence.
Dependencies	The system requires a database in which to store the settings.

Description	The bot should use Reddit credentials configured via a web app to execute Reddit API requests.
Criticality	If this condition is not met, then either the user cannot input their credentials or the system will be unable to make api requests.
Technical Issues	Implementing a REST API and web interface to take in the credentials. Implementing OAuth login to authorise the requests.
Dependencies	Required a database in which to store the credentials and a webapp to allow them to be input.

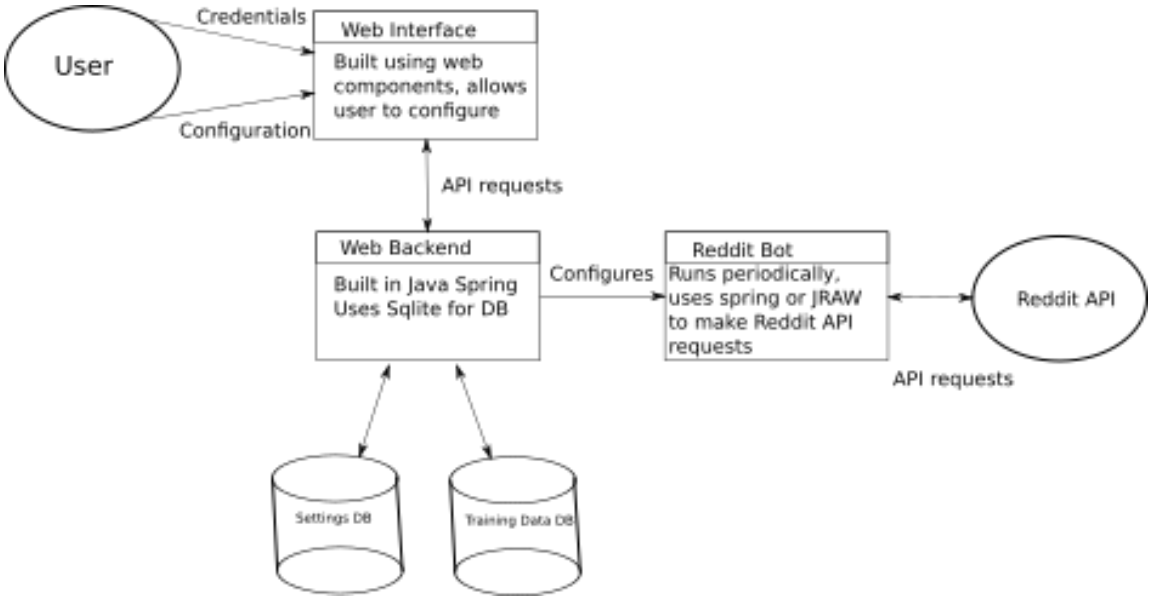
Description	The system should retrieve comments from configured subreddits in a timely manner.
Criticality	The system needs to retrieve reddit comments to act on them.
Technical Issues	There is a rate limit of 60 requests per minute. Duplicate comments must be avoided.
Dependencies	Database, Reddit OAuth

Description	The system needs to be able to classify incoming posts as one of a set of configurable classes.
Criticality	The system will not know how to respond to comments without this.
Technical Issues	Storing the sample data in a database. Storing precomputed values in the database to ensure quick classification.
Dependencies	Database, Reddit API requests

Description	The system needs to be able to detect anomalous or spam comments.
Criticality	The system will not know how to respond to comments without this.
Technical Issues	Storing the sample data in a database. Storing precomputed values in the database to ensure a fast run time.
Dependencies	Database, Reddit API requests

Description	The system needs to be able to flag, reply to and remove reddit comments.
Criticality	The system will be unable to react to comments without this.
Technical Issues	Authenticating with and making calls to the Reddit API.
Dependencies	Database, Reddit API requests, Reddit OAuth

4. System Architecture



The system's web interface will be built using Web components, and use Ajax to make requests to the system's api.

The server will be written in Java, using Springboot as a web framework and sqlite for the database.

The Reddit bot will be a scheduled task running in the background of the server app. It will use Spring's OAuthContext to authenticate with the Reddit API.

5. High-Level Design

The data for the system will be stored in a sqlite database. The settings will be key/value pairs, whereas rules will be stored in a dedicated table with attributes for each of the possible options.

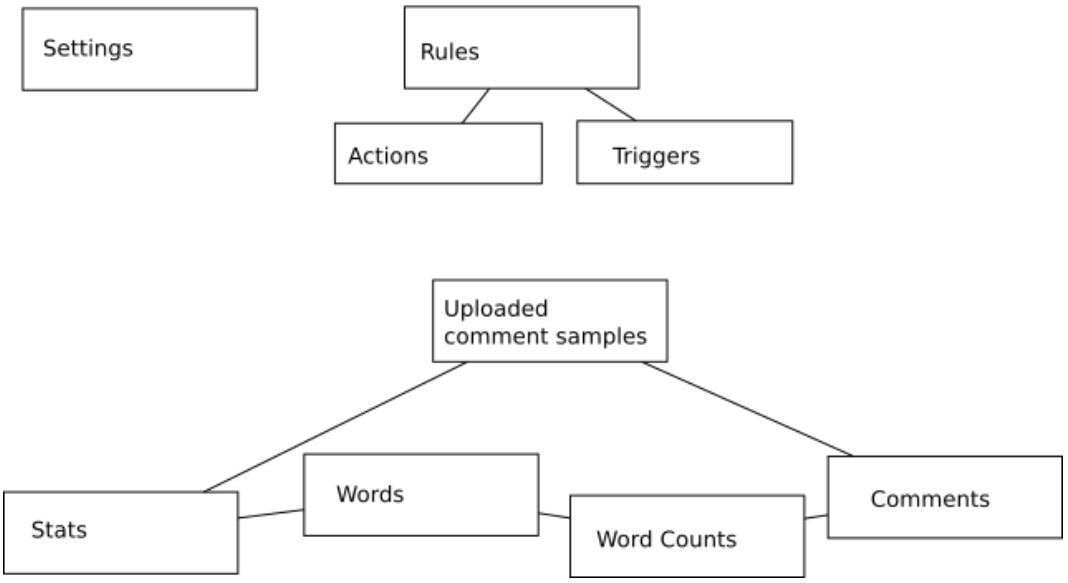
The comment samples will be stored pre-processed, with the word counts table storing a relationship between individual words, the number of occurrences in a comment, and the full comment. Comments will have a foreign key into the sample they came from. This will allow the construction of a posterior probability array at runtime as shown below:

word	*comment1*	*comment2*	*comment3*	*totals*
word1	*2*	*0*	*3*	*5*
word2	*3*	*2*	*7*	*12*
word3	*5*	*1*	*2*	*8*
totals	*10*	*3*	*12*	*25*

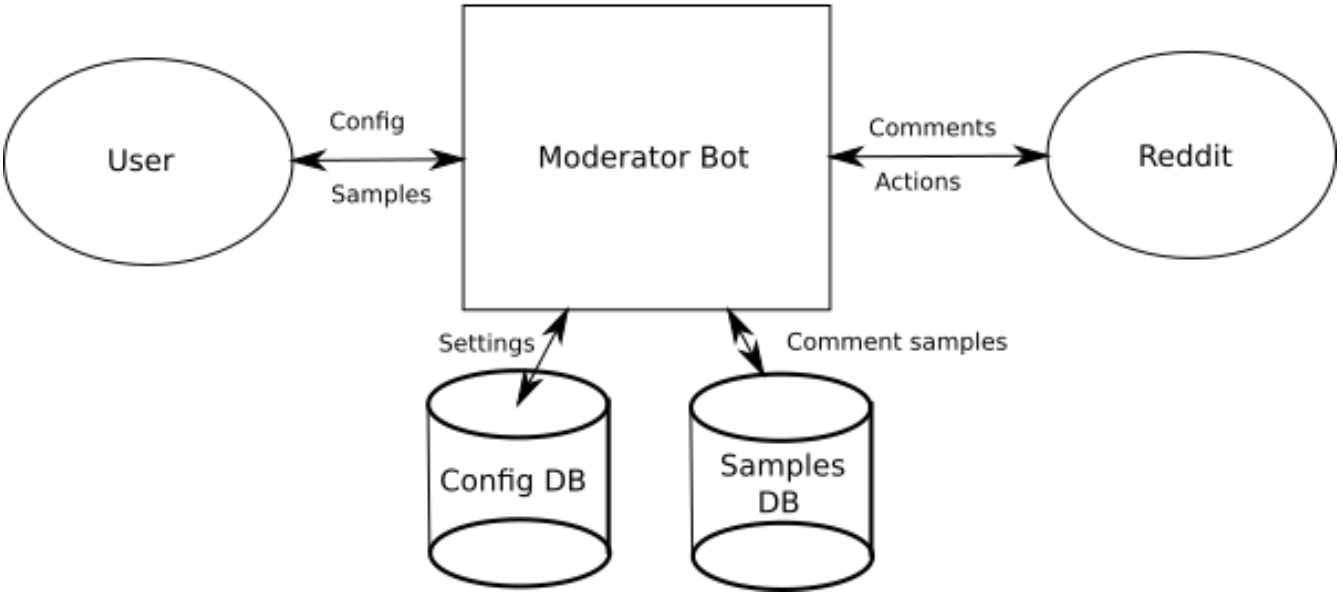
IE. the probability of word1 in this set of comments is 5/25.

The stats table will hold the mean, standard deviation, median and median difference from median of each word in each sample, which may be useful for analysing word frequencies or for some types of anomaly detection. [7]

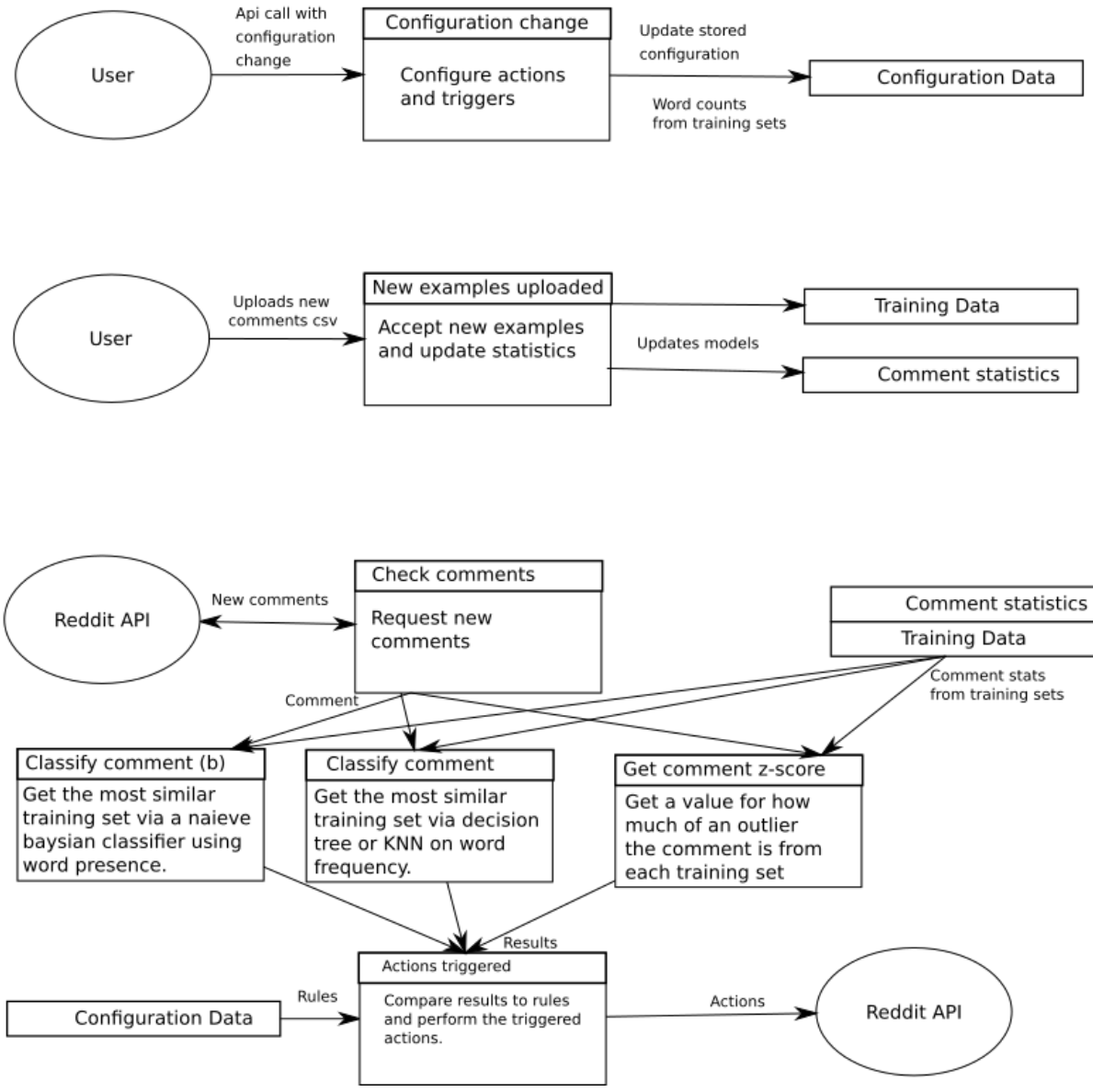
Data Model



Context Diagram

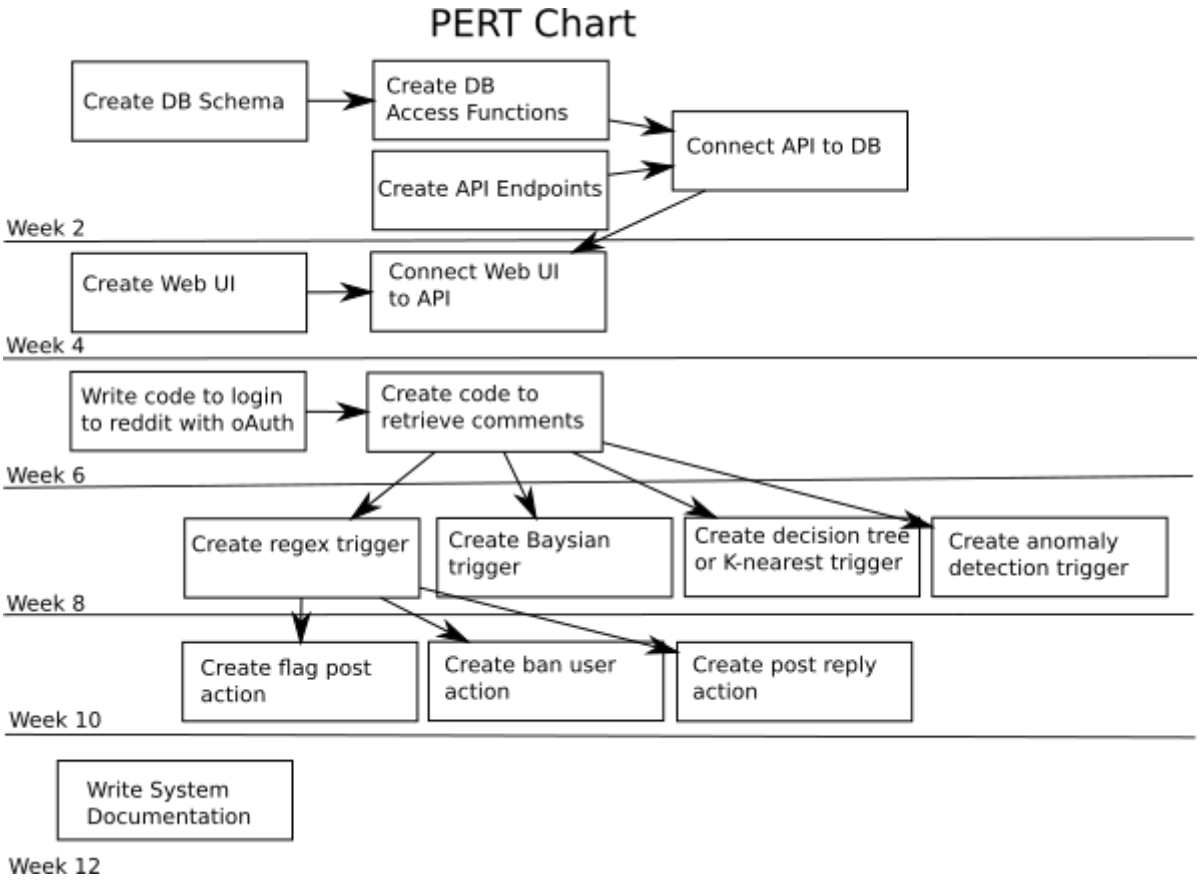


DFD of core functions



6. Preliminary Schedule

The project must be fully completed, tested and documented within a 12 week semester. A preliminary PERT chart showing the order tasks should be completed in and the estimated time of completion is below.



7. Appendices

Links:

1. http://i.imgur.com/dPozEQc.png
2. http://mikedoesweb.com/sandbox/reddit/?r=dogs&n=500
3. http://mikedoesweb.com/sandbox/reddit/?r=sneks&n=500
4. https://www.reddit.com/dev/api/
5. https://www.reddit.com/wiki/automoderator/full-documentation
6. https://github.com/reddit/reddit/blob/master/r2/r2/lib/automoderator.py
7. http://nlp.shef.ac.uk/Completed_PhD_Projects/guthrie.pdf