Università
della
Svizzera
italiana

Faculty
of Informatics

Bachelor Project

November 29, 2020

Experimental Apparatus for a Digital Health Literacy Experiment

(DRAFT)

Robert Jans

Abstract

In this project we implement an software toolkit needed for an experiment planned by the *Institute of Communication and Health (ICH)*. The institute is part of the *Faculty of Communication Sciences* at the *Università della Sizzera Italiana*. Health communication is a relatively new and multidisciplinary field, which includes the study and use of communication to inform and affect decision making at the individual and community level in order to improve the quality of healthcare [1]. The purpose of the planned experiment is to investigate the digital health literacy of participants in the area of sleeping disorders.

This project seeks to provide the experimental apparatus which will allow the research team to record data an test its hypoteses. The project's main tasks are indexing a predefined set of websites, creating a user interface similar to common search engines that can be configured to selectively show different subsets of the corpus according to the experimental conditions under inestigation and setting up an experimental environment allowing to conduct controlled experimments (e.g. prevent participants from accessing non-corpus sites) and to record salient data such as search logs and click-stream.

Advisor
Prof. Marc Langheinrich
Co-Advisor
Prof. Peter Schulz

Advisor's approval (Prof. Marc Langheinrich): Date:

1 Introduction

1.1 Motivation

Since the beginning of modern science researchers needed special tools in order to conduct experiments. These tools consist mainly in devices for taking measurements and triggering phenomena under controlled conditions. Whereas in the past the experimental apparatuses comprised almost exclusively physical devices, nowadays increasingly more software is involved. The experiment for which the result of my project is going to be used is a case in which the process of setting up the experimental conditions an collecting the result data depends heavily on the software toolkit. It is therefore essential for the software to be robust and reliable.

My personal motivations Come from two sides: I have a general interest in science an the scientific method, but in practice, rather than as a scientist, I see myself as a technician, who in the area of of software development seeks to build useful applications. Int that sense this project perfectly fits my interests, as it involves the creation of software to be used in the context of scientific research.

1.2 Outline

TODO:

2 Requirements

Below is a description of the requirements as defined by the advisor. The requirements include three main tasks as well as eight milestones; the milestones are divided into the categories *Must have*, *should have*, and *Nice to have*.

2.1 Main Tasks

1. Spidering (i.e. creating a full or partial local copy of) a predefined set of websites that provide sleeping disorder information as an experimental corpus.
2. Creating a Google-like search interface to the corpus that can be configured to selectively show/rank different sets of corpus sites, according to the experimental conditions under investigation.
3. Setting up an experimental environment (e.g. using the *SafeExamBrowser*, or using a proxy server) to conduct controlled experiments using the corpus (e.g. to prevent participants from accessing non-corpus sites) and to record salient data (e.g. search logs, click-stream).

2.2 Milestones

1. **(Must have)** A website which simulates a search engine that lets users enter keywords into a search form and returns results (snippets) from a predefined corpus of websites/links, which can then be clicked on / followed.
2. **(Must have)** A result generator and a simple way to configure it (e.g. using a text file) on a per-group basis (i.e. participants in Group 1 receive result from lists R1, R2, and R3; Group 2 participants receive results from R4, R2 and R3).
3. **(Must have)** A report describing the system setup/installation and the architectural design.
4. **(Should have)** A result generator that can detect identical, repeated queries (or minor variations of otherwise identical queries, detectable via stop word removal and stemming) upon which it will generate the same response.
5. **(Should have)** A detailed log engine which allows the experimenter to track key experimental results for each participant, such as search terms entered, the time spent on a given result list and any clicks on results (when, which order).
6. **(Should have)** A visual presentation of the search entry and result section that mimics a known search engine.
7. **(Nice to have)** A result generator that can handle non-related searches using a pass-through to a real search engine.
8. **(Nice to have)** A web interface to the log engine that allows for convenient inspection, analysis, and export of experimental results.

3 Project design

3.1 General structure

Given the close relationships between the experiment configuration, conduction, and analysis, I decided to include most of the implied functionalities into a single web application, called *HSE (Health Search Engine)*. The user management functionality of HSE allows participants to access only a search interface, while experimenters have access to pages for managing corpora and experiments. If needed, participants can be prevented from accessing non-corpus sites by restricting the browser to a whitelist based on the corpus used for a given experiment. The following subsections briefly describe the usage modalities and the related user interfaces.

3.2 Typical usage workflow

The typical workflow includes four steps: preparing the document corpora, setting up an experiment, running the experiment and finally evaluating the resulting data. **Figure 1** shows this usage scenario. To each usage step corresponds a dedicated user interface.

For preparing the corpora the experimenter provides text files containing the web URL's of the chosen documents. After setting the names for the corresponding document collection, the application takes care of downloading the contents and creating the inverted indices needed for retrieval. Setting up an experiment involves defining test groups and assigning participants to each group. Moreover for each test group it is needed to set the document collections from which the retrieval mechanism will select the results to be displayed to the participants. The groups cab be defined either manually or by uploading a configuration file. The U.I. for experiment execution includes a control button for starting, stopping, or resetting an experiment, as well as a tabular display for real time monitoring, showing the current number of queries and clicks for each participant. Starting an experiment enables the participants tos log in, and initiates the data collection mechanism; after stopping the experiment, the participants are logged out, and transient data is saved. The experiment evaluation interface allows for quick inspection through data summaries and visualizations. Moreover it allows to export both the raw data and the summaries as files.

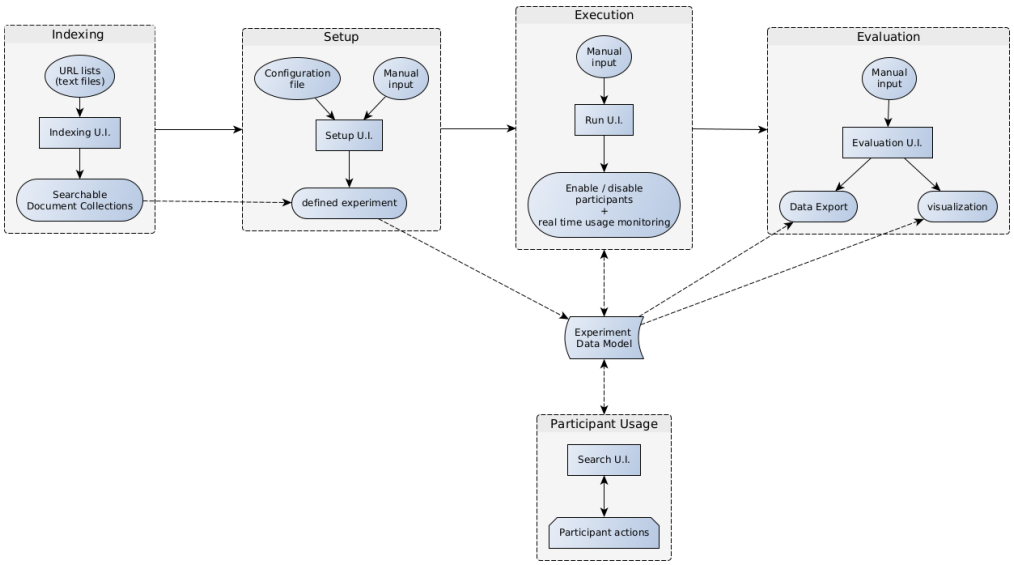
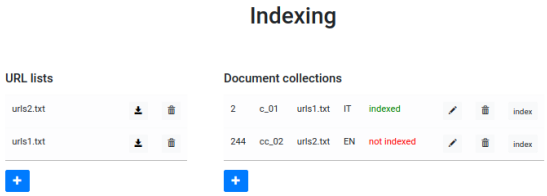


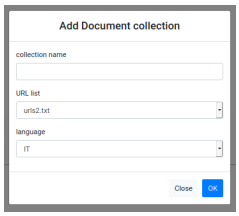
Figure 1. Typical usage workflow

3.3 Defining the document corpora

In order to define the document collections (corpora) to be used during subsequent experiments, an experimenter can upload text files containing lists of web URL's. The files are stored, so they can be reused for defining multiple document collections. Via a popup menu a new document collection can be defined by providing a name, the collection's language the related url list. Clicking on the "index" button initiates the indexing process, which includes data download and the creation of a index data structure for retrieval. The details of the indexing process are explained in section xxx **TODO: link actual section**. **Figures 2a and 2b** show the relevant parts of the interface.



(a) indexing u.i



(b) popup for defining document collections

3.4 Setting up an experiment

The U.I. for experiment setup allows to define the details of an experiment to be carried out. This step involves creating test groups with associated participants and document collections. Groups can be defined either manually or by using an uploaded configuration file. In both cases the test group configuration can later be edited manually. The interface allows to link each group to a set of previously indexed document collections. **Figure 3** shows this U.I.

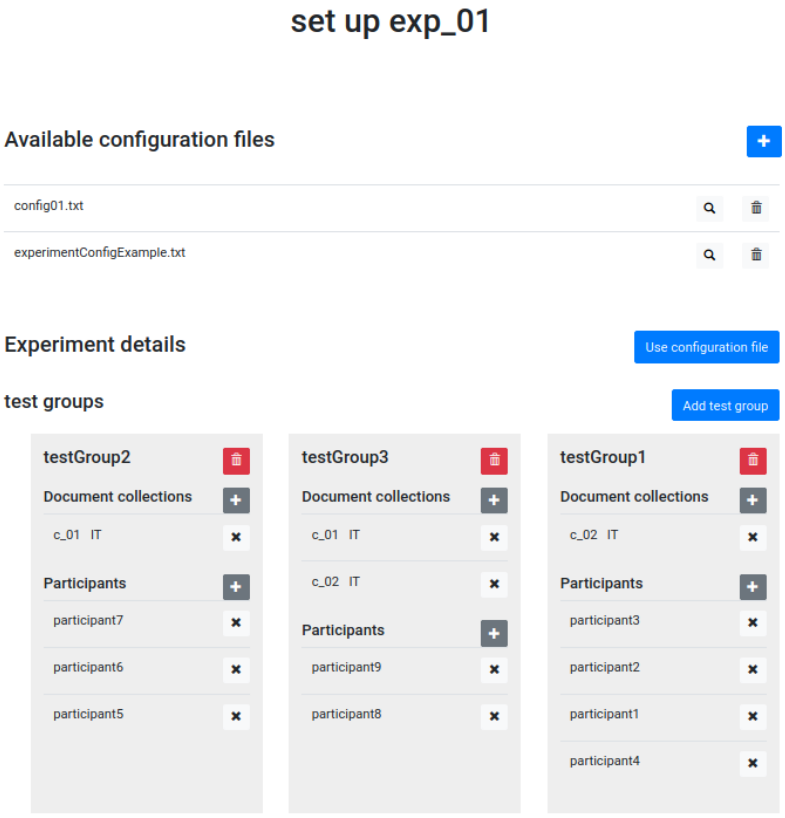


Figure 3. U.I. for experiment setup

3.5 Running an experiment

The U.I. for experiment execution inccludes a start/stop/reset button, a timer, and a tabular display showing the current participant activities. Clicking the start button starts the timer, enables the participants to log in and initiates the data collection process. While the experiment is running, all queries and click carried out by the participants are stored as database records including a timestamp, user id, group id, and query/document related data. The details of the data collection process are described in section xxx **TODO: link actual section**. When the stop button is clicked the participants are logged out and all transient data is saved to the database. after the experiment is complete the related evaluation page becomes available. In case something goes wrong, the experiment can be reset. This causes all query and click data to be deleted, while the experiments configuration is preserved. **Figure 4** shows this U.I.

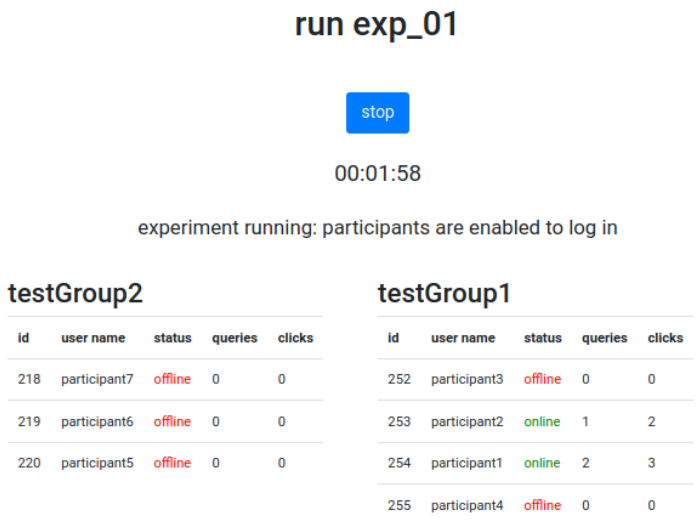


Figure 4. U.I. for experiment execution

3.6 Search user interface available to participants

The interface available to the participants looks similar to the main page of most known search engines. It simply includes a search text bar and a button for entering queries, and displays the results as a list of links accompanied by short summaries (snippets) with highlighted query terms. **Figure 5** shows the U.I. after a query has been entered.

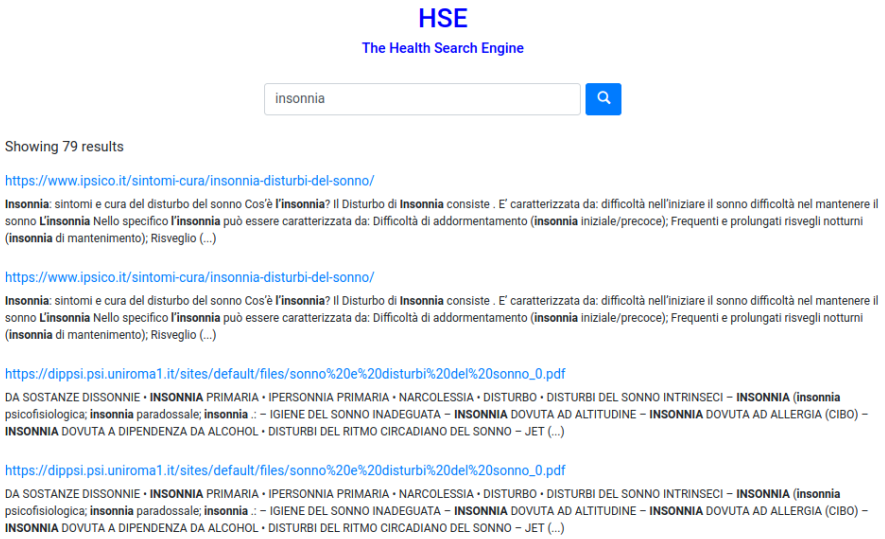


Figure 5. User interface available to participants

3.7 Evaluating an experiment

After an experiment has been conducted, the related evaluation interface becomes available. From on page experimenters can inspect the experiment's results and export the complete raw data or pre-processed data summaries. The U.I. includes visualizations of the most relevant data features.

The raw data can be exported either in csv or json format, and consists i a list of all user actions occurred during the experiment. Each record has a timestamp, a user id and a group id. Query event records include the query text and the proportions in which the data collections are represented in the result list. Document click event records include the document id, its URL, and the document collection to which it belongs.

The data summaries include overall experiment statistics and per-group statistics. Moreover the individual user histories can be exported in the same format as the raw data.

Per-experiment statistics include the total count of clicks and queries as well as averages, medians, and standard deviations for queries per user, clicks per user, clicks per query, time per query and time per click.

Per group statistics include the same metrics as the per-experiment statistics, plus totals, averages, medians and standard deviations for clicks per document collection.

TODO: describe in more detail

4 Architecture and employed frameworks

4.1 Architecture

4.2 Lucene: an open source information retrieval library

TODO: library description

4.3 SpringBoot: a Java framework for web applications

TODO: framework description

5 Implementation details

5.1 Indexing

5.2 Retrieval

5.3 Usage Tracking

5.4 Data pre-processing

6 Under the hood: The information retrieval process

TODO: describe indexing, scoring formula, vector space retrieval model, etc...

7 Quality assessment and evaluation

8 Conclusions

9 Future work

References

[1] I. of Communication and Health. *ICH web site*. https://www.ich.com.usi.ch/en/about-us/institute.