Social Information Retrieval Experiment & Exercise

Fact Pack for Homework Assignment #3

Christoph Fuchs

2015-05-20

Source: https://vmschlichter24.informatik.tu-muenchen.de/ha3/

Homework Assignment #3

Due: **10.06. 23.59 CEST**

Covered areas in this document

Please take a look at this week's fact pack for further information on how to use the tools.

1. **Build and index your own individual information space** - this might take a while until it finishes
   - Use `generate_whitelist` and `history_crawler` from the software kit to set up your private information space (consisting of the websites you visited)
   - Once `history_crawler` has finished, use the `indexer` tool to calculate a LDA topic model on your information space
2. **Build product database**
   - Use `AmazonViewedProductsToFile`, `AmazonParser.js` and `AmazonBoughtProductsToFile` to generate files
3. **Visualize the topics** in your information space using `cloudcreator`.
   1. Look at the generated HTML files, evaluate whether the topics represent areas of expertise or interest. Save the output of the last HTML file in a text file.
   2. Summarize whether - from your point of view - topic models appear to be a suitable tool to detect topics in information spaces (~50-100 words, PDF).
4. **Start the manual query process**
5. Uploads:
   1. Upload the calculated topic model (task #1)
   2. Upload the product database and your analysis on the topics (tasks #2, 3)

Build and index your own individual information space

1) Generate Whitelist (1/4)

- **Close your web browser** completely (all windows!)
- **Copy your browser's history file** to the directory where you extracted the software kit (see next slides for possible paths of history files for different browsers)
- Use "`generate_whitelist`" to generate a whitelist (whitelist.txt) which lists all the domains in your browser history (which will form your private information space)

```
./generate_whitelist --history PATH_TO_BROWSER_HISTORY
```

PATH_TO_BROWSER_HISTORY (example, full list on next slides):

Chrome, Mac OS X:

/Users/**christoph**/Library/Application\ Support/Google/Chrome/Default/History

Firefox, Mac OS X:

/Users/**christoph**/Library/Application\ Support/Firefox/Profiles/**tibhr9q0.default**/places.sqlite

1) Generate Whitelist – Chrome (2/4)

Windows XP

Google Chrome: C:\Documents and Settings\%USERNAME%\Local Settings\Application Data\Google\Chrome\User Data\Default\History
Chromium: C:\Documents and Settings\%USERNAME%\Local Settings\Application Data\Chromium\User Data\Default\History

Windows 8 or 7 or Vista

Google Chrome: C:\Users\%USERNAME%\AppData\Local\Google\Chrome\User Data\Default\History
Chromium: C:\Users\%USERNAME%\AppData\Local\Chromium\User Data\Default\History

Mac OS X

Google Chrome: ~/Library/Application Support/Google/Chrome/Default/History
Chromium: ~/Library/Application Support/Chromium/Default/History

Linux

Google Chrome: ~/.config/google-chrome/Default/History
Chromium: ~/.config/chromium/Default/History

1) Generate Whitelist – Firefox (3/4)

Windows XP

C:\Documents and Settings\<**Windows login/user name**>\Application Data\Mozilla\Firefox\Profiles\<**profile folder**>\places.sqlite

Windows Vista and later

C:\Users\<**Windows login/user name**>\AppData\Roaming\Mozilla\Firefox\Profiles\<**profile folder**>\places.sqlite

Mac OS X

~/Library/Application Support/Firefox/Profiles/<**profile folder**>/places.sqlite
~/Library/Mozilla/Firefox/Profiles/<**profile folder**>/places.sqlite

Linux

~/.mozilla/firefox/<**profile folder**>/places.sqlite

1) Generate Whitelist (4/4)

- Life is easier when you copy the history / places.sqlite file to a suitable location
- It is not possible to access the original history / places.sqlite file while the respective browser is still open…
- **Open whitelist.txt and remove all the URLs you don‘t want to include in the crawling process to build your private information space**

2) Download the visited websites

- Start the crawling process with the following command:

```
./history_crawler --history PATH_TO_BROWSER_HISTORY
```

- All your visited websites from domains listed in `whitelist.txt` (required to be in the same directory) are downloaded and stored inside the `data/` directory
- Depending on your internet connection speed and size of browsing history, this might take some time…

3) Calculate Topic Models

- Identify topics of the downloaded websites by running

```
./indexer --data_dir ./data/
```

If the command above does not work, please try
`./indexer/indexer --data_dir ./data/`

- This tool uses LDA (Blei, 2012) to identify topics in your browsing history
- Output data is stored in the current directory (theta.mm, topic_model_lda.zip, dict.dict)
- It might take some time to calculate the topics

Blei, David M. (2012): Probabilistic topic models, Communications of the ACM Vol. 55(4) p77-84

4) Upload output files to web interface

- Go to https://vmschlichter24.informatik.tu-muenchen.de/lda_upload/ and upload output files

Remarks on Privacy:

- The uploaded data does not contain the visited URLs of your history but only derived information (frequencies of words)
- The data is only used to calculate a matching between a query from one of your friends (“data request”) and your data
- In the end, you will be the one who authorizes a response to the data request in a case-by-case basis; no participant will be able to see parts of your dataset without your explicit authorization
- You will have the chance to reply anonymously to a data request

Build product database

1) Generate List of Viewed Items (products_viewed.json) (1/2)

- Close your web browser
- Copy the browser's history file to the folder of the extracted software kit (if not already done so)
- Use "`AmazonViewedProductsToFile`" to generate a JSON file with the products you looked at on Amazon

```
./AmazonViewedProductsToFile --history NAME_OF_BROWSER_HISTORY_FILE
```

```
NAME_OF_BROWSER_HISTORY_FILE:
Chrome: History
Firefox: places.sqlite
```

Full list of path to history files for Firefox / Chrome on different systems -> see slides 5+6

1) Generate List of Viewed Items (products_viewed.json) (2/2)

- Edit products_viewed.json with a text editor to remove all those things you “looked up for friends”… ☺ - but don‘t destroy the beautiful JSON schema (i.e. only remove complete chunks including opening and closing brackets and commas)

products_viewed.json — User A


```
        "visit_count": 1,
        "name": "POLAR Herzfrequenz Sensor H7",
        "last_visit": "2015-04-23 18:05:30"
    },
    {
        "category": "Sport & Freizeit",
        "URL": "http://www.amazon.de/Runtastic-Brustgurt-Heart-Monitor-RUNBT1/dp/B00B84JQSE/ref=sr_1_2?ie=UTF8&qid=1
        "visit_count": 1,
        "name": "Runtastic Brustgurt Heart Rate Combo Monitor, RUNBT1",
        "last_visit": "2015-04-23 18:05:32"
    },
    {
        "category": "Schuhe & Handtaschen",
        "URL": "http://www.amazon.de/Brooks-Herren-Laufschuhe-Silber-Schwarz/dp/B00H0DA9VA/ref=sr_1_5?ie=UTF8&qid=14
        "visit_count": 1,
        "name": "Brooks Ghost Herren Laufschuhe",
        "last_visit": "2015-04-23 18:05:57"
    },
    {
        "category": "Schuhe & Handtaschen",
        "URL": "http://www.amazon.de/Brooks-Ghost-GTX-Herren-Laufschuhe/dp/B00R43QUVE/ref=pd_sbs_a_2?ie=UTF8&refRID=
        "visit_count": 1,
        "name": "Brooks Ghost 7 GTX Men Herren Laufschuhe",
        "last_visit": "2015-04-23 18:07:04"
    }
]
```

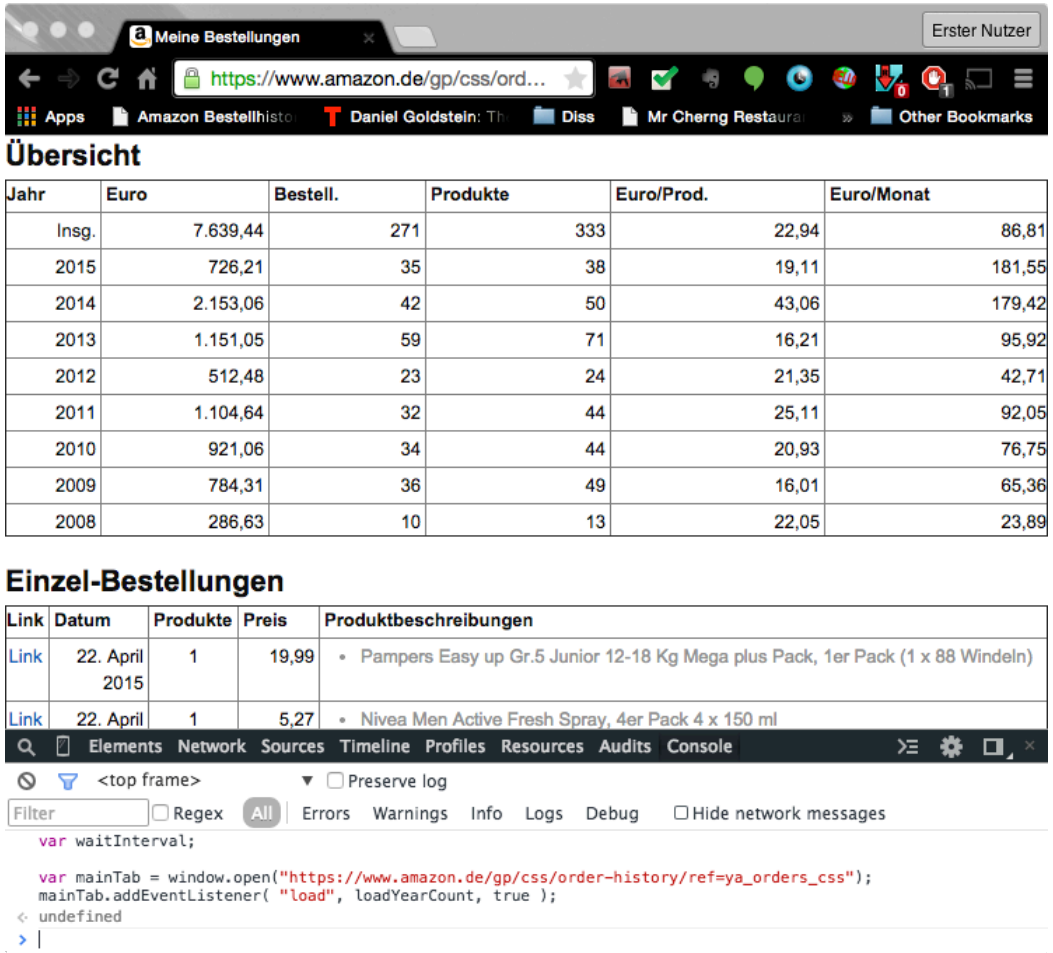

Line: 1:2 | JSON | Soft Tabs: 4

2) Generate List of Bought Items (products_bought.txt) (1/3)

- **Easy way I:** Export items directly from Amazon using Developer Mode of your browser (Java Script Console), a special JavaScript code snippet and a Python script to export the data (you will also get nice statistics about the amount of money you already spent at Amazon...)

- **Log in to Amazon**, navigate to order history, open JavaScript console in developer tools (e.g. by choosing „Inspect element“ in Chrome‘s context menu or using Firebug in Firefox)
- **Paste content** of `AmazonParser.js` (part of software kit) in console window and press ENTER
- A lot of tabs will open automatically, **your browser will be really busy** for a couple of minutes...

2) Generate List of Bought Items (products_bought.txt) (2/3)



- After some time, a nice result page is shown in the browser...
- Before you start to reconsider your spending behavior, save the page (file, save as, ...)
- Run „`AmazonOrderAnalyzer`“ to extract the bought items to a file called „`products_bought.txt`“

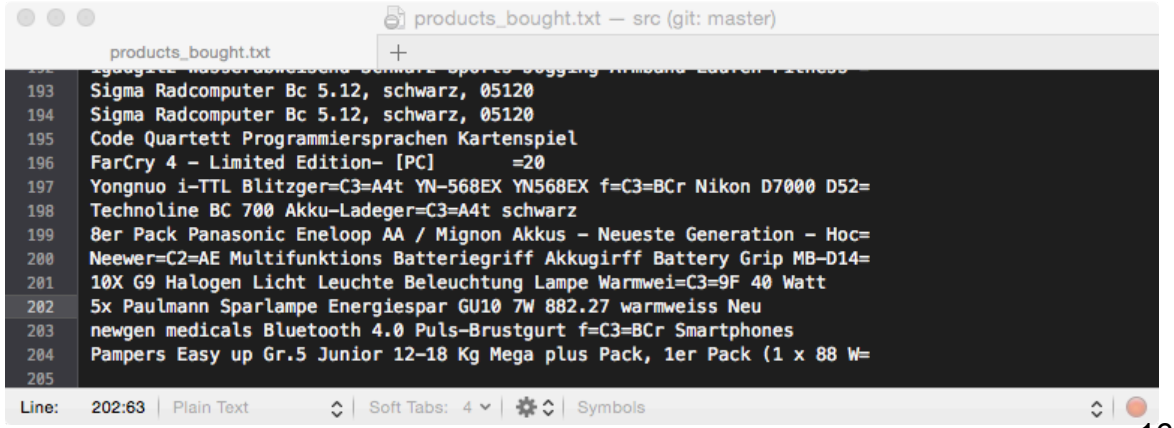
```
./AmazonOrderAnalyzer --htmlfile PATH_TO_SAVED_HTML_FILE
```

ALTERNATIVE

2) Generate List of Bought Items (products_bought.txt) (3/3)

- **More complicated way:** Create products_bought.txt file manually
  - Fire up your favorite text editor
  - Open your mail program & search for Amazon shipping confirmation mails
  - Copy the names of the shipped items to the text file, use a new line for each item

- Even if you use the easy way, it might make sense to review the products_bought.txt file before uploading it (feel free to delete/correct items)



Visualize the topics

1) Run cloudcreator (1/2)

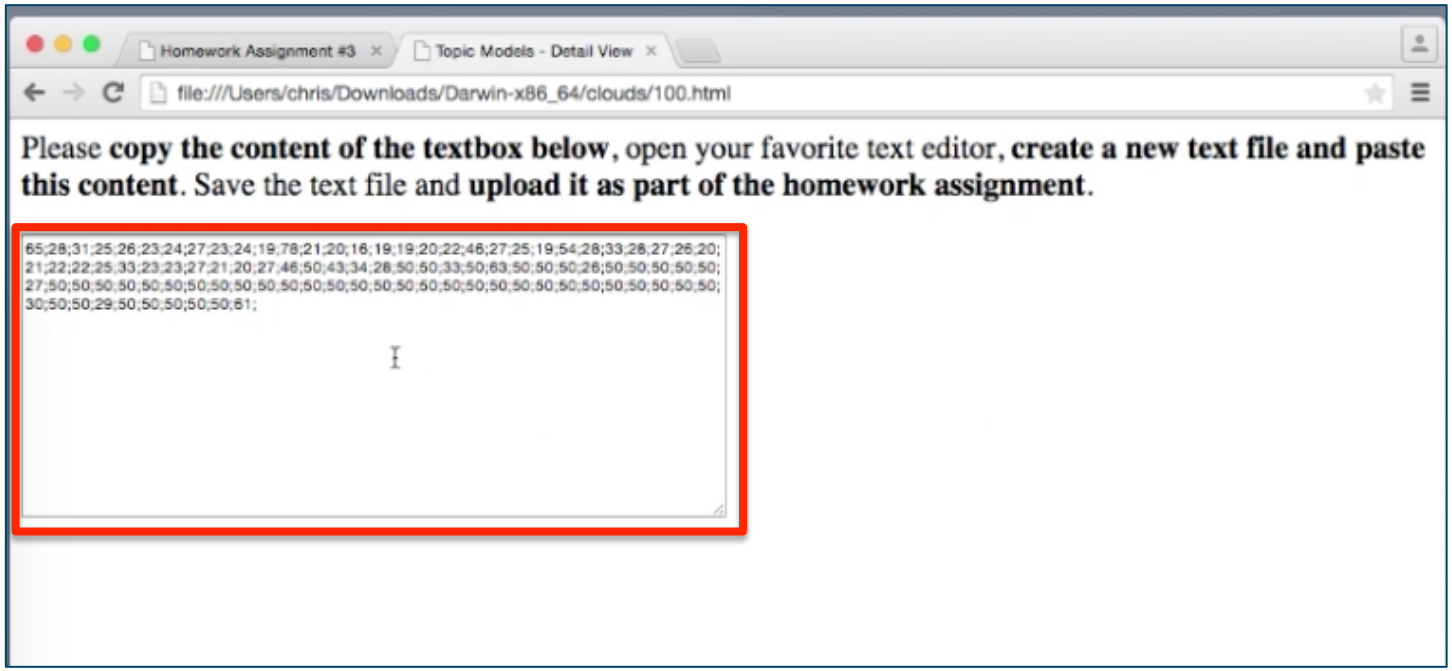
- Run **cloudcreator** from the directory where you extracted the software kit to - this assumes that the files for the topic model (e.g., lda_model.model) are in the same directory, otherwise please adjust accordingly):

```
./cloudcreator --topic_model_dir . --output_dir clouds
```

- (This will take a few minutes)
- After it finished, please open `clouds/0.html` in your browser
- Indicate the quality of the topic using the vertical slider on the right side of the window
- Click on “Next Topic”

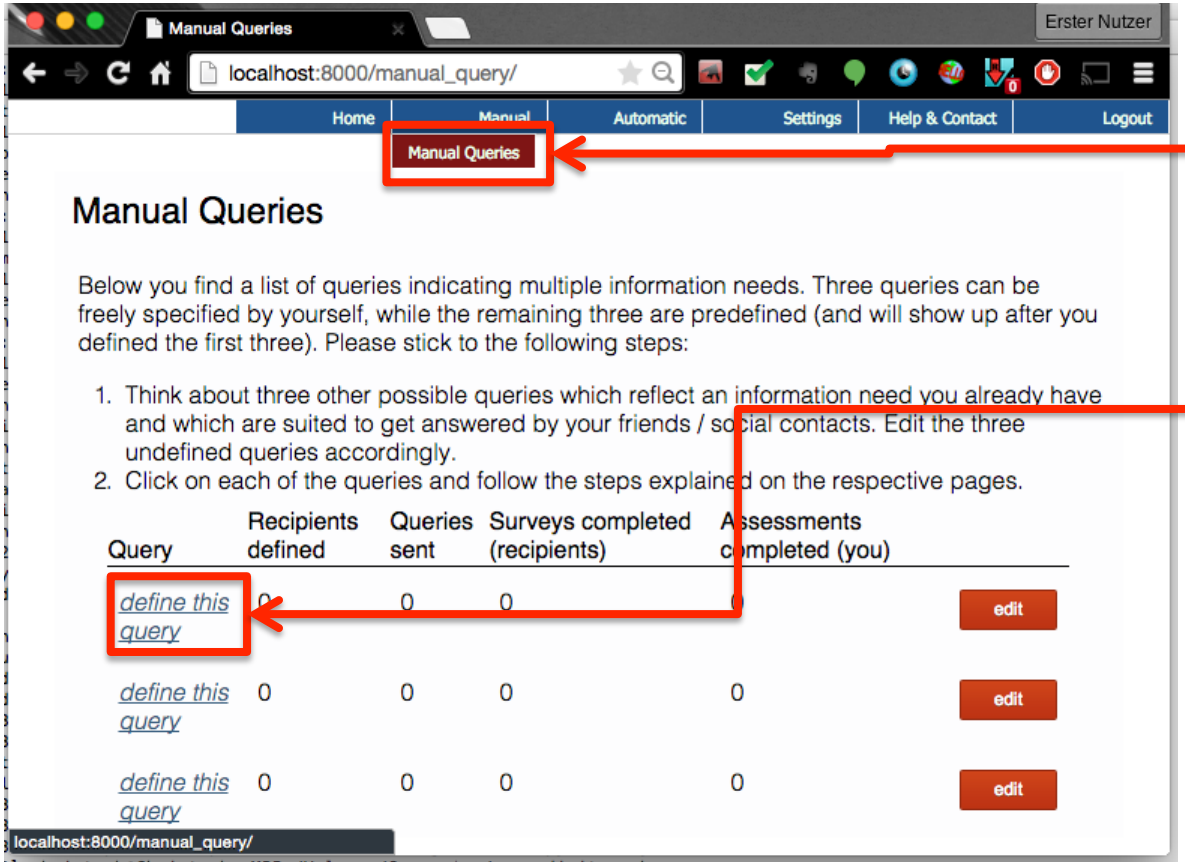
1) Run cloudcreator (2/2)

- Once you are done with all topics, please copy the content of the text box on the final HTML file to a newly created text file and save it



Start the manual query process

1) Log into Web System



1) Select „Manual Queries“ from the menu – you see an overview of active queries

2) Click on one of the queries to define it (if not done already)

2) Define Query in Manual Mode

Define and save query

Subscription

Erster Nutzer

localhost:8000/manual_query_edit/?q...

Please choose a **query** which **reflects an information need you already have** and which is suited (in your opinion) to get answered by your friends / social contacts.

You will **manually send the query to selected recipients** within your social network, so it is **possible to phrase it as a question**.

Save query

Erster Nutzer

utomatic | Settings | Help & Contact | Logout

mation needs. Three queries can be
are predefined (and will show up after you
ps:

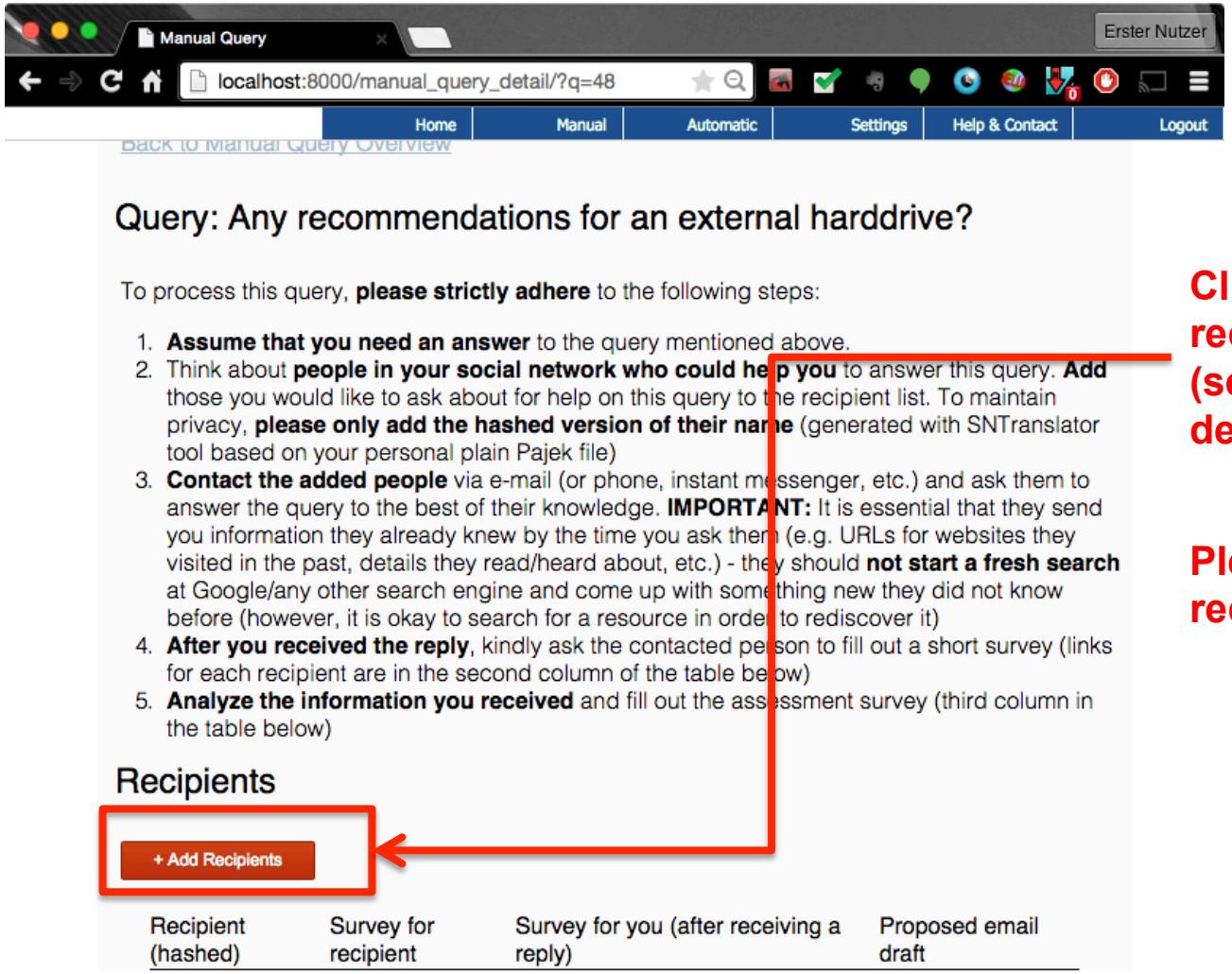
lect an information need you already have
and which are suited to get answered by your friends / social contacts. Edit the three
undefined queries accordingly.

2. Click on each of the queries and follow the steps explained on the respective pages.

Query	Recipients defined	Queries sent	Surveys completed (recipients)	Assessments completed (you)	
define this query	0	0	0	0	edit
define this query	0	0	0	0	edit
Any recommendations for an external harddrive?	0	0	0	0	edit

Newly added query

3) Add Recipients for Query

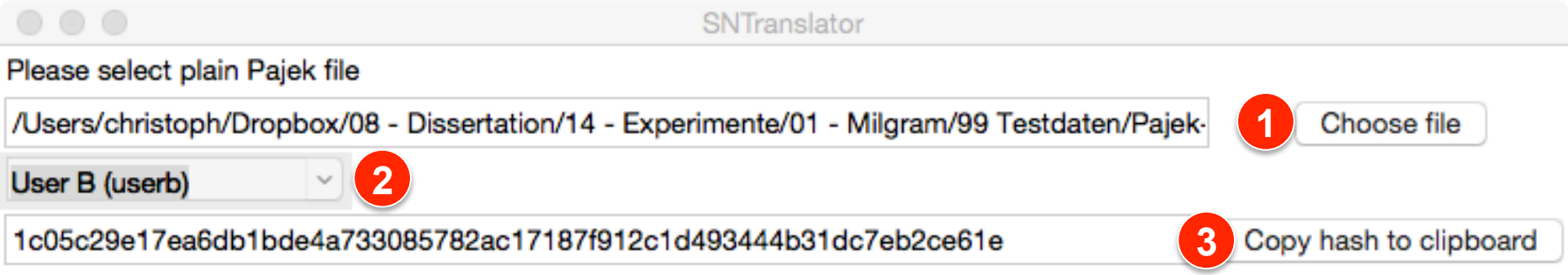


Click here to add new recipients for the query (see next chapter for details)

Please add at least one recipient to each query

4) Calculate Hash ID for recipient

- Run `SNTranslator` from the software kit



1. Choose plain pajek file (stored before) – (default filename: my_plain_network.net)
2. Select recipient for manual query
3. Copy hash to clipboard (and insert it in form field on web application when adding a new recipient, see next slide)

5) Paste ID / hash here

Query: Which movie would you recommend to watch with a close friend?

Provide the **hash of the person you would like to add as a recipient** to the query *"Which movie would you recommend to watch with a close friend?"*. You can calculate the hash code based on your plain Pajek file and the SNTranslator tool. (FIXME: Add link). Please also provide a **short justification** why this person seems to be a good choice to answer this query.

Please be aware that **you have to contact this person later**.

Name (hashed):

Why do you think that this recipient is a good choice?

Add recipient

Paste ID / hash here

6) Surveys

Home | Manual | Automatic | Settings | Help & Contact | Log

Query: Any recommendations for an external harddrive?

To process this query, **please strictly adhere** to the following steps:

1. **Assume that you need an answer** to the query mentioned above.
2. Think about **people in your social network who could help you** to answer this query. **Add** those you would like to ask about for help on this query to the recipient list. To maintain privacy, **please only add the hashed version of their name** (generated with SNTranslator tool based on your personal plain Pajek file)
3. **Contact the added people** via e-mail (or phone, instant messenger, etc.) and ask them to answer the query to the best of their knowledge. **IMPORTANT:** It is essential that they send you information they al Ls for websi es they visited in the past, deta **not start a fresh search** at Google/any other se they did no know before (however, it is o over it)
4. **After you received th** out a short survey (links for each recipient are i
5. **Analyze the informati** urvey (third column in the table below)

1) Click here for an email draft to send to the recipient containing all required links and information for the recipient

2) Click on this link to fill out a survey once you received and reviewed a reply from this user

Recipients

+ Add Recipients

Recipient (hashed)	Survey for recipient	Survey for you (after receiving a reply)	Proposed email draft	
bbc4200d9...	http://localhost:8000/manual_query_reply/UFwAo9JHN0/	fill out survey	open e-mail draft	edit

PREPARATION – NOTES & REMARKS (1/2)

We take Data Privacy seriously: We don't do anything without your permission & we are transparent on what we do:

1. **We don't know the names of your Facebook friends**: We only ask you to upload the hashed version of the network file[1]
2. **None of your data** (evaluation of ties, names, products, etc.) **will ever be related to you**
3. **We don't get to know your browser history** – the uploaded topic models only contain document IDs and word vectors

1 The social network built from all individual ego-networks will get hashed again and randomly changed (edges and nodes changed/removed) before used in exercise to ensure that it is not possible to identify nodes

PREPARATION – NOTES & REMARKS (2/2)

4. The data will only be used to **conduct sound scientific research** (and student exercises)
5. You will use a (**obfuscated**) version of the dataset to **solve homework assignments**
6. Your **name, matriculation number, email address**, etc. **is only used to grade the homework** – it **<u>will be deleted afterwards</u>**
7. We will **never give** the data to anyone
8. Publications based on this dataset will only contain **highly aggregated information**
9. We will **hash the Amazon products after the experiment** and delete all copies of the original dataset