

# **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

**Covered areas in this document**

Due: **10.06. 23.59 CEST**

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

#### 4) Upload output files to web interface

- Go to [https://vmschlichter24.informatik.tu-muenchen.de/lda\\_upload/](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

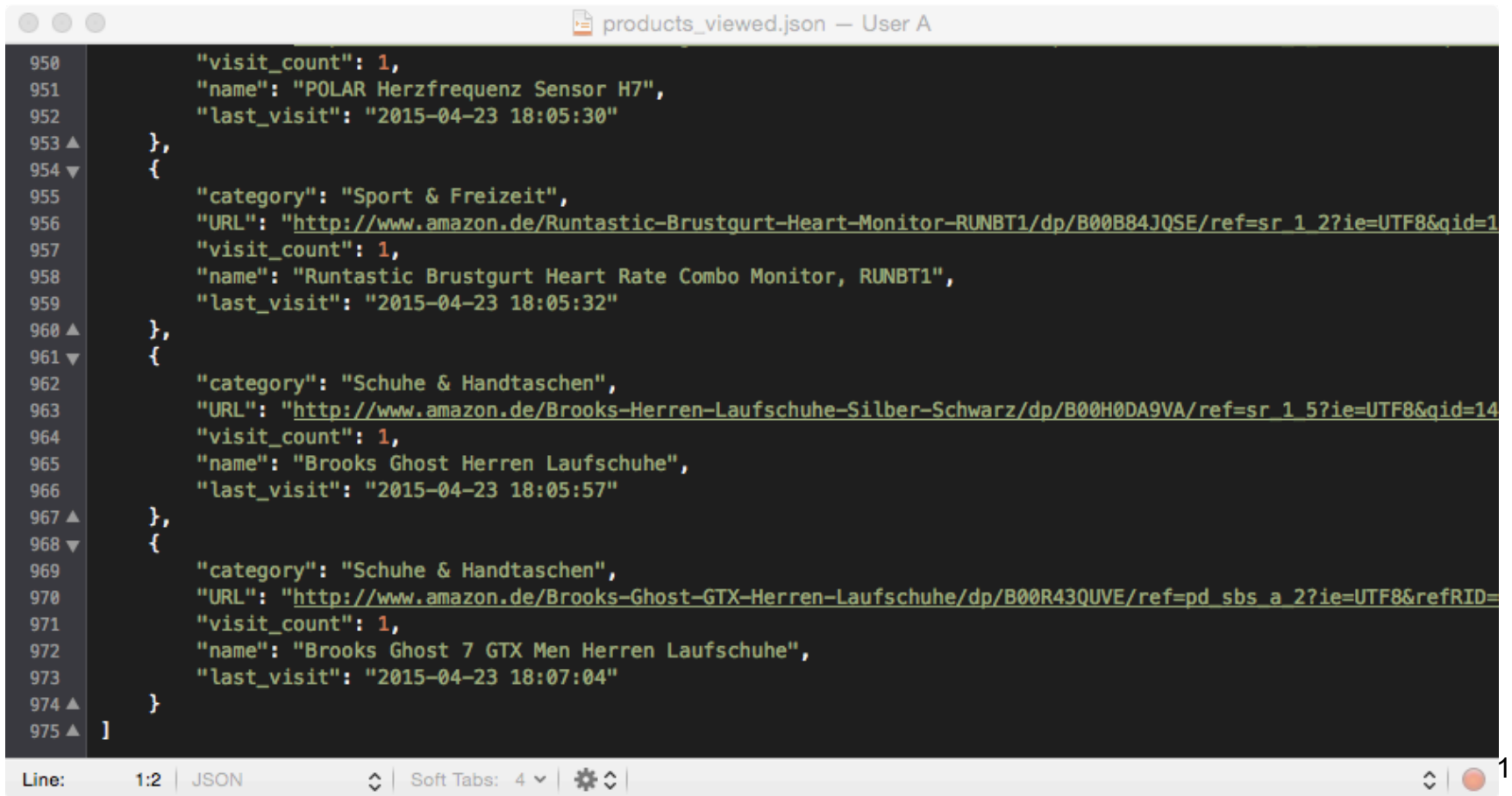
```
./AmazonViewedProductsToFile --history NAME_OF_BROWSER_HISTORY_FILE
```

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
950     "visit_count": 1,
951     "name": "POLAR Herzfrequenz Sensor H7",
952     "last_visit": "2015-04-23 18:05:30"
953 },
954 {
955     "category": "Sport & Freizeit",
956     "URL": "http://www.amazon.de/Runtastic-Brustgurt-Heart-Monitor-RUNBT1/dp/B00B84JQSE/ref=sr_1_2?ie=UTF8&qid=1444444444",
957     "visit_count": 1,
958     "name": "Runtastic Brustgurt Heart Rate Combo Monitor, RUNBT1",
959     "last_visit": "2015-04-23 18:05:32"
960 },
961 {
962     "category": "Schuhe & Handtaschen",
963     "URL": "http://www.amazon.de/Brooks-Herren-Laufschuhe-Silber-Schwarz/dp/B00H0DA9VA/ref=sr_1_5?ie=UTF8&qid=1444444444",
964     "visit_count": 1,
965     "name": "Brooks Ghost Herren Laufschuhe",
966     "last_visit": "2015-04-23 18:05:57"
967 },
968 {
969     "category": "Schuhe & Handtaschen",
970     "URL": "http://www.amazon.de/Brooks-Ghost-GTX-Herren-Laufschuhe/dp/B00R43QUVE/ref=pd_sbs_a_2?ie=UTF8&refRID=XXXXXXXXXX",
971     "visit_count": 1,
972     "name": "Brooks Ghost 7 GTX Men Herren Laufschuhe",
973     "last_visit": "2015-04-23 18:07:04"
974 }
975 ]
```

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

## 2) Generate List of Bought Items (products\_bought.txt) (1/3)

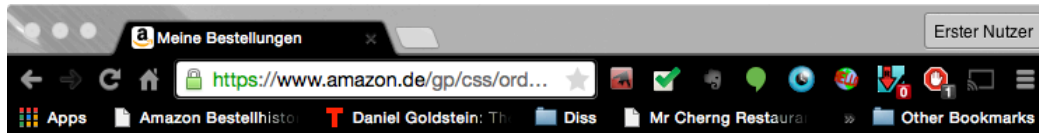
- **Easy way I:** Export items directly from Amazon using Developer Mode of your browser (Java Script Console), a special JavaScript code snippet and a Python script to export the data (you will also get nice statistics about the amount of money you already spent at Amazon...)

Please allow  
popup windows  
for 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)

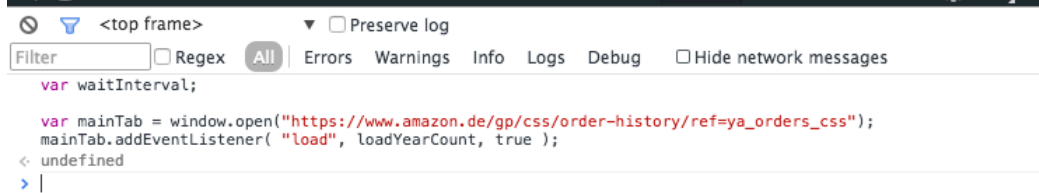


### Übersicht

Jahr	Euro	Bestell.	Produkte	Euro/Prod.	Euro/Monat
Insg.	7.639,44	271	333	22,94	86,81
2015	726,21	35	38	19,11	181,55
2014	2.153,06	42	50	43,06	179,42
2013	1.151,05	59	71	16,21	95,92
2012	512,48	23	24	21,35	42,71
2011	1.104,64	32	44	25,11	92,05
2010	921,06	34	44	20,93	76,75
2009	784,31	36	49	16,01	65,36
2008	286,63	10	13	22,05	23,89

### Einzel-Bestellungen

Link	Datum	Produkte	Preis	Produktbeschreibungen
<a href="#">Link</a>	22. April 2015	1	19,99	• Pampers Easy up Gr.5 Junior 12-18 Kg Mega plus Pack, 1er Pack (1 x 88 Windeln)
<a href="#">Link</a>	22. April	1	5,27	• Nivea Men Active Fresh Spray, 4er Pack 4 x 150 ml

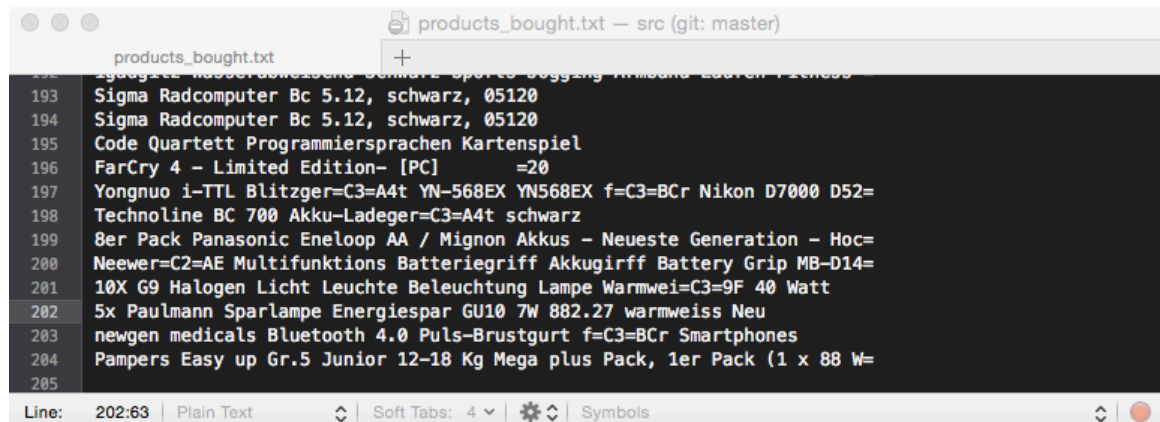


- After some time, a nice result page is shown in the browser...
- Before you start to reconsider your spending behavior, save the page (file, save as, ...)
- Run „AmazonOrderAnalyzer“ to extract the bought items to a file called „products\_bought.txt“

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

## 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)



The screenshot shows a text editor window titled "products\_bought.txt - src (git: master)". The file content is as follows:

```
193 Sigma Radcomputer Bc 5.12, schwarz, 05120
194 Sigma Radcomputer Bc 5.12, schwarz, 05120
195 Code Quartett Programmiersprachen Kartenspiel
196 FarCry 4 - Limited Edition- [PC] =20
197 Yongnuo i-TTL Blitzger=C3=A4t YN-568EX YN568EX f=C3=BCr Nikon D7000 D52=
198 Technoline BC 700 Akku-Ladeger=C3=A4t schwarz
199 8er Pack Panasonic Eneloop AA / Mignon Akkus - Neueste Generation - Hoc=
200 Neewer=C2=AE Multifunktions Batteriegriff Akkugirff Battery Grip MB-D14=
201 10X G9 Halogen Licht Leuchte Beleuchtung Lampe Warmwei=C3=9F 40 Watt
202 5x Paulmann Sparlampe Energiespar GU10 7W 882.27 warmweiss Neu
203 newgen medicals Bluetooth 4.0 Puls-Brustgurt f=C3=BCr Smartphones
204 Pampers Easy up Gr.5 Junior 12-18 Kg Mega plus Pack, 1er Pack (1 x 88 W=
205
```

The editor interface includes a status bar at the bottom showing "Line: 202:63 | Plain Text | Soft Tabs: 4 | Symbols".



**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):

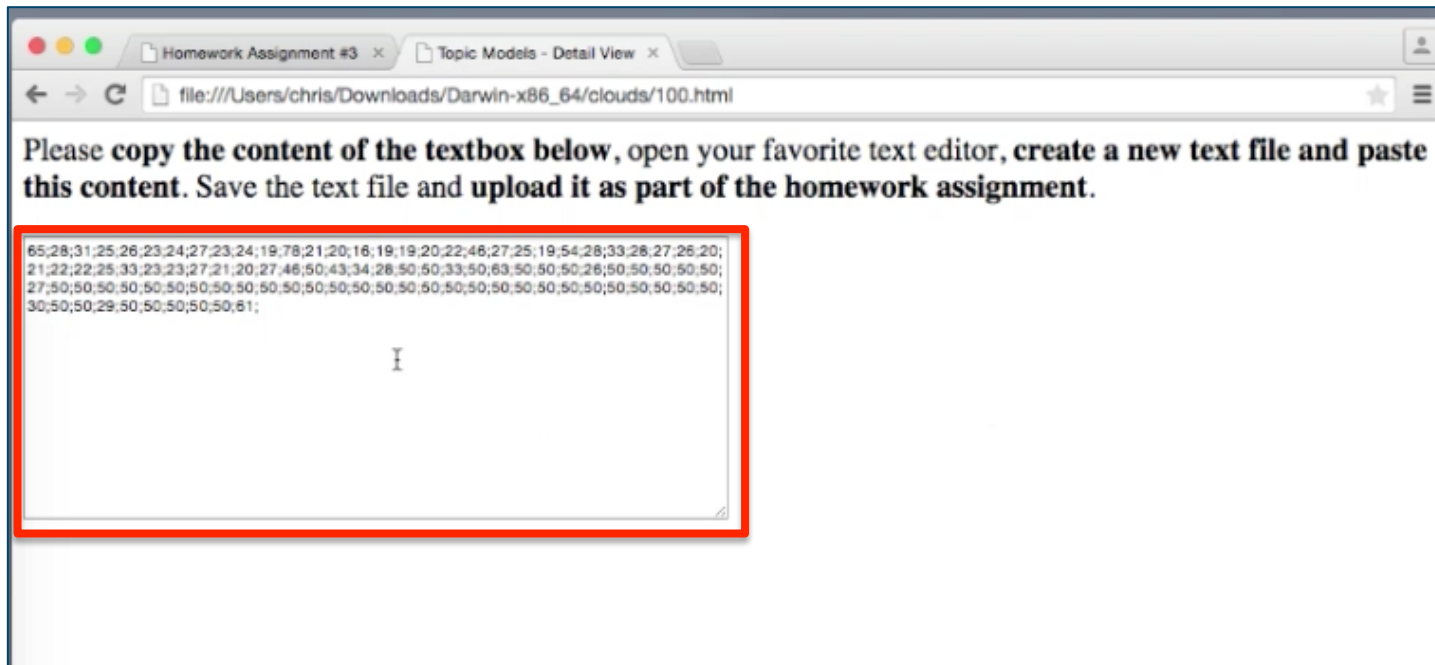
```
./cloudcreator --topic_model_dir . --output_dir clouds
```

- (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”



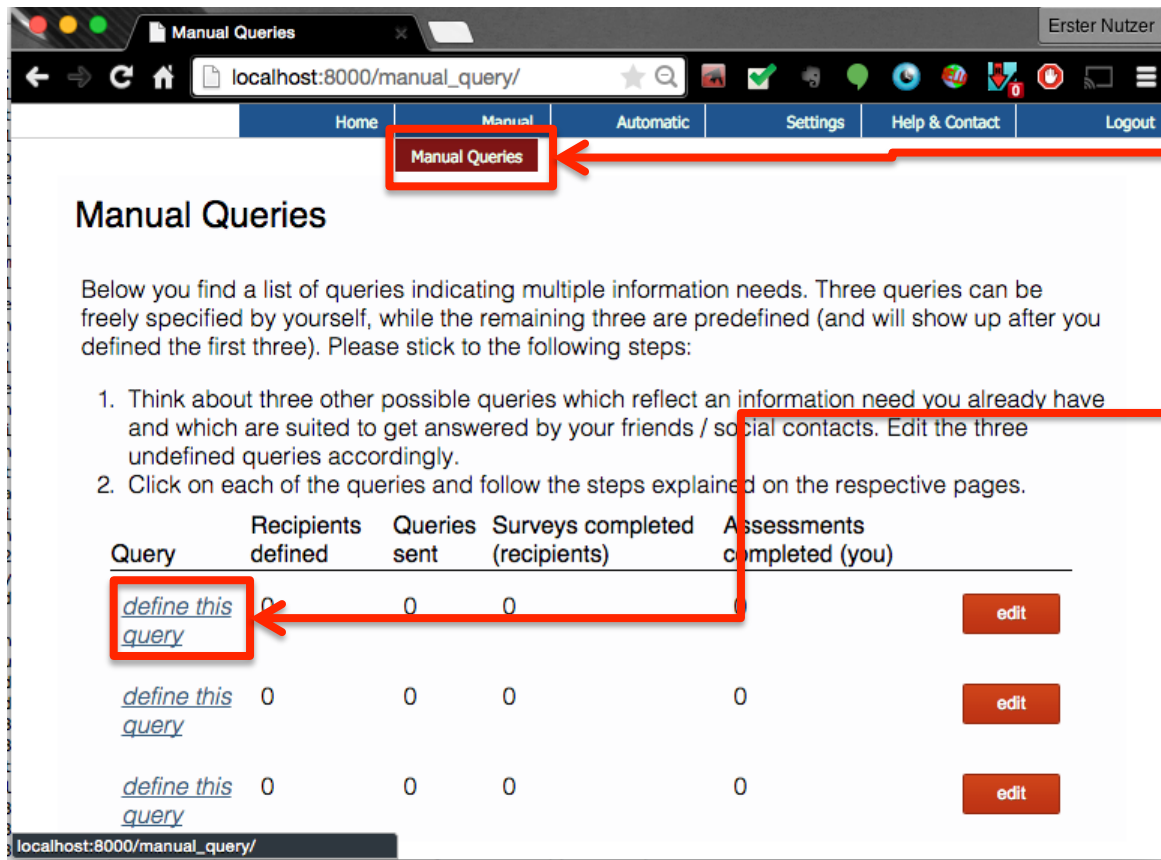
## 1) Run cloudcreator (2/2)

- 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



The screenshot shows a web browser window with the URL `localhost:8000/manual_query/`. The page has a navigation bar with links: Home, Manual, Automatic, Settings, Help & Contact, and Logout. The 'Manual' link is highlighted with a red box and an arrow pointing to it from the right. Below the navigation bar, the page title is 'Manual Queries'. A paragraph of text explains the purpose of the queries. Below this, there are two numbered instructions. The first instruction is highlighted with a red box and an arrow pointing to it from the right. Below the instructions is a table with the following columns: Query, Recipients defined, Queries sent, Surveys completed (recipients), and Assessments completed (you). The table contains three rows, each with a 'define this query' link and an 'edit' button. The first row's 'define this query' link is highlighted with a red box and an arrow pointing to it from the right.

Query	Recipients defined	Queries sent	Surveys completed (recipients)	Assessments completed (you)
<a href="#">define this query</a>	0	0	0	0
<a href="#">define this query</a>	0	0	0	0
<a href="#">define this query</a>	0	0	0	0

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

localhost:8000/manual\_query\_edit/?q...

Erster Nutzer

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

Automatic Settings Help & Contact Logout

information needs. Three queries can be predefined (and will show up after you log in):

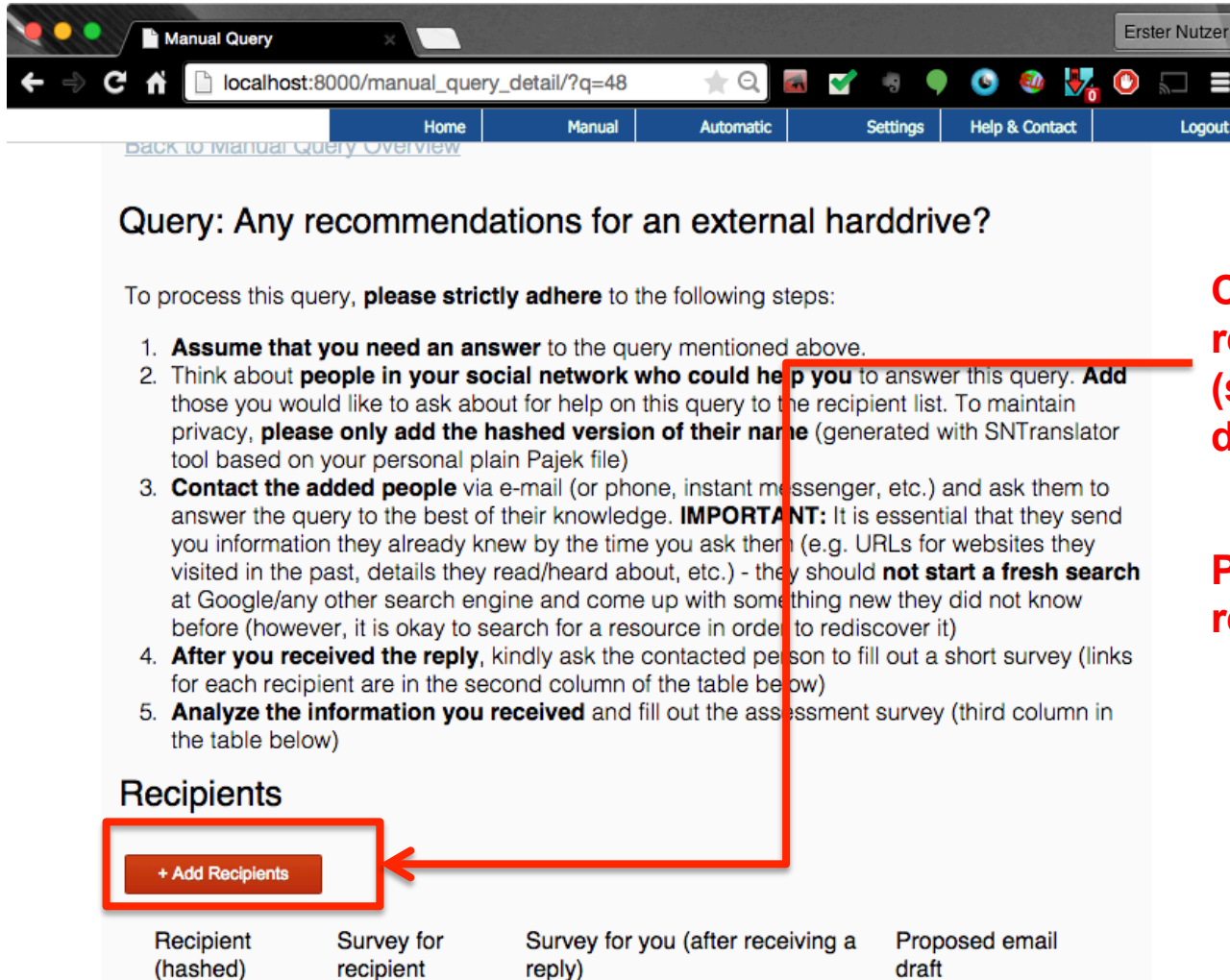
Select an information need you already have and which are suited to get answered by your friends / social contacts. Edit the three predefined queries accordingly.

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)	
<a href="#">define this query</a>	0	0	0	0	<a href="#">edit</a>
<a href="#">define this query</a>	0	0	0	0	<a href="#">edit</a>
<a href="#">Any recommendations for an external harddrive?</a>	0	0	0	0	<a href="#">edit</a>

Newly added query

### 3) Add Recipients for Query



The screenshot shows a web browser window with the URL `localhost:8000/manual_query_detail/?q=48`. The page title is "Manual Query". The user is logged in as "Erster Nutzer". The navigation bar includes links for Home, Manual, Automatic, Settings, Help & Contact, and Logout. A link "Back to Manual Query Overview" is also present.

**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 already knew by the time you ask them (e.g. URLs for websites they visited in the past, details they read/heard about, etc.) - they should **not start a fresh search** at Google/any other search engine and come up with something new they did not know before (however, it is okay to search for a resource in order to rediscover it)
4. **After you received the reply**, kindly ask the contacted person to fill out a short survey (links for each recipient are in the second column of the table below)
5. **Analyze the information you received** and fill out the assessment survey (third column in the table below)

**Recipients**

[+ Add Recipients](#)

Recipient (hashed)	Survey for recipient	Survey for you (after receiving a reply)	Proposed email draft

**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

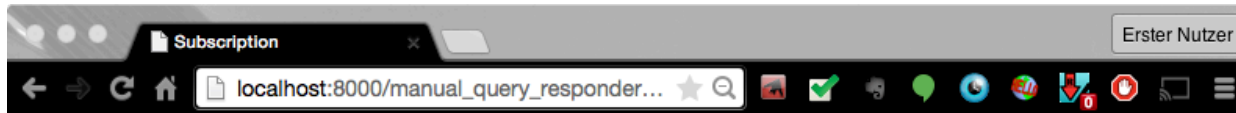


The screenshot shows the SNTranslator application window. The title bar reads "SNTranslator". The main content area has a label "Please select plain Pajek file" above a text input field containing the file path: "/Users/christoph/Dropbox/08 - Dissertation/14 - Experimente/01 - Milgram/99 Testdaten/Pajek-". To the right of the input field is a red circle with the number "1" and a button labeled "Choose file". Below the input field is a dropdown menu showing "User B (userb)" with a red circle and the number "2" next to it. At the bottom, a text field displays a long hexadecimal hash: "1c05c29e17ea6db1bde4a733085782ac17187f912c1d493444b31dc7eb2ce61e". To the right of this field is a red circle with the number "3" and a button labeled "Copy hash to clipboard".

- 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 SN Translator 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 already know (e.g. links for websites they visited in the past, data at Google/any other search engine they used before (however, it is ok if they did not know over it) but a short survey (links to the survey (third column in the table below)
4. **After you received the information** for each recipient are in the table below
5. **Analyze the information** (the table below)

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

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

### Recipients

+ Add Recipients

Survey for you (after receiving email a reply) Proposed draft

Recipient (hashed)

Survey for recipient

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<sup>1</sup>
- 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

<sup>1</sup> 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 **will be deleted afterwards**
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