Mael Brocher Logbook
Interactive Visualisation of English and French Semantic in Passwords

25th May

I’m finally done with exams and I can now focus on my thesis.
But after talking to Shujun about it , I am not doing my thesis about Passwords in a universe as the previous student has not submitted his work in time. I will work on Interactive heatmaps.

16th June

I went through most of Stephanie's work and I found that there are a LOT of bugs present, for example some make the webpage freeze, or I simply couldn’t find a lot of functionalities described in her paper.
Maybe the version of the software I got isn’t her final work.
Despite all of these, the project is quite easy to handle and I could run some tests and analysis of the password database quite rapidly.
I also think that we could deploy the project on the internet to make it more accessible.

20th June

I discovered a strange bug due to the button of upload and setting being movable. Whenever the user would click on one of these two buttons and move a little his mouse, it would move the button and not perform the action. Before the fix one in three times the action would not be performed.

21th June

I changed cells with frequency of 0 to impossible to make the heatmap easier to look at.

30th June

We reviewed my work about the early deliverable. I clearly didn’t understand what I needed to do with the project and focused too much on Stephanie Schmid version and UI improvements but it was out of scope.
I removed the changement of cells with frequency of 0 considered as impossible because it’s not the same thing. Impossible means that nothing could ever happen in a cell considering the database, and frequency of 0 only means that no password contains the character at a certain position but could be present in the database.

I have to re-do my tests on Gary Read works because we think that Stephanie version

2nd July

I made the div for the title of heatmaps longer so we can quickly understand which database we are looking at.
I implemented some of Stephanie's work into Gary code: the copy button and when a database is uploaded it’s selected by default.

7th July

We got a meeting and conclude that I need to set difference of two database to see what removing passwords in common does

8th July

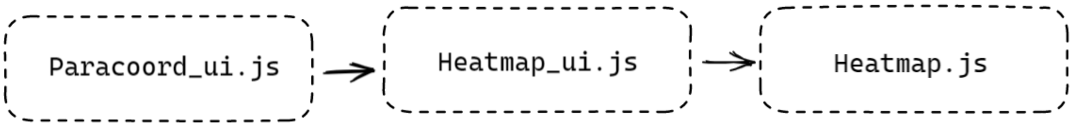
Since coding the set difference was fairly easy I also added the set intersection.

15th July

We just got a meeting, I will have to work with axis reordering.

20th July

I think this is the most difficult part of the thesis because it is the first time I really go in the code and I had no idea how the tool was working.



I tried for hours to get the heatmap.js function from the paracoord_ui.js file but this isn’t possible and I didn’t know it. Gary didn’t specify in the code that heatmapUI and heatmap class were different. In fact each heatmapUI has a heatmap and they can call the heatmap function. So I had to create a lot of duplicate functions in heatmap_ui.js to have heatmap.js data from paracoord_ui.js.
Once I understood that, I changed the way of handling axis Order and created a small section in the parameters to modify it.
It was hard to make it work because the base string used for is stored in a config file. So I decided to process the said string according to the wanted Order before drawing the heatmap.
I also added a way to put a custom axis with a text field that adds to the start the wanted word.

22th July

We just got a meeting and I have to do a quick modification right after the meeting. Heatmap that just got reordered should be a child and not a copy of the heatmap.
Shujun told me to look for wordninja, wordsegment and symspell.
He also suggested that I could do heatmaps on a keyboard and color keys like I color heatmap's cells. It sounds cool but I have no idea how to do it and maybe not enough time.

23th July

Wordninja and wordsegment are pretty interesting and give good results for english.
After some test I saw that wordninja is way quicker and better than wordsegment to split words in a password.
I have to find a dictionary ordered by frequency to make it work in other languages.
Also wordninja is in python and the only way I found to call python from Js is to deploy a python server and make ajax call from Js.
I also discovered Spacy that does part of speech tagging in many languages.

27th July

As the original wordninja regex is not accepting characters outside [A-z] and the French language is full of à, é, è, ê, or ô I needed to download the source code and modify it.
Also I duplicate wordninja, one for English language and the other for French language to avoid loading a different dictionary each time the user wants a semantic analysis.
I created a small section withs button that just makes the ajax call in English or french.
Everything is ready on the server side, I can make a complete array of POS tags from a heatmap I just sent from the webpage.
I decided to make it multithreaded and the server processes one language per thread and adds the spacy analysis if the language detected matches the language of the thread .

28th July

Shujun validated what I did but I need better dictionaries because the one provided by wordninja has no real sources and in french I used a frequency ordered dictionary based on books.
So he advised me to look at wiktionary and I could find some with trusted sources.

1st August

I did a quick scraper with python and selenium that wrote the words in a file.
He also told me to deploy the index with the rest of the server made with flask.
I have some issues finding a module that could help me to detect language in python.
Most of them are API based and are limited in number of calls per day (e.g. googletrans).
Langdetect seems to be fine, it's offline but quite slow.

5th August

I finally made my first semantic heatmap, it seems that I have an issue with wordninja as many Spacy detect a lot of unknown words.

I made an alternative of processing, instead of processing when the user want, data is processed when the user upload a database on the webpage
So I made a quick animation on each heatmap showing if the result of the semantic analysis is ready or not.
I got a bug where the first semantic heatmap is empty.

11th August

I resolved the bug by making this call synchronous as this is almost instant.
I remade the entire code for semantic heatmap, by adding a semantic version of the components required to make a heatmap. Making the histogram and the cell highlight function is working for semantic heatmap.

12th August

I remade the way the python server was handling semantics heatmaps, by storing them like the webpage keeps them.
It went from a list of words with duplicates to a list with words with their frequency.

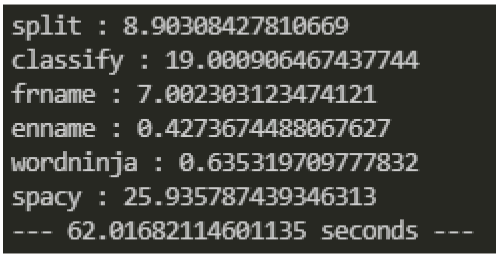
15th August

I had a hard time with parallel coordinates because it is drawing normal heatmap parallel coordinates on top of a semantic heatmap.
I finally made the parallel coordinates work for semantic heatmaps, I simply forgot to check if I was dealing with semantic heatmap or not before drawing the lines.

19th August

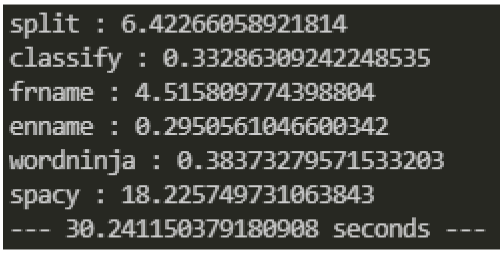
I just made my programs 8 times faster by changing my language detection module and switching from list to dict as python is quicker with dict and removes some useless spacy analysis.
Screenshots are from tests with the wordlists “10k most common with frequency.txt”.

before any change :



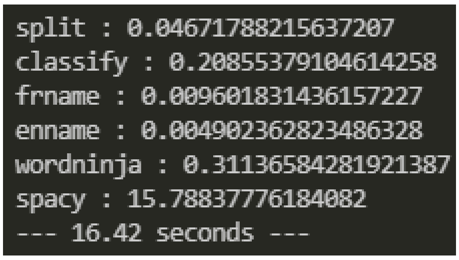
```
split : 8.90308427810669
classify : 19.000906467437744
frname : 7.002303123474121
enname : 0.4273674488067627
wordninja : 0.635319709777832
spacy : 25.935787439346313
--- 62.01682114601135 seconds ---
```

after changing from langdetect to fasttext :



```
split : 6.42266058921814
classify : 0.33286309242248535
frname : 4.515809774398804
enname : 0.2950561046600342
wordninja : 0.38373279571533203
spacy : 18.225749731063843
--- 30.241150379180908 seconds ---
```

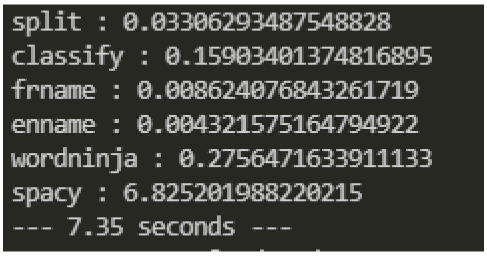
switch from list to dict :



```
split : 0.04671788215637207
classify : 0.20855379104614258
frname : 0.009601831436157227
enname : 0.004902362823486328
wordninja : 0.31136584281921387
spacy : 15.78837776184082
--- 16.42 seconds ---
```

.

cleaning the Spacy pipeline :



```
split : 0.03306293487548828
classify : 0.15903401374816895
frname : 0.008624076843261719
enname : 0.004321575164794922
wordninja : 0.2756471633911133
spacy : 6.825201988220215
--- 7.35 seconds ---
```

I also changed how I process a heatmap. Now it processes only one time the database and the language detected is stored in the array with the raw words, his frequency and the semantic analysis.