Task 25/08/2020 :

As discussed in the last meeting, the aim was to create a model that takes only the 20 most occurring websites in the whole dataset since we have already filtered and kept only the relevant domains, subdomains and tags :
https://docs.google.com/document/d/1ZumFo_s8Urrl0H9sAx245fDHfmyA_tyaWQlnoqE_nqI/edit?usp=sharing

The original number of articles they had was **86346** out of **118487** total records. After filtering out the irrelevant ones, I was left with **36193** records which is still good for building the model and the maximum number of occurrences of each article was **25.** This indicates that we got rid of most of the irrelevant advertisements and comment blogs.
This whole filtering can be later extended to all **179** domains.

Fine-tuning :

When it comes to the model, I tried at first with the normal data preprocessing but I found out that there are some words occurring a lot in the topics' keywords but irrelevant to the topic itself (screenshot below). So, I added them to the stopwords and deleted them from the records which gave better topic selection with better interpretation.

```
stopwords_verbs = ['say', 'get', 'go', 'know', 'may', 'need', 'like', 'make', 'see', 'want', 'come', 'take', 'use', 'would', 'can']
stopwords_other = ['one', 'mr', 'bbc', 'imag', 'de', 'en', 'caption', 'also', 'copyright', 'something',
                   'nh', "getti", "pa","don","ap","afp","reuter","pictur", "ve", "didn", "share","septemb","august","octob","jo","thoma"
                  "st","ms"]
my_stopwords = SW + stopwords_verbs + stopwords_other
```

I tried the model for multiple number of topics between 8 and 20 but the best one in terms of interpretability is the one with 16 topics :

1. (polic attack man kill die murder death offic arrest incid) : **Crime OR Terrorism**
2. (win play candid final fan team film season tv seri) : **Sports**
3. (health school children work univers hospit patient cancer student care) : **Health - The education system**
4. (court case claim prison hear jail abus sentenc month charg) :**Judicial system**
5. (royal queen dress wear black harri princ event palac white): **Royal family**
6. (citi water countri game unit target nation forc climat player) : **The environment, climate and energy issues**
7. (babi feel good night start week life woman hand women) **lifestyle**
8. (credit news email uk sun pay mum stori flight address), **vacation ??**
9. (brexit parti mp deal tori vote minist eu labour leadership), **brexit**
10. (pension hous council build work street live properti local plan) **Pensions**
11. (licenc pay free govern cost uk servic compani busi fund) **Taxation - Government debt**
12. (london protest england scotland june juli independ street action celebr), **Rising prices/inflation/cost of living OR Unemployment**
13. (love famili star daughter son mother instagram father life friend), **lifestyle**
14. (trump presid donald visit birthday meet hous morgan russian fox), **related to politics**
15. (media twitter comment social post facebook tv devic write video) **social media**
16. (car road servic park west bu driver south passeng drive), **vacation ?**

As you can see above, most topics could be labelled but some still need more investigation. So, the best way was to increase the number of topics in the training so that we go deeper into the subclasses and identify them easily. I am trying to find the optimal number of topics since the training and fine-tuning takes time but it's almost done.

In parallel, I created the word embeddings and now searching for a way to incorporate them in the model to enhance topic modeling.

For next week, I'll finish the topic labelling and try to implement lda2vec and word2vec to compare with the one we already have.

My questions are :

- Is there another file containing user ids and the urls they navigated? If yes , is it possible to share it with me so that I start checking the matching technique (topic user profiles)?
- What do you think about extending the 20 most occurring newspapers list into the total number ?
- Any other suggestions, please let me know.