Report for Week 1 – 01 April – 05 April

I. Actions performed

1. We had a meeting with Ms. Largeron where we discussed the plan for the internship and the main steps that need to be followed
2. I took some time to understand each step, started researching about step 1. **Vectorial representation of textual contents**
3. I did read the **lecture** about Text Mining provided by Ms. Largeron, to get accustomed with the first approach in encoding words into vectors of real numbers.
4. I did read some articles / tutorials about Bag of Words – Tf, Tf-Idf, with normalization. Limitations: what if we have a super large corpus with a lot of different words?
5. I did read some articles about Word Embedding – Word2Vec, GloVe, fastText. Shallow Neural Network with one hidden layer, one input layer and one output layer. Two flavors, Continuous Bag of Words vs Skip-Gram. We will use the word embedding models trained on the English TenTen corpora – still need to figure out how to use it in practice.
6. I did the **practical exercises** about Text Mining to get accustomed with the R functions for dealing with text.
7. I got accustomed with the structure of Reddit, what is a subreddit, what is upvoting, divergent opinions etc. and mapped the elements we will consider for each node of the tree (tuples – user, content, timestamp)
8. Some nice subreddits: https://www.reddit.com/r/brexit/, https://www.reddit.com/r/politics/, https://www.reddit.com/r/ukpolitics/
9. Started looking for bibliography (related works to this one) – for now, I could only search for similar papers w.r.t the first step – vectorial representation – saved them in Zotero.
   - Cataldi et al. - 2010 - Emerging topic detection on Twitter based on tempo.pdf
   - Choi et al. - 2015 - Characterizing Conversation Patterns in Reddit Fr.pdf
   - Klenin and Botov - Comparison of Vector Space Representations of Docu.pdf
   - Guille et al. - 2013 - Information diffusion in online social networks a.pdf
   - Villegas et al. - Vector-based word representations for sentiment an.pdf

II. Encountered Difficulties

1. How will we choose the Dictionary for the Text Mining approach? All unique words from the corpus? What if there are a lot of them? Limit them to a number?
2. What if the matrix will be sparse (high chances)? Should we use n-grams?
3. Data?

4. Reading papers is not my cup of tea

III. Plans for next week

1. Start implementing the first step, once I have the data.
2. Start researching about the second step**: Non-Supervised Clustering of the users based on the content they posted**.
3. Search for bibliography regarding the second step.