Master's in Modelling for Science and Engineering **Mathematics and Big Data**
Text Mining : Document Classification Techniques *Curs* 2021 – 2022*: Practica* 2

1 Document Classification and Representation

Following tasks are to be performed with Python packages `numpy, pandas, scipy, re, string`.

2 Case Study : Whatsapp Chat

2.1 Term Frequency and Inverse term frequency

1. Count the frequency of each term w_i in document d_i.
2. Print the list of terms with their frequencies.
3. Find inverse term frequency for words in each document.
4. Plot the term frequencies.
5. Visualize the terms with wordcloud.

2.2 Document Term Matrix

1. Generate a doucment term matrix for all the documents d_i and terms w_i in each document.
2. Visualize the words with wordcloud and bar plot.

2.3 Document Similarity

1. By using distance based matching find the most similar documents in corpus C.
2. Apply angle based matching on the documents found and compare the results.
3. By using the measures of precision and recall find the most similr documents and compare with previous results.

Delivery 2: Document Classification and Representation - Twitter Tweets

1. Read twitter tweat data from file tweets.txt.
2. Create a corpus of documents by using this data.
3. What does this data reflects?
4. Preporcess the data for required cleaning up to gain insight about data.
5. Represent the document in matrix form.
6. Find the most relevant words from all documents.
7. Visualize the word frequencies found in document term matrix using wordcloud.