* Form a group of 4 and send me an e-mail (d.kurowicka@tudelft.nl) with names and student numbers of students in each group.
* Go to the website Trump Twitter Archive <http://www.trumptwitterarchive.com/archive> which contains 32929 tweets. Choose one year of data (you can save it in Jason or CVS format, depending which format you want to use later ). It would be interesting to see if main subject of tweets change from one year to the other, so different groups can take different years)
* You can use R package, Matlab or Python (Python is as far as I know the most efficient in memory use. This might be not too important for this assignment but in analysis where really large amount of data is available you might be forced to use this software)
* First you have to pre-prepare the file (remove stopwords, do stemming etc… to extract the text for the analysis where you have only ‘important’ words and similar words are identified)
* Analyse these tweets (check if everything went well with stem procedure by comparing original tweets  and words after stemming, analyse the most frequent words etc. )
* Perform clustering analysis with different methods.
* Compare the results from different clustering algorithms.
* Analyze the optimal number of clusters, and what different clusters are about.
* Check few randomly chosen, original tweets to see if indeed the tweets were properly divided by subject with clustering algorithms.
* Be resourceful and try to impress me with your analysis.
* Send me report with your analysis - deadline 15.03.2018 at 24.00.
* Include the copy of the implementation in the end of the report. Also send the file with tweet analysis implementation.
* Since it is a group work I would like you to indicate in your report who did what (group assessment where you will evaluate the group members for the involvement in this assignment).