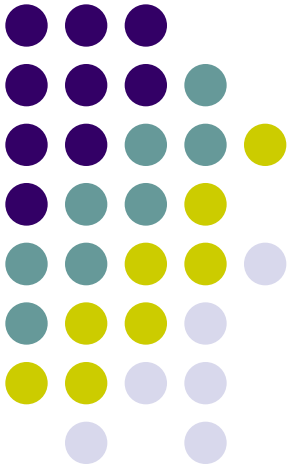
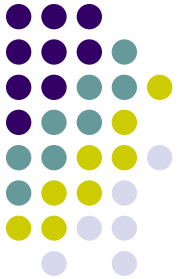
Polimetrics

Fifth Assignment



Deadline: 13 November 2018



First part: Dictionaries

Focus on the UK party manifestoes in the 2015 elections

1. Identify which have been the most negative (and the most positive) documents *in absolute terms* [not needed, but appreciated: focus also on *relative terms* (discounting the total number of positive and negative words according to the number of words employed in each party manifesto)]
2. *By applying a dictionary*, try to estimate the positions of the UK parties with respect to some policy domain you are interested about *in absolute terms* and briefly comment your results [not needed, but appreciated: focus also on *relative terms* (discounting the total number of words dedicated to discuss a particular policy domain according to the number of words employed in each party manifesto)]

Deadline: 13 November 2018

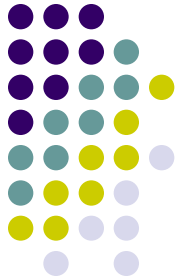
Second part: Classifiers

The dataset "Trump-orig3.csv" is a sample of tweets mentioning the official account of Donald Trump "@realDonaldTrump", on dates 7–13 June 2016, written in English and coming from the US. This dataset include a sample of around 472 tweets that have been manually codified by a group of students. The coding stage involved detecting the sentiment towards Trump (negative, positive, neutral).Use this dataset as your **training-set**

As a **test-set** use the dataset "Trump_tweets2.csv", that referst to a sample of 1000 tweets written in English about Trump and published since 1.17.2018 till 1.19.2018

Of course this is just an exercise! Not a great idea to use a training-set of 2016 to classify a test of almost 2 years later!

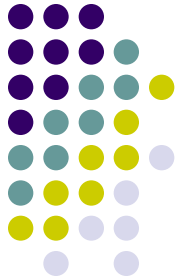
Deadline: 13 November 2018



Second part: Classifiers

Then: 1) estimate both a Random Forest as well as a Naïve Bayes Model trained on the training-set to estimate the test-set; 2) then run a k-fold cross-validation (with k=2; if you want to do it with k=3 you are welcome!) for just one of the two Models (either RF or NB)

Deadline: 13 November 2018



Second part: Classifiers

REMEMBER: a) a classifier such as Random Forest Model requires that your dependent variable is not a character, but a factor (take a look at which type of variable is “sentiment” in the "Trump-orig3.csv” dataset). To transform a character in a factor use the command: *as.factor(name of the character variable)*; b) to sum two corpus is easy! Just remember the discussion in our first class! For example, *corpus3 <- (corpus1+corpus2);* c) to select a corpus including only missing values for a given variable *x*, type is.na(x); to select a corpus including all these values that are NOT missing for a given variable *x*, type !is.na(x)