ICYBM 201 2019-2020

The main objective is to reproduce some of the results of the reference paper:
Fame for sale: efficient detection of fake Twitter followers, S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi. arXiv:1509.04098 09/2015. Elsevier Decision Support Systems, Volume 80, December 2015, Pages 56-71

It consists in the automatic detection of Twitter accounts specifically created to inflate the number of followers of some target account (so called fake Twitter followers). The paper makes available an annotated database of user accounts, available here:
https://botometer.iuni.iu.edu/bot-repository/datasets/cresci-2015/cresci-2015.csv.tar.gz

Accounts are tagged as human versus fake, and characterized by a number of attributes, including behavioural data, profile description, etc. By means of tools from data mining, the goal will be to reproduce some of the results from the paper, such as those of table 16:

		evaluation metrics					
	algorithm	*accuracy*	*precision*	*recall*	*F-M.*	*MCC*	*AUC*
Class C classifiers that use all the features							
RF	Random Forest	**0.994**	**0.997**	0.990	**0.994**	**0.987**	**0.999**
D	Decorate	0.993	0.993	**0.993**	0.993	**0.987**	**0.999**
J48	Decision Tree	0.992	0.991	0.992	0.992	0.983	0.993
AB	Adaptive Boosting	0.987	0.988	0.987	0.987	0.975	**0.999**
BN	Bayesian Network	0.960	0.965	0.954	0.960	0.921	0.991
kNN	k-Nearest Neighbors	0.971	0.963	0.979	0.971	0.941	0.990
LR	Logistic Regression	0.986	0.985	0.988	0.986	0.972	0.996
SVM	Support Vector Machine	0.987	0.983	0.991	0.987	0.974	0.987
Class A classifiers that use only Class A (profile) features							
RF	Random Forest	**0.987**	**0.993**	0.980	**0.987**	**0.967**	**0.995**
D	Decorate	0.984	0.987	**0.981**	0.984	0.964	**0.995**
J48	Decision Tree	0.983	0.987	0.979	0.983	0.962	0.983
AB	Adaptive Boosting	0.972	0.975	0.969	0.972	0.941	**0.995**
BN	Bayesian Network	0.966	0.969	0.963	0.966	0.928	0.991
kNN	k-Nearest Neighbors	0.957	0.961	0.953	0.957	0.914	0.978
LR	Logistic Regression	0.971	0.964	0.978	0.971	0.942	0.987
SVM	Support Vector Machine	0.961	0.972	0.950	0.961	0.923	0.961

Table 16: Performance comparison for 10-fold cross validation. Training set: BAS.

Evaluation will be based on a written report, where the students introduce the thematics, describe their results and interpret them. All codes should also be provided. **Deadline: 15/03/2020**.

Reports and codes should be sent by e-mail to vsevolod.salnikov@unamur.be

Some interesting visualisations to play with

- Network vaccination
- Random Walks & Diffusion
- Network models and connected components

Slides