IST Austria: Data Science and Scientific Computing
Track Core Course 2016/17 – Segment 1: "Predictive Models"

Instructor: Christoph Lampert <chl@ist.ac.at>
TAs: Bor Kavcic <bor.kavcic@ist.ac.at>, Katharina Ölsböck <katharina.oelsboeck@ist.ac.at>,

Final Project

Project Description

The goal is to construct the best classifier for predicting who wrote a certain tweet from text and/or meta-data.

Procedure

1) form small working groups. Each group should contain at most 2 participants who want ECTS points. A third participants can join who just audits the course. Groups are encouraged to interact with each other and share experience, but ultimately each group must produce its own model and presentation.

2) download the JSON file `http://www.ist.ac.at/~chl/courses/DSSC-TCC17/AllTweets-train.json` (or the corresponding `csv` file, which contains less meta-data though). It contains 55539 tweets from different 7503 writers (specified in "user" field).

3) identify all writers with more than 1000 tweets (there should be 9 of them).

4) the goal of the project is to produce a classifier that for any new tweet can predict who wrote it (which of the 9 frequent writers or "somebody else"). You can use any model, method, software package, etc., that you want for this, relying on text, meta-data or both. You can also rely on additional resources, e.g. other text representations, but you may not make use of additional tweets to what is provided in the above file.

5) when you have produced a classifier that you consider final, send an email to the TAs declaring that you are done. You may not change you classifier anymore after this step, even if you find bugs, or similar!

6) the TAs will reply to you with a password that allows you to unzip the archive `http://www.ist.ac.at/~chl/courses/DSSC-TCC17/AllTweets-test.zip`, which contains new data of the same formats as above.

7) apply your classifier the evaluation data and evaluate its multi-class accuracy.

8) identify other interesting tasks or question you could ask, e.g.who is commonly confused with who? Which tweet by Hillary Clinton sounds most like Donald Trump? Or build an interface where users can enter their own tweets and the system tells them which writer it resembles, etc.

9) prepare a short team presentation (10 minutes plus 5 minutes for questions) that explain all steps you did to reach the model and why you made them. In particular, report the results you achieved and put them into context (how good are they? what would a trivial baseline method achieve, e.g. predicting the majority class? how good would a person do in comparison?). If you identified problems with the model after step 5), report and discuss them.

10) the presentations will take place on March 22nd. All team members must actively participate in this.

Time Frame

- March 8: official begin of project period
- March 15, March 20: no lectures, but time for Q&A
- March 22: team presentations