HPSA Projects

Roberto Trani

roberto.trani@isti.cnr.it

HPSA Projects

- Consist of applying concepts and tools learned in this course to real use-cases
- Involve employing a data mining pipeline on real data in a distributed environment
- Six projects are available, although new projects can be proposed to the tutor before May, 4th

Rules

- The projects must be developed in teams
  - Each team is composed of at most six people
  - Each team works on a different project
  - Team members and project must be communicated via e-mail to the tutor **before May, 4th**

Submission

- The teams must submit the following files via e-mail to the tutor **before 23:59 CET of May, 19th**:
  - A final report (at most two pages) with: *i*) introduction and objective, *ii*) data understanding and preparation, *iii*) feature extraction and modelling, *iv*) assessment.
  - A script or a jupyter notebook with the source code needed to replicate all experiments.

Spam classification of SMS

- **Objective**: build a SMS spam detection model predicting if a given SMS is spam or not.
- **Dataset**: https://www.kaggle.com/uciml/sms-spam-collection-dataset
  - ~5,600 SMS messages

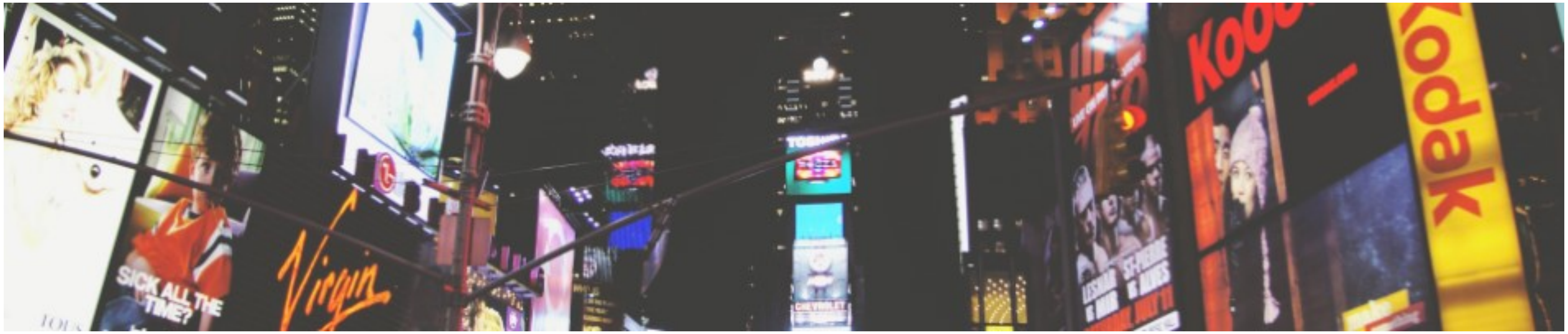
Sentiment analysis of movie reviews

- **Objective**: build a sentiment-analysis model of movie reviews predicting if a given review is negative or positive.
- **Dataset**: https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data
  - ~156,000 reviews

Identification of beer categories

- **Objective**: build a beer classification model predicting if a beer is an "*American IPA*" or not based on its properties.
- **Dataset**: https://www.kaggle.com/jtrofe/beer-recipes
  - ~74,000 beers

Stock prediction using news



- **Objective**: build a stock prediction model predicting if a given stock goes up or down by incorporating news data.
- **Dataset**: https://www.kaggle.com/aaron7sun/stocknews
    - ~2,000 days of stock data and ~76,000 news

Segmentation of bank customers

- **Objective**: segment the bank customers based on their credit card usage behaviour.
- **Dataset**: https://www.kaggle.com/arjunbhasin2013/ccdata
  - ~9,000 customers

Segmentation of e-commerce customers

- **Objective**: segment the e-commerce customers based on their transactions.
- **Dataset:** https://www.kaggle.com/carrie1/ecommerce-data
  - ~542,000 transactions

Evaluation

- What affects the grading:
  - achievement of the project objective
  - usage of the different tools shown in the course
  - original ideas are a plus
- What does not affect the grading (if it is not so bad…):
  - quality of the results
  - quality of code and report

Roberto Trani

roberto.trani@isti.cnr.it