Web Mining for Business Application
Fall 2019

Course Project

Form:	Jupyter notebook file
Language:	English
Requirements:	The report should be clear, readable and include all code documented
Submission:	Jupyter notebook via Moodle. The file name should include the students' ids
Contact:	simohanouna@gmail.com Shimon Hanouna
Deadline for submission:	1.3.2020

Students will form teams of two or three people, and submit a single project for each team. The same score for the homework will be given to each member of the team.

Submit your solution in the form of an Jupyter notebook file (with extension ipynb). The analysis and prediction results should be clearly presented as part of the notebook report. Images should be submitted as PNG or JPG files. Python 3.6 should be used.

The goal of this homework is to let you practice different tools which you learnt during the course.

Submission: The report should contain an overview of the process, a detailed response to each question, challenges you faced and conclusions. The source code should be included and documented.

Project 1: Twitter Sentiment Classification

This project is based on Kaggle competition: https://www.kaggle.com/kazanova/sentiment140.
Question 1: Download the dataset. Perform text pre-processing. Present data exploration: class distribution, terms frequency for the different sentiments.
Question 2: Train a machine learning model to predict the sentiment of the tweet. Evaluate two models\approaches, and tune parameters. One of the models should be based on 'deep learning' approach. Evaluation metrics: accuracy. Present train and test accuracy for the different models and pre-processing combinations.
Question 3: Use Twitter streaming API to collect 15,000 tweets from Twitter. Repeat the same pre-processing you implemented in Question 1 for the collected tweets. Analyze the most popular terms for each sentiment for this test dataset as well. Present terms frequency and discuss the similarity with the train. Clarify how you label the collected tweets' sentiment.
Question 4: Use the best sentiment classification prediction model which was trained in question 2 to predict the sentiment of the collected Tweets. Present the prediction results and your conclusions.

Project 2: Stock Market Prediction

This project is based on Kaggle competition https://www.kaggle.com/aaron7sun/stocknews. The goal is to predict if the Dow Jones Industrial Average (DJIA) index will increase based on news headlines.
Question 1: Download the dataset. Present data exploration: class distribution, terms frequency for each class.

Question 2: Build a machine learning classifier to predict if the DJIA index will increase (including no change) or decrease per day, based on news headlines. Evaluate two models. One of the models should be based on 'deep learning'. Present train and test accuracy for the different models and pre-processing combinations as well as for different combinations of input datasets (news, and historical data). Evaluation metrics: AUC.

Question 3: Crawl updated news data from Reddit WorldNews Channel for 10 days. DJIA data is available from Yahoo! Finance. Repeat the same data exploration you implemented in question 1 for the crawled news data and present the results. Discuss similarity and differences with the train.

Question 4: Use the best machine learning classifier trained in question 3 to predict the DJIA direction per day based on the crawled data.

Good luck