Homework 3: Web Mining for Business Applications Project

Form:	Written report (PDF) + source code
Language:	English
Submission:	PDF sent through email to the TA
Contact:	asnatm@post.bgu.ac.il
Deadline for submission:	**February 11, 2018**

Students will form teams of two or three people each, and submit a single homework for each team. The same score for the homework will be given to each member of the team. The students should choose one of two projects or suggest a project of your own (should be confirmed).

The goal of this homework is to use the different tools presented during the course for business predictions.

The report should contain an overview of the process, design document for the code, a detailed response to each question, challenges you faced and conclusions. The source code should be attached.

Submission: Submission will be done via Moodle. The project needs to be entirely in English. The deadline for submission of the project is set to February 11, 2018.

Project 1: Twitter User Gender Classification

This project is based on Kaggle competition: https://www.kaggle.com/crowdflower/twitter-user-gender-classification.

Question 1: Download the dataset. Perform text pre-processing. Present data exploration: class distribution, terms frequency for the different genders.

Question 2: Train a machine learning model to predict the gender of the tweet author. Evaluate three models, and tune parametrs. One of the models should be based on 'deep learning' with Keras. Evaluation metrics: accuracy. Present train and test accuracy for different model and pre-processing combinations.

Question 3: Use Twitter streaming API to collect 15,000 tweets from the country which was most popular in the training data. Optional: you can filter based on an hashtag which is expected to be related for a specific gender. Repeat the same pre-processing you implemented in Question 1 for the collected tweets. Analyze the most popular terms for this test dataset as well. Present terms frequency and discuss the similarity with the train.

Question 4: Use the best gender classification prediction model which was trained on Question 2 to predict the gender of the authors of collected tweets. Present your conclusions. 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: Identify 4 most correlated terms in Google Trends (using Google Correlate) to the DJIA dataset. Present the Pearson correlation for each term with the dataset.

Question 3: Build a machine learning classifier to predict if the DJIA index will increase (including no change) or decrease per day, based on news headlines, historical DJIA data and Google Trends. Evaluate two models. One of the models should be based on 'deep learning' with Keras. Present train and test accuracy for different model and pre-processing combinations as well as for different combinations of input datasets (news, Google trends, and historical data). Evaluation metrics: AUC.

Question 4: Crawl updated data for 10 days. Use the best machine learning classifier trained in question 3 to predict the DJIA direction per day. News data should be crawled from Redit. DJIA data is available from Yahoo! Finance Yahoo! Finance. Present the results and your conclusions.

Good Luck