Programming Practice

Automated Prediction System - *Stock Market*

Professor: Javier Paris

Text Mining 4-ADE

Contents

1 Task Description

This practice consists of the implementation, in teams, of an **automated prediction system** for data related to the *Stock Market*. The objective is for the student to learn how to use the **text mining techniques** covered in class to solve a real prediction problem.

Teams will consist of a maximum of 3 members.

2 Learning Objectives

1. **Understand** the information workflow and the importance of data collection.
2. **Practice** the text mining techniques covered in class.
3. **Design** an automated predictive system.

3 Development

3.1 Description

This practice focuses on implementing an automated prediction system based on various methods of data collection and processing. The system must be able to **predict data** related to the topic: *Stock Market*.
The teams could be composed of a **maximum of 3 members**.
Each **team will choose a topic** for the project. The topic must be approved by the professor before starting.
The practice will consist of **a single submission** at the end of the semester, including a written report, the code necessary to reproduce the results obtained, and an in-person defense of the practice.
The final system must be written in **Python**, be capable of using various sources to make predictions, and meet the following requirements: use preprocessed historical data, obtain recent data through text mining techniques such as web scraping and APIs, and predict new data using a Machine Learning model.

3.2 Phases

To facilitate the development of the practice, the following phases are recommended:

1. **Topic Selection** Choose a topic aligned with the team's interests and knowledge.
2. **Preliminary Study** Study the chosen topic. It is highly recommended to clearly define what the system will and will not do.
3. **Historical Data Collection** Find datasets of historical data, and preprocess them to generate a base dataset.
4. **Automated Data Collection via Web Scraping** Use web scraping techniques to collect recent data to update the base dataset.
5. **Automated Data Collection via API** Use APIs to collect recent data to update the base dataset.
6. **Predictive Model** Create a Machine Learning model (using scikit-learn) that predicts information based on the dataset.

7. **Model Evaluation** Evaluate the model to understand its effectiveness.
8. **Conclusions** Reflect on the results obtained and possible future improvements.

3.3 Result

The optimal outcome of this practice will be a computer program ('.ipynb'), which can automatically retrieve **updated** information and use both this and historical data to predict a result.

3.4 Current Data

It is important to consider that the system must automatically collect data. This means the resulting value will depend (to some degree) on the moment of execution.
Therefore, it is necessary to search for updated sources of information. This means at least one of the following requirements must be met (the more, the better):

- The prediction is for the future (or present).
- Web scraping is performed on current data.
- The API uses current data.

3.5 Predictive Model

The predictive model does not need to be highly complex. Simple models covered in the Machine Learning course can be used, or even analytical models (following a formula instead of an algorithm) if their use is well justified.
For example, to evaluate the price of an asset, regression models or technical analysis like trends or patterns can be used.

3.6 Topic

The topic of the practice is restricted to: *Stock Market*. Within this field, any desired subtopic can be chosen.
Before starting the project, it is necessary to review the chosen topic with the professor to ensure it is appropriate and sufficient.
Some examples of topics are:

- Trend Prediction
  - Predict whether a stock value will rise or fall (or by how much).
  - Predict the stability of an asset based on news articles.
  - etc.
- Currency Prediction
  - Predict the value of a specific currency relative to another.
  - Predict the adoption of a cryptocurrency.

    - etc.
- Prediction of Organizations or Entities
    - Predict company movements.
    - Identify the most relevant topics in industrial or stock market domains.
    - etc.
- etc.

3.6.1 Example Pages

Here are some example pages you can use:

- **Web Scraping**:
    - **Yahoo Finance**: https://finance.yahoo.com/
    - **Investing**: https://www.investing.com/
    - **MarketWatch**: https://www.marketwatch.com/
    - **CoinMarketCap**: https://coinmarketcap.com/
- **API**:
    - **Crypto APIs**: https://www.coingecko.com/en/api
    - **Alpha Vantage**: https://www.alphavantage.co/
    - **marketstack**: https://marketstack.com/

These pages may be useful for their simplicity or accessibility, but any other source of information can be used depending on the chosen topic.

4 Follow-up Sessions

Throughout the course, there will be certain follow-up sessions for the practice. During these sessions, the progress of the practice, the decisions made or to be made by the team, and milestones for future sessions will be reviewed to ensure the practice is progressing correctly. These sessions will be held outside regular class hours and will not count for attendance, but they may be relevant to the final grade for the practice, in addition to being very beneficial to ensure the practice is on the right track.

5 Submission and Defense

5.1 Notebook

The practice requires **a single submission** of a *Jupyter Notebook (.ipynb)* document. This notebook must contain *Python* cells with **all** the necessary code to collect, prepare, mine and predict

the data. The notebook must also include interspersed *Markdown* cells with explanations and justifications for each step.
The notebook must be self-contained and portable, meaning it should be executable from start to finish without user interaction or dependence on any other file or folder.
The notebook must execute without errors.

5.2 Defense

Each team will defend their practice in a 15-45 minute session with the professor, justifying the decisions made and answering any questions posed.
The defense will evaluate the **justification** of the decisions made. The professor may request specific cases to be checked or minor modifications to verify the team's understanding of the steps and results.
The defense also aims to verify the authorship of the practice, ensuring that all team members have participated and understand the practice's operation.

6 Grading

6.1 Requirements

These are the minimum requirements that the practice must meet:

- **Historical Data Collection**: Historical data relevant to the chosen topic must be collected, preprocessed, and used.
- **Automated Data Collection via Web Scraping**: Data collection through web scraping techniques must be automated.
- **Automated Data Collection via API**: Data collection through APIs must be automated.
- **Predictive Model**: A Machine Learning predictive model must be used. Python libraries or very simple models can be used, provided their use is justified.
- **Current Data**: At least one requirement mentioned in section 3.4 must be met.

6.2 Scoring

This task accounts for **30%** of the final grade for the course. To achieve a passing grade, the system must function automatically with no or minimal errors.
The grade will be calculated as follows: If one requirement is not met, the maximum score will be 7. If two requirements are not met, the maximum score will be 5.
The rest of the grade will depend on the **complexity** of the task, the **justification** of decisions made, and the **quality** of the report and code submitted.

6.3 Extra Credit

The practice will be graded positively if it meets the following criteria:

- The chosen topic is relevant to current events.
- Use of text sources or textual information.
- Text analysis such as classification or sentiment analysis.
- Use of advanced Machine Learning techniques.

6.4 Teams

In general, the grade will be the same for all team members. Attendance at follow-up sessions and participation in these and the defense may affect the individual grades of team members, with a variation of up to ± 2 points.
In cases where it can be demonstrated or inferred that a team member has not contributed to the practice, the grade of that member may be significantly negatively impacted.

6.5 Academic Fraud

Any deliverable, whether code or text, must be original and created by the team members. The professor reserves the right to ask questions about any part of the practice, assuming that at least one team member is familiar with all details of the work.
If academic fraud is detected in the practice, whether in the final submission or partial submissions, the practice will receive a score of 0 for all team members. Depending on the severity, a disciplinary action may also be taken.
The use of AI tools to generate deliverable content (code, text, etc.) is not allowed.
The use of such tools for support (resolving doubts, searching for information) in the practice is allowed and encouraged.

7 Example

Below is a **short and simple** example of a topic and structure for the practice:
Study of DogeCoin Relevance

1. **Preliminary Study** Study cryptocurrencies and DogeCoin. Search for news websites about cryptocurrencies. Search for APIs providing cryptocurrency values.
2. **Historical Data Collection** Collect historical values of DogeCoin. Create a dataset.
3. **Automated Data Collection via Web Scraping** Use web scraping to count how many times DogeCoin is mentioned in cryptocurrency news articles. Add this information along with its value to the dataset.
4. **Automated Data Collection via API** Use an API to obtain the value of other cryptocurrencies (Bitcoin, Ethereum, etc.). Add this information along with its value to the dataset.

5. **Predictive Model** Train a linear regression Machine Learning model to predict the price of DogeCoin based on the number of mentions and the value of other cryptocurrencies.

6. **Model Evaluation** Use validation techniques (Train-Test split, R^2, etc.) to evaluate the reliability of the model.

7. **Conclusions** Discuss the results obtained and possible improvements.