System Requirements Specification

Project: Sentiment Analysis

Author: Damian Lebiedź

Date: April 29, 2025

Table of contents

1. Introduction

This document constitutes the System Requirements Specification (SRS) for the "*Sentiment Analysis*" project. The document defines the objectives, functional and non-functional requirements, interfaces, vocabulary, use cases, and user stories for a system conducting sentiment analysis of data using data science techniques and *the Reproducible Research* approach.

2. System Objectives

The project's goal is to create an IT system enabling sentiment analysis of entries in the r/wallstreetbets thread on Reddit during periods of highest volatility on the American stock exchange, measured by the VIX index value. To achieve this, the following should be used: text and stock market data analysis, sentiment analysis, clustering, and results visualization. The system should enable:

- Downloading and cleaning text and stock market data.
- Conducting sentiment analysis of text data.
- Performing clustering of text data.
- Visualizing results using charts and wordclouds.

3. Functional Requirements

- The system enables downloading stock market data from a specified period, cleaning it, and presenting it in charts.
- The system identifies the day with the highest volatility in the given market.
- The system enables downloading text data from a Reddit thread on the day identified as having the highest volatility in the given market and cleaning this data.
- The system performs sentiment analysis of the downloaded entries using the DistilBERT language model.
- The system performs PCA & KMeans clustering of the downloaded entries.
- The system visualizes the average sentiment of entries from a given day on a chart.
- The system visualizes the most frequent words using word clouds, both with and without division into clusters.
- The system allows the user to run the analysis on their own input data using the controller.py module, which enables changing system parameters.

4. Non-Functional Requirements

- The system should be implemented in Python and have clear, commented source code.
- All input and output data should be stored in formats that allow for the analysis to be repeated and reproduced (Reproducible Research).
- The system should be able to run on a standard personal computer with Windows, Linux, or MacOS.
- The project and documentation should be placed in a public GitHub repository.
- The system should allow for easy expansion with additional text analysis modules or data sources.

5. User Interfaces and Data Requirements

- The system is launched from a Python script or Jupyter Notebook.
- The user can specify the date range and system parameters via the controller.py module.
- Input data is downloaded automatically using the yfinance library and the Reddit API. Output data: HTML report exported from a Jupyter Notebook file.
- The user interface is limited to the parameters passed in the controller.py module.

6. Documentation Vocabulary

- **API** - Application Programming Interface; a set of rules enabling data transfer between applications.

- **Sentiment** - emotional classification of content (positive/negative/neutral).
- **VIX** - volatility index of the American stock market.
- **Wordcloud** - a graphical representation of the frequency of word occurrence in a set of texts.
- **Reproducible Research** - an approach ensuring the ability to repeat the analysis on the same data and code.
- **DistilBERT** - a machine learning model for text analysis, used for sentiment classification.
- **PCA & KMeans Clustering** - statistical methods used for grouping data.

7. Use Cases

UC1: Sentiment analysis of entries on Reddit

- Actor: User
- Description: The user runs a script that downloads entries on Reddit, analyzes their sentiment, performs clustering, and generates word clouds.

UC2: Analysis of asset volatility on the stock exchange

- Actor: User
- Description: The system downloads stock market data from a specified period, visualizes its volatility, and finds the day of greatest volatility of that asset.

UC3: Generating a wordcloud

- Actor: User
- Description: The system creates a word cloud for a given dataset in a given period, with or without division into clusters.

8. User Stories

- **As a data analyst**, I want to download and analyze entries on Reddit concerning the American stock exchange to identify changes in investor sentiment during periods of high volatility.
- **As an investor**, I want to check the correlation between investor emotions and volatility on the American stock exchange, measured, for example, by the VIX index.
- **As a member of the project team**, I want the code and documentation to be available in a GitHub repository and for the installation and launch of the project to be well described in README.md to facilitate understanding of the code and its implementation.