SDAIA Academy T5C04 Bootcamps: Data Science

Linear Regression and Web Scraping Module Project Proposal

Data Scraping: Predict the number of likes on a LinkedIn post

Date: October 6, 2021

Prepared for: Dmitry Denisov

Prepared by: Nada Rambu

Objective

The goal of this project is to develop a linear regression model that can estimate the number of likes on a given LinkedIn post. After training the model on a set of LinkedIn posts, the model will be able to generate predictions based on some information about each post.

Data Description

The data for this project will be gathered using common web scarping techniques. The obtained dataset will have 10 features, which are listed below:

- Likes - the number of "likes" the post received. This represents the target variable.
- Followers - the number of followers in the network of the user who made the post.
- Age - the age of the user who made the post
- Gender - the gender of the user who made the post
- PostType - the type of post (Photo, Video, etc.)
- PostFrequency - the average number of posts made in a single day for the person making the post.
- Hashtags – the hashtags found with the post
- Education - the highest educational level of the individual making the post
- Date - the date of the post
- Time - the time of the post

Tools

The following tools will be used to carry out the project:

1. **Beautiful Soup** and **Selenium** libraries will be used to scrape the data from LinkedIn posts.
2. **Pandas** library will be used to create data frames for easier data manipulation.
3. **Scikit-learn** library will be to implement the linear regression model.
4. **Matplotlib** library will be used to visualize and discuss the results of the analysis.