Practical Data Science
Assignment-1 Report

Name: Dhanushraj Murugan
Reg. No: F3352509

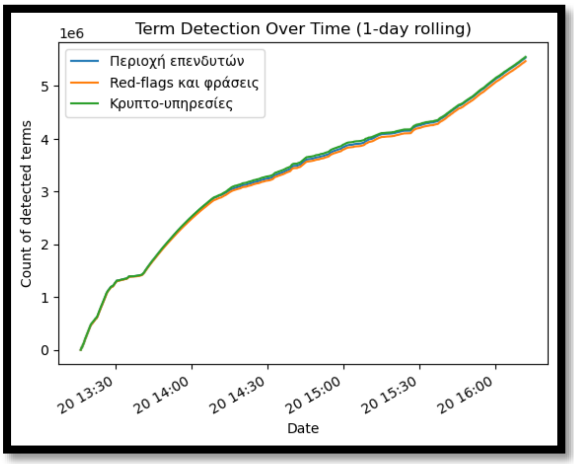
Project Title: Greek Websites Crawling and Term Detection Analysis

In the Part-A of this assignment, I have used an Asynchronous Crawler for scraping through multiple websites at a same time with a lot of workers which in turn drastically reduces the total time required for crawling. So, for this, I have used '**asyncio**' and '**aiohttp**' libraries for implementing the crawler. I have used 10 websites with no limit for the max pages to be scrapped. So, the crawler scraped totally 44K+ websites and all the scrapped text (cleaned) is stored in the .csv file at the same time immediately. So, if I interrupt the crawler at any time, the data will still be stored in the .csv file. Ethical crawling is followed strictly.

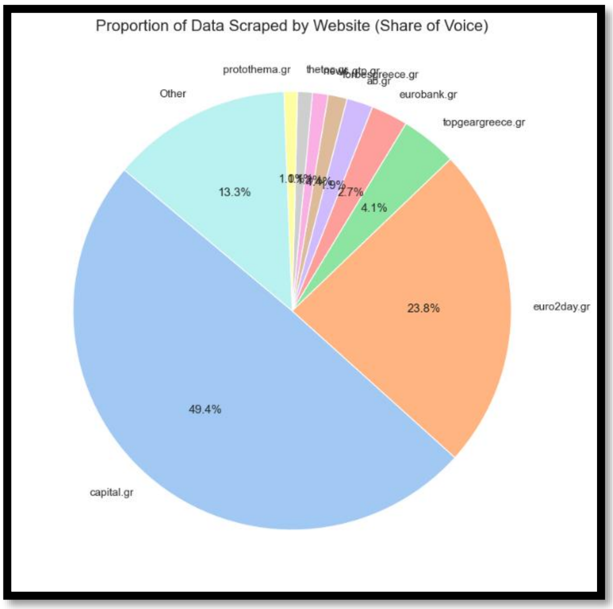
In the Part-B of this assignment, I have expanded the given terms according to the relevant words in each category. For matching the terms in file, I have used an NLP Pipeline which does a list of process like Greek Normalization, Lemmatization, Text preprocessing, Term Matching and after this performing the correction categories of these terms and saving these results into a pickle file, which will be used later for visualizations.

Using the results.pkl file, I have performed different visualizations for the results. The list of visualizations is as follows:

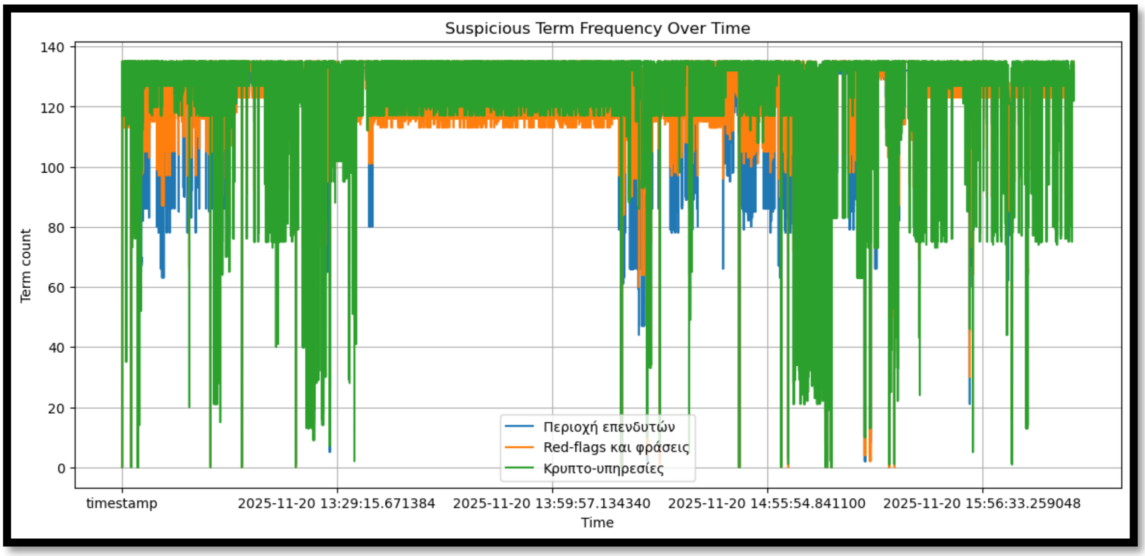
1. Time-series plot of Term Detection Over Time:



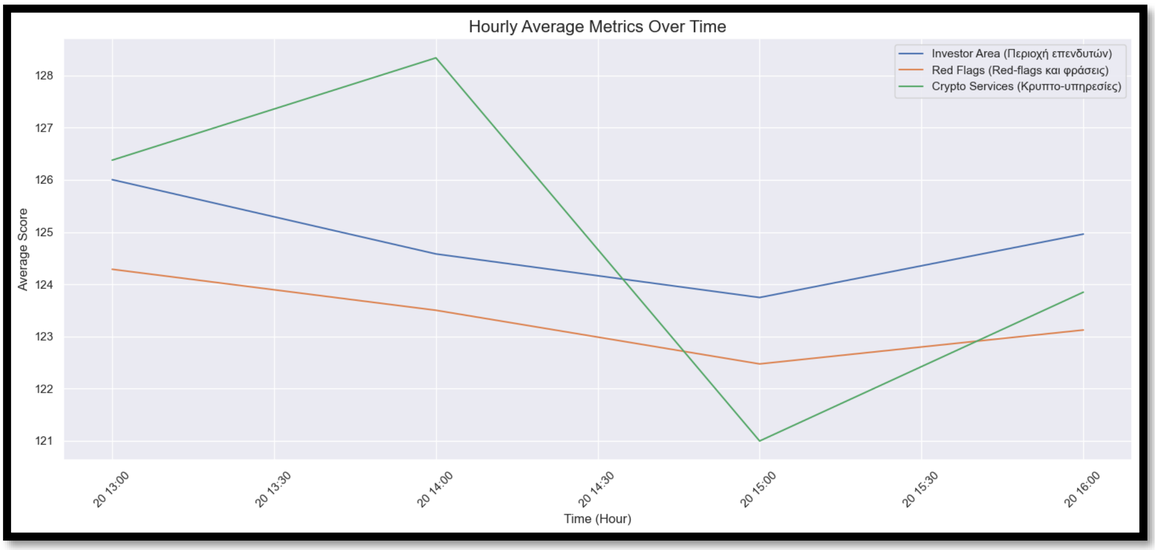
2. Proportion of Data from each website:



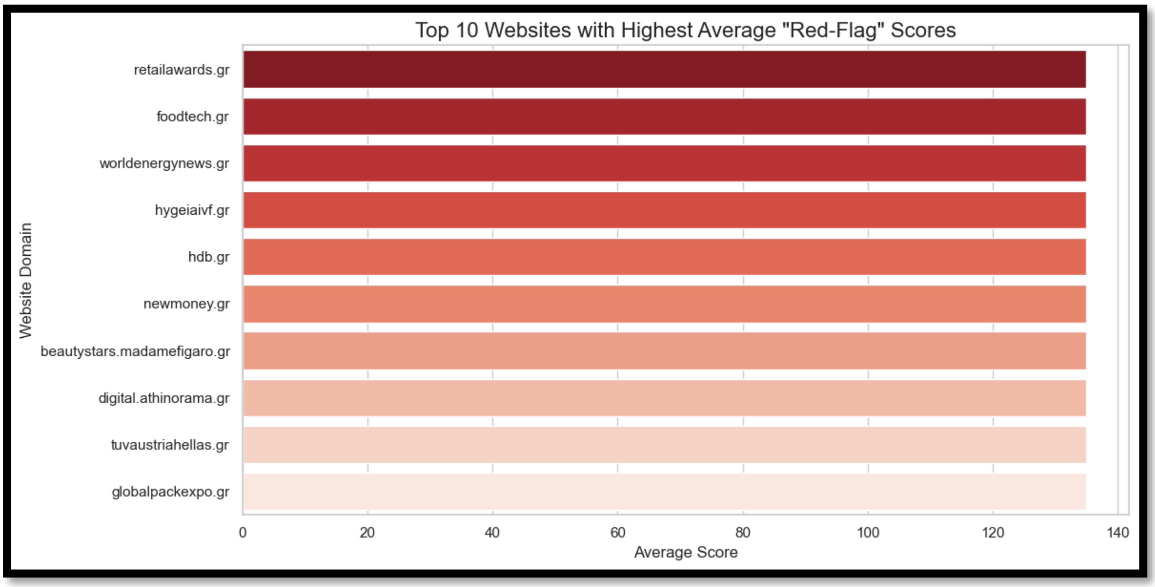
3. Term Frequency Over Time:



4. Hourly Average Metric:



5. Top 10 Websites which has more Red-Flag terms:


Top 10 Websites with Highest Average "Red-Flag" Scores
Website Domain
retailawards.gr
foodtech.gr
worldenergynews.gr
hygeiaivf.gr
hdb.gr
newmoney.gr
beautystars.madamefigaro.gr
digital.athinorama.gr
tuvaustriahellas.gr
globalpackexpo.gr
Average Score