A2. Crawling and classifying YouTube comments from Greek pages

I. Language detection

1. Build a language detector (Greek/Greeklish/English/Other) using RegExps.
2. Create a ground truth evaluation dataset and assess your classifier. (Hint: you are free to manually extract data from online sources.) This dataset will be submitted as a CSV named as **gold.csv**.

II. Crawl YouTube for videos with Greek posts

1. Apply your language detector to the page's title.
2. Parse all the comments of the page but only if the title is in Greek/Greeklish.
3. Use a strategy to jump to other pages that *will likely* have Greek/Greeklish titles.
4. Form a CSV with the crawled information, to be submitted named as **crawl.csv**.

III Improve language detection

1. Benchmark text classification (scikit) algorithms for the language detection task, outperforming your rule-based classifier and naive baselines. For evaluation, use the dataset created in the 1st step. (Hint: you are free to annotate more data or augment your training dataset otherwise.)
2. Apply your best classifier to each post to annotate mechanically the language of each comment and explore the annotated data. (Hint: use visualisations and extract insightful findings that would not be visible without your mechanical annotations.) A report named **report.pdf** should comprise these.

IV Toxicity classification

1. Use prompting to create a toxicity classifier that classifies each post from 1 (not) to 5 (toxic). Report all your prompts (from the one your started to the one you ended up with) as a PDF named **prompts.pdf**.
2. Use your prompting classifier to annotate for toxicity scale each crawled post. (Hint: you are allowed to use prompting for a sample that you will use to train a scikit classifier.) The submitted **crawl.csv** should comprise these annotations.
3. Report (in **report.pdf**): (a) the most toxic language, (b) the page with the more/highest rate of toxic posts, (c) the page where toxicity is uniform over time, (d) the page where toxicity increases over time.