SI 507

Final Project Proposal
Xi Li

Data Source: Medium / Digital Design Page / In Case You Miss It

The data sources I intend to use are pages with hot topics in digital design area in Medium (https://medium.com/topic/digital-design), which is a online articles/stories sharing platform. For each article detail page, readers can also find tags, highlights by other readers, how many people "clap" for the article, and so on.

I want to scrap the data from top 100 articles in In Case You Miss It part, where a full list of articles is listed. The goal of the project is to understand the top most salient terms in those articles' contents, and the trend in digital design area shown from those words.

Method: Crawling & Scraping

From Digital Design / In Case You Miss It, along with those article's details pages. Fields include: title, author name, author's brief introduction, releasing time, full text of the article, and tags of the article.

Challenge Score: **8**

Information Displayed:

(1) **Expecting Output: A bar chart of top-30 most salient terms' frequencies in articles' contents of Medium - Digital Design**

(2) * (optional, will have a try if possible) A multi-transform graph with x-axis as top-30 most salient terms, y-axis as each term's total frequency (there should be a scatter plot by now), and circle area with center on each dot representing the amount of articles under 5 most salient tags which include this term inside. The 5 most salient tags will be shown in different colors. (output example: https://plot.ly/python/multiple-transforms/)

Presenting Tool: Graph on Plotly

The final presenting work will be a bar chart graph (or a multi-transform graph) showing the statistical information of top-30 most salient terms in top 100 articles' contents of Medium - Digital Design.