Project Stage 2
CS839 Data Science

Team :

1. Shruti Nambiar
2. Ushmal Ramesh
3. Anuja Golechha

Web Data Sources

Our two web data sources are :

1. GSMArena - `https://www.gsmarena.com`
2. Flipkart - `https://www.flipkart.com`

GSMArena is a website that is an online repository of information about smart devices like cellphones and tablets. Users can quickly query for detailed information regarding configurations and specifications of any device they plan to buy.

Flipkart is a prominent eCommerce website in India. Their market is neck-to-neck with Amazon India, and they are popular vendors for smart devices like cellphones and tablets.For our purpose,both sources contain data about smart devices in the form of device specifications. This is what we have decided to extract

Extraction of Data

1. **Extraction from GSMArena:**
   To begin with we looked at the DOM structure of our web page. We used selenium to emulate a web user. Web elements were selected using xpath and css selectors and individual pages were crawled into. For every data instance, the page source of the respective page was extracted into a file. A series of such files were collected as input data and fed into the next stage of our pipeline. The second stage of the pipeline was an extractor program the extracts contents from the source code into a structured schema format.The device specifications were stored in a with id 'specs-list' in a tabular format. The name of the device was extracted separately.The result would be an excel sheet which contains the schema.
2. **Extraction from Flipkart:** A similar approach was followed for flipkart. Selenium was employed for crawling. DOM structure for Flipkart was different,and the changes were incorporated into our crawler. The page sources of every data instance was extracted.The extractor scanned through the page source, and based on the rules created by looking at the DOM structure, data was extracted. An excel sheet with data of all smart devices is thus prepared.

Entity Extracted

We have chosen to extract data about **smart devices** (cell phones, tablets, smart watches etc). From both websites, we extracted all possible specification information into our tables. The features that were not common between two data instances within a website were assigned "NaN"

values to ensure maximum information is retained. To bring both tables into the schema, we retained missing columns from both tables, assigning "Nan" as the default missing value.

Number of Tuples in Table1(GSMArena):3273
Number of tuples in Table 2(Flipkart):3507

Open Source Tools used

1. BeautifulSoup - `https://www.crummy.com/software/BeautifulSoup/`
   Beautiful Soup was used to parse the HTML files and extract attributes from within tags.

2. Selenium - `https://www.seleniumhq.org`
   Selenium is used to emulate a web user by automating firefox browser. To enable automation of Firefox, we used geckodriver.