CS 839 - Spring 839 - Project Stage 2

Team - Abhinav Garg, Bidyut Hota, Chirayu Garg

Data Sources

We extracted data for video games. We selected the following sources to retrieve structured data:
Source A - Gamespot (https://www.gamespot.com/)
Source B - How Long To Beat (https://howlongtobeat.com/)

Data Extraction

We extracted the structured data from the sources in two stages:
1) Crawl websites to download html pages.
2) Extract structured data from downloaded html pages.

Data Extraction from Source A (Gamespot)

We first crawled 692 pages on the gamespot website that had links to reviews for video games. Crawled pages have url like "https://www.gamespot.com/reviews/?page=1". From these 692 pages we extracted unique valid links for video game reviews. We crawled these video game review links to get links to unique video game information page. We then crawled these video game information links to download html pages with structured data. We downloaded around 9000 html pages.

We then used the downloaded html pages to extract structured data of interest using xpath in python scrapy package. We dropped the tuples which contained non-ascii characters and formatted data properly. We finally extracted 8884 tuples.

Data Extraction from Source B (How Long To Beat)

This website have no direct links to video games. However, website assigns an id to each game and thus the webpages for games can be retrieved by sending a request to url like "https://howlongtobeat.com/game.php?id=1". We sent the requests for id=1 to id=50000, which resulted in around 40,000 valid html pages being downloaded. We then realised that not all the downloaded pages are valid video games, so we pruned the downloaded pages to keep only the valid video game pages. We finally ended up with around 8800 html pages.

We then used the downloaded html pages to extract structured data of interest using xpath in python scrapy package. We dropped the tuples which contained non-ascii characters and formatted data properly. We finally extracted 8481 tuples.

Entity Type
We have extracted **Video Games** from the two sources. The two tables for two data sources have following seven attributes for each tuple.

Attribute Name	**Attribute Meaning**
Title	Name of the video game
Developer	Individuals/Studios who developed the video game
Publisher	Studio who published the video game
Platform	Platform on which video game is available. Example: XBox, PlayStation, Wii
Genre	Genre of the video game. Example: Action, Sports, Puzzle
ReleaseDate	Date on which game was released
Rating	Average user rating on a scale of 10

Tuples extracted:
Table A contains tuples extracted from Source A (Gamespot). - **8884 tuples**
Table B contains tuples extracted from Source B (How Long To Beat) - **8481 tuples**

Open Source Tools used

- python urllib module - urllib is a package that collects several modules for working with URLs. We specifically used urlopen from urllib.request for opening http urls for the purpose of website crawling.
- BeautifulSoup4 - Beautiful Soup is a python library for extracting data from html pages. We used it while crawling websites to extract links from opened http urls.
- Scrapy - Scrapy is an application which can be used to crawl website and extract structured data. We did not use scrapy to crawl website. We used scrapy to extract structured data from already crawled pages. We generated json files for the extracted structured data.
- pandas - Pandas is a python package which makes working with data easy and visually intuitive. We used pandas to explore the extracted structured data and to generate csv files.