CS 839 – Project Stage 2

Akshata Bhat, Felipe Gutierrez Barragan, Pratyush Mahapatra

1. Data Sources

We chose to extract academic paper metadata and specifically focusing on the field of Computer Vision. Our sources were as follows:

1. **arxiv.org**: It is an automated electronic archive (pre-prints) for research articles. We filtered our search by looking only for Computer Vision articles in the 2017-18 date range.
   Search Link : *https://arxiv.org/search/advanced?advanced=&terms-0-term=cs.cv&terms-0-operator=AND&terms-0-field=all&classification-physics_archives=all&classification-include_cross_list=include&date-filter_by=date_range&date-year=&date-from_date=2011&date-to_date=2018&date-date_type=submitted_date&abstracts=show&size=200&order=-announced_date_first*

2. **openaccess.thecvf.com:** It is a repository for the Open Access versions provided by Computer Vision Foundation of papers submitted to CVF sponsored conferences (CVPR and ICCV). We filtered our search for articles published in CVPR from 2014-2018
   Link *: http://openaccess.thecvf.com/CVPR<Year>.py*

2. Data Extraction

Both data sources had a uniform DOM structure for all the entities. This made our data collection relatively easy because we did not have to deal with too many exceptions. We used XPath to extract the required entities from the webpages.

- *For arxiv*, we start with the search link provided above and extracted all the entities from the resulting search. We crawl the search till we reach 10,000 entries which is the maximum limit for a search in arxiv.
  - The script that crawls arxiv is *arxivExtractor.py*.

- *For openaccess.thecvf*, we crawl through the CVPR pages of 2014-2018. The CVPR pages have information on each paper published during that year, including the citation information (i.e. bibtex) associated to each paper. We extract and store that information. Additionally, in order to get abstract and full journal reference (i.e. journal page numbers) we also crawl through the webpage corresponding to each paper.
  - The script that crawls the CVPR pager is *cvprExtractor.py*.

3. Explaining the Data

1. arxiv :
   a. Tag - Refers to the arxiv tag associated with the paper
   b. Title - Title of the academic paper
   c. Authors - Names of all the authors of the paper
   d. Month - Submitted month of the paper
   e. Year - Submitted year of the paper

f. Journal Ref - Refers to the journal in which the paper is published
g. Abstract - Contains the abstract of the paper

Number of entries : 10,000

2. cvpr :
   a. Tag - Refers to the bibtex tag associated with the paper
   b. Title - Title of the academic paper
   c. Authors - Names of all the authors of the paper
   d. Month - Accepted month of the paper
   e. Year - Accepted year of the paper
   f. Journal Ref - Refers to the journal in which the paper is published
   g. Abstract - Contains the abstract of the paper

Number of entries : 3547

4. Open Source Tools Used

We used the lxml python library (https://lxml.de/index.html) to parse the html page and then store in a tree structure which can then be accessed using XPath. We also used the Requests module to access webpages.