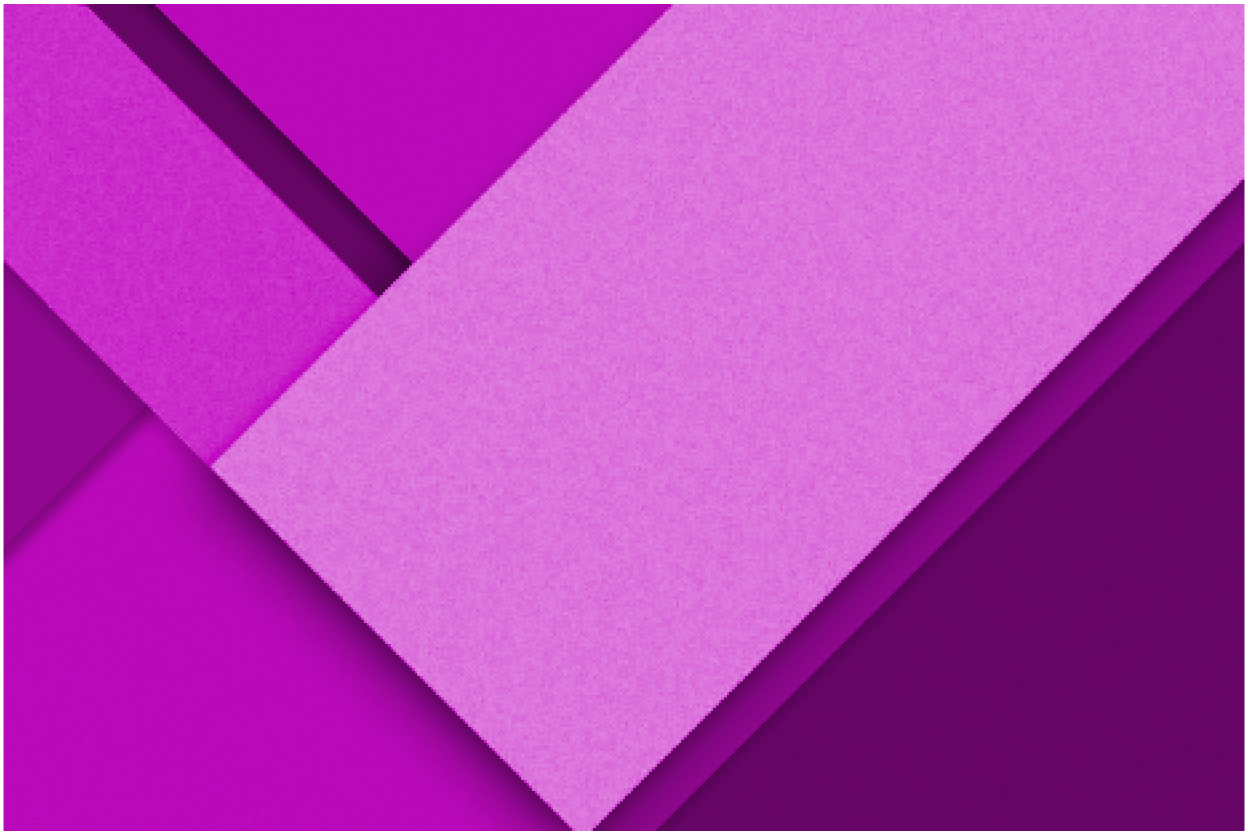

Multi Objective Dataset Generation & Validation

12.04.2019

Shashank Gupta
BTech, Information Technology (Second Year)
Vellore Institute of Technology
Vellore, Tamil Nadu

Overview

Data is fundamental for model performance. However for specific tasks, labelled data in not available. Hence there is a need for data to be annotated by experts for specific tasks. We can crawl open web data, but availability of such resources and removal of noise from such data is not trivial. Hence, there is a need to build a pipeline to remove noisy data and optimise learning over multiple objectives.

Goals

1. Pipeline to crawl image data from online then filter, segment, stitch entities for required task.
2. Text annotation data for solving multiple tasks as a common objective function.

Specifications

- Image retrieval will be of following types: -
    1. Attribute-based: It uses context and or structural metadata values. Example:
        a. Find an image file name '123' or
        b. Find images from the 17th of June 2012
    2. Textual: It uses textual information or descriptors of the image to retrieve. Example:
        a. Find images of sunsets or
        b. Find images of President of India
    3. Visual: It uses visual characteristics (color, texture, shapes) of an image. Examples:
        a. Find images whose dominant color is orange and blue or
        b. Find images by taking the example image.

- Avoid re-downloading media that was downloaded recently
- Specifying where to store the media (filesystem directory, Amazon S3 bucket, Google Cloud Storage bucket, IBM Cloud Object Storage Bucket)

(Optional)

- Convert all downloaded images to a common format (JPG) and mode (RGB)
- Check images width/height to make sure they meet a minimum constraint

Platforms

I. Python and Scrappy framework

Scrapy is an open source web scraping and crawling framework written in Python. We will be using to design the pipeline to crawl image data from online then filter, segment, stitch entities for required task. Custom Python scripts will enable text annotation data for solving multiple tasks as a common objective function.

II. Node.js or Django framework

A web application will be made for the end user to interact and give input for the data type and filters (if needed). Django framework is preferred for development of the web application due to its compatibility with Python based projects, but there is no objection with using Node.js.

Working

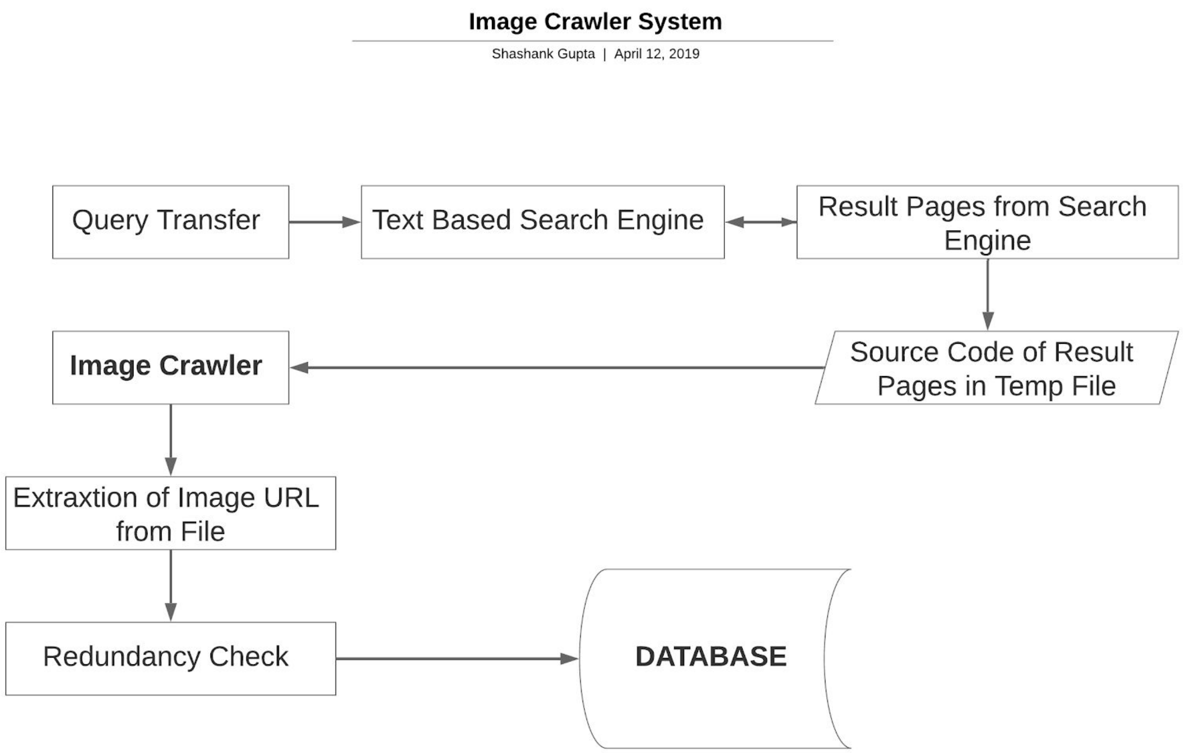
Image Crawler System
Shashank Gupta | April 12, 2019
Query Transfer
Text Based Search Engine
Result Pages from Search Engine
Source Code of Result Pages in Temp File
Image Crawler
Extraxtion of Image URL from File
Redundancy Check
DATABASE

The description of the each module of the above figure as follows:

- IMAGE CRAWLER: It is a search based tool where it requires only a keyword or phrase from the user to present the relevant images according to the user requirements.
- The tool "crawls" or "spiders" the web and then the user can browse through the search results.
- QUERY TRANSLATOR: The query is converted into the format specific to the search engine it is dealing with the object and the results are obtained in the form of an HTML page.
- TEXT BASED SEARCH ENGINE: The tool requires only a keyword or phrase from the user to present the relevant images according to the user requirements.
- REDUNDANCY CHECKER: Extraction of different urls leads us to the same content. As the check needs to be fast, all URLs are kept in memory, and are comparing character by character quickly
- DATABASE: These results are entered into a database sheet with the key as the url and the corresponding disk path.

Contact

Shashank Gupta

Email 1: shashankgupta2611@gmail.com

Email 2: priyamgupta165@gmail.com

Phone: +91 8084915252

LinkedIn: www.linkedin.com/in/shashankgupta2611

GitHub: github.com/ShshnkGpta