Data acquisition, extraction, and storage

Pierre Senellart

20 December 2024

This exam lasts three hours. The only documents allowed are ten A4 sheets (both sides), with the content of your choice. Communicating devices are strictly forbidden. When writing code, imprecision in the syntax of the languages will be tolerated. The exam is formed of a single problem, graded out of 20 points.

A dataset of images

You are tasked with creating a dataset of images from the Web in order to train an object detection system. For its image we need to collect:

- the original image data in whatever format it was available;
- the original filename;
- textual meta-data about this image (description, surrounding text, etc.);
- information about the license of the image, when available;
- original source of the image (URL of the image, URL of the page containing the image, Web site).

We assume we are given a list of thousands of Web domains from which to retrieve such images (we are not interested in any other contents). Images may be marked within an HTML page by a fragment such as:

```
<div class="image">
  <p>Hippos are <em>fascinating</em> animals with humongous
  mouths, as illustrated by the following picture
  of a hippo eating a watermelon:</p>
  <div><img alt="Hippo with broad open mouth where
    a half-watermelon is deposited by two hands" src="hippowm.jpg"></div>
  <p class="license"><small>
    <a rel="license" href="https://creativecommons.org/licenses/by-nc/4.0/">
      CC-BY-NC 4.0</a>, avdeevaeee, GoodFon.
  </small></p>
</div>
```

which may be rendered as:

Hippos are *fascinating* animals with humongous mouths, as illustrated by the following picture of a hippo eating a watermelon:

CC-BY-NC 4.0, avdeevaeee, GoodFon.

This is just an example, it is not expected that all or even most images from the domains we are crawling are coded in HTML in exactly the same way.

1. (2 points) Compared to a standard scrapy crawler such as the one we built during the hands-on session on mathematical genealogy data, what changes need to be made to crawl this dataset? In other words, without writing any code, but highlighting the specificities of this particular task, outline the general design of the crawler: what are the seed URLs? where to find links and which links to crawl? what to do with crawled data?

2. (1 point) If there are roughly 5 000 Web sites to crawl, each contains roughly 100 000 pages and 20 000 images and each image data (metadata and everything included) takes on average 200 kB, what is the expected size of the entire dataset?

3. (1 point) Give an order of magnitude of the time required to crawl the entire dataset, with the objective of accelerating its acquisition while following crawl ethics.

4. (1.5 points) If the image is encoded exactly as in the example given on the previous page and several such images may be present on a given page, give XPath selectors for each of the metadata of interest: filename, textual meta-data, license URL.

5. (2 points) In order to build the final dataset, we need to convert the image data to a standardized format and resolution (PNG, and maximum image size of 640×480). Assume your data is stored on a distributed filesystem on a cluster of machines. Provide a MapReduce or Spark program that processes all images to produce such standardized format. You can assume the existence of a library function for image conversion to a specific format and resolution and make any reasonable assumption on the input format. Give an order of magnitude of the runtime of the approach, based on the fact that the library function takes 0.1 s including disk I/O costs, depending on the number N of machines in the distributed system.

6. (1 points) Is it feasible to store the entire data in a regular DBMS? What about the metadata (which excludes original or reformatted image data)? Justify.

7. (2 point) Propose a relational schema for the metadata. Assuming such a schema, give a SQL query that lists all CC-BY or CC-BY-NC images whose meta-data indicate they are about hippos.

8. (2 points) For GDPR-related reasons we need to be able to delete any image containing personal information upon request. We want an approach that allows us to process such requests while being able to maintain the list of images satisfying some conditions (such as those of the last question), without having to recompute these lists. Outline a possible approach.

9. (3 points) We now need to build a ground truth annotation of (part of) the image dataset, each annotation indicating that a specific object (say, a watermelon, a hippo) is present in the image within a specific region. Propose a general approach to do this, assuming the availability of human annotators who can be paid for this annotation task. Consider the fact that human annotators may make mistake. What interface to design? What questions to ask human annotators? How to combine their annotations?

10. (2 points) Discuss legal aspects arising from the acquisition and eventual use of this dataset.

11. (2.5 points) Once the image detection model is built, it is expected users can have access to it through a RESTful API where the user provides the URL of an image from which to extract objects and the server processes the image with the image detection model and serves as a JSON or XML file a description of all objects identified in the image. Give an example of such interaction on the example of the image of a hippo eating a watermelon: What is the form of the request made by the user? What is the full content of the response returned by the server? *You need to provide the entire content of the input and output for this specific example, but you do not need to provide the code of the function implementing the service.*