

Information Retrieval and Web Search - Project Phase 1

Lukas Pfahler

Tejas Umakanth

February 8, 2014

1 Collaboration Details

We both do everything. clever...asdfa sckfjalöksdj fölakdjf öaljdf,a,jhdf a abdf afas dasfdghalkjd-hfa

adsfadsf

sfg asd

sdfg

2 Description



For our project, we decided to crawl and index data found on Instagram¹. Instagram is a popular² social media application that allows users to publish photos and videos and search the media published by other users. Each user is identified by a unique user name. Each media item can have a caption, a list of comments by other users or a location. As known from other popular social networks like Twitter and Facebook, captions and comments can contain hashtags.

¹ instagram.com

² Instagram reported 150 million users in September, 2013.

3 Crawling

Instagram offers a developer API that allows us to subscribe to real time updates and crawl new media items as soon as they are posted.³ All communication between the crawler and instagram is done using the HTTP protocol: The crawler is itself a http-server and once we subscribe to media updates at [instagram.com](https://www.instagram.com/developer/realtime/), their servers start connecting to our crawler. Everytime there is new media, they issue a http POST request. The post does not contain the actual media, but is merely a notification that there has been an update. We then grab the actual data by requesting all recently added media from their servers in a http GET request.



The response is a JSON file containing a list of media items. We save each of those items in a separate JSON file. We also save a thumbnail image for each media item.

Our crawler is based on the software platform Node.js, which allows us to setup a http-server and request data in an asynchronous fashion with very little javascript code.

Node.js is a platform built on Chrome's JavaScript runtime. [...] It uses an event-driven, non-blocking I/O model that makes it lightweight and efficient, perfect for data-intensive real-time applications that run across distributed devices.⁴

Instagram allows developers to issue 5000 requests per hour⁵, thus we introduced a politeness factor p : Only every p -th time our crawler is notified we actually request the new data. This has two effects: First, the number of requests is reduced and second, the number of new media items per response is higher. It also means that we probably miss a fraction of the data. However, if one would setup a larger system with multiple crawlers the coverage

³[http://instagram.com/developer/realtime/](https://www.instagram.com/developer/realtime/)

⁴<http://nodejs.org/>

⁵[http://instagram.com/developer/endpoints/](https://www.instagram.com/developer/endpoints/)

would increase.

4 Indexing