Information Retrieval and Web Search - Project Phase 1

Lukas Pfahler Tejas Umakanth

February 7, 2014

1 Collaboration Detail

2 Description

For our project, we decided to crawl and index data found on Instagram[1]. Instagram is a popular[2] 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.

3 Crawling

The crawler is itself a webserver. Once we register a subscription at instagram.com, their servers start connecting to our crawler. Everytime there is new media, they issue a http POST request to our server. 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 using a http GET request.

[1] `instagram.com`

[2] 150 million users reported in September, 2013

The response is a JSON file containing a number of media items.

using node.js[3]

4 Indexing

[3] http://nodejs.org/