10/5/2010

# Data Mining Digital Media

## Team Members

Mackenzie Boyd

Brian Cullinan

Mike Austin

## Project Description

We will provide a simple search engine for media, such as movies, music, pictures. We would initially focus on movies and video media and if we were highly successful, continue on with music and pictures.

This search engine would show results for related genres and use related words to search the database cache of the video media. The search results will be displayed as a jQuery list that continuously updates as the database is updated.

## Data Source

Our data sources would be RSS feeds and sites that are regularly updated with new data. Likely data sources include, but are not limited to:

* Reddit.com
* Youtube.com
* Soundcloud.com
* Hulu.com

Data will also be scraped from HTML pages, retrieving media titles and descriptions and other related words to categorize the media files. The actual video and audio content will remain hosted on the service it was obtained from.

## Methodologies

Media will be categorized by file-type and file-extension initially. Then the header and content-type will be analyzed. Finally, HTML pages the media comes from will be analyzed for context information such as titles and descriptions of the media. Regular expression and XML parsing will play a large role in categorizing and caching the data.

## Milestones

* Script to find media in RSS feeds
* Scripts to update the database
* Cron job integration
* jQuery UI front end similar to <http://www.google.com/realtime>

## Refrences

* <http://www.google.com/landing/realtime/>
* <http://docs.jquery.com/Main_Page>
* <http://en.wikipedia.org/wiki/RSS>
* <http://en.wikipedia.org/wiki/Regular_expression>

## Example HTTP Header:

HTTP/1.1 200 OK

Date: Mon, 23 May 2005 22:38:34 GMT

Server: Apache/1.3.3.7 (Unix) (Red-Hat/Linux)

Last-Modified: Wed, 08 Jan 2003 23:11:55 GMT

Etag: "3f80f-1b6-3e1cb03b"

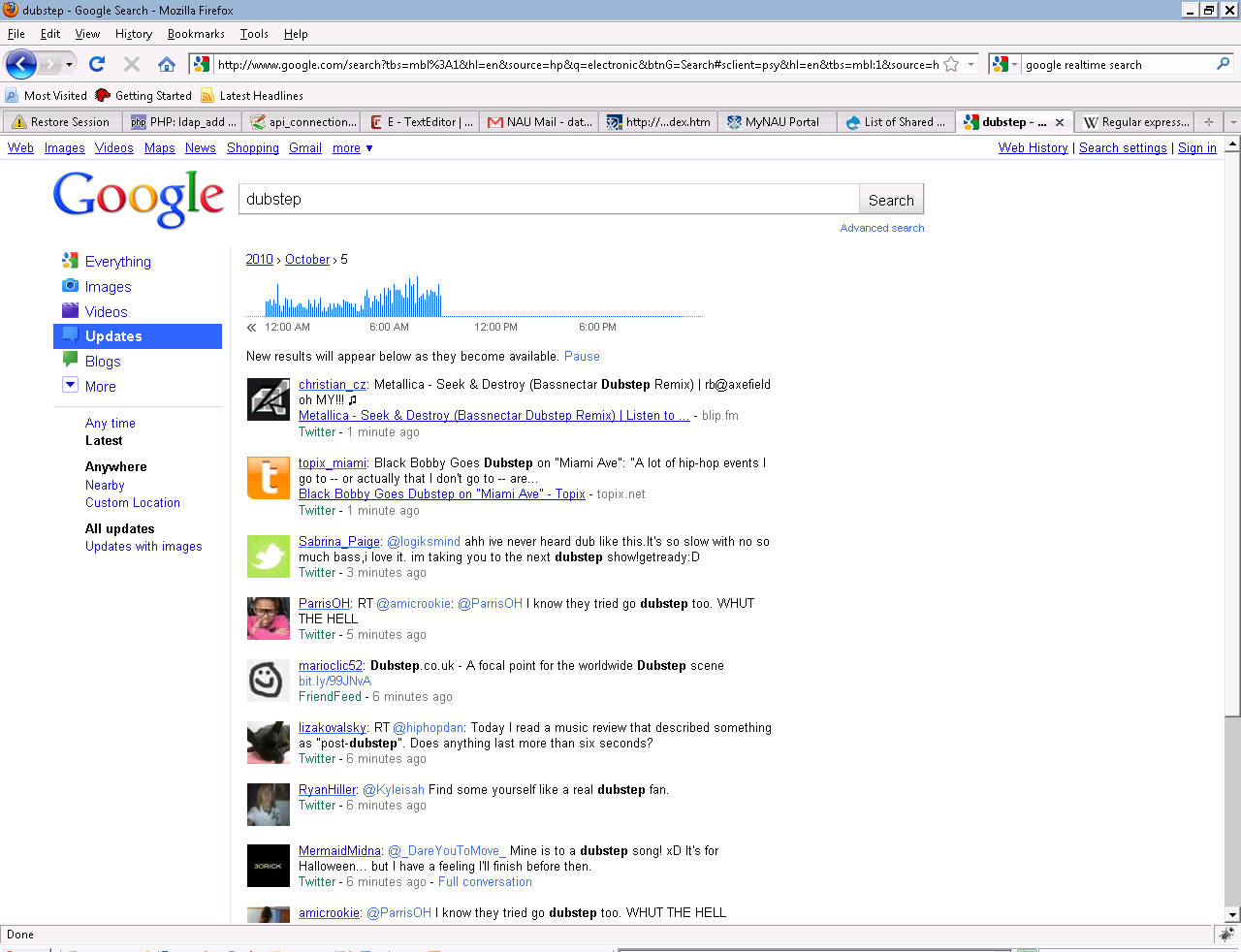
Accept-Ranges: bytes

Content-Length: 438

Connection: close

Content-Type: text/html; charset=UTF-8

## Google Real-time Search:



## Reddit Music Feed:

