Overall Summary

YiLin Liu

March 30, 2017

Objective:

To explore various APIs (facebook, google maps, twitter, flickr and weather data) through R and explore their usefulness to a data scientist.

Summary:

An API is a set of protocols, tools and routines used for software building. APIs specify how components should interact with each other. In regards to this report. An API specifies the protocol used to extract data from the various data sources. The data sources investigated includes facebook, google maps, twitter flickr and weather data. A ranking from the most useful to least useful for data scientists is produced by exploring the APIs, the packages provided in R and relevance to potential projects.

1. Based on the capabilities of google maps API as well as the usefulness to current projects, google maps API is ranked first. The API provides route information between 2 points, geocoding services for points, elevation and more. Given that one of the current projects at DAC involves optimizing routing for ambulances in Ottawa, the accurate routing information as well as traffic modeling that the google maps directions API provides is of essential importance. Another potential project at DAC aims to predict the next calls for a delivery service based in USA. The ability of the google maps geocoding API to standardize differently formatted addresses as well as providing specific geocoding information is useful for future analysis of data.
2. Twitter API ranks 2nd due to the majority of data being publicly available. Users as well as tweets are typically publicly available through the API. This means that sentiment analysis can be performed on existing tweets returned by the search API. New tweets and updates can be monitored using the stream API and potentially stored in a database for long term usage. Friendship data is publicly available through searching for users' followers and friends,which can lead to social network analysis. A major downfall with the twitter API is that the search API only returns tweets from up till 1 week ago, however there is no time limit to tweets searched via timelines of users.
3. Flickr API ranks 3rd due to the details of the data. Flickr is an photo hosting/sharing website meant for professionals. The data (photos) are not readily available for public access due to privacy settings, but the public search API offers fine control in terms of locating the desired response. The metadata for the photos are typically available which is what differentiates Flickr from many other data sources where the metadata is stripped to conserve space.
4. The weakest API is the facebook API. The facebook API provides massive amounts of data via the ability to search for public groups, pages, posts, comments and replies to comments. The downfall is that user information is unavailable with the new API unless the proper permissions have been obtained, meaning it would be hard to categorize and investigate the data.
5. The exploration for weather data was done through web scrapping and not through an official API. It was shown that parsing HTML source code is possible through R and be analysed. The framework can be useful for future web scrapping needs.

Details:

Overview:

All APIs explored relies on sending an **HTTP GET request** with a OAuth token to access the API data. The response recieved is typically in JSON or XML format. The GET and OAuth token is generated based on the package httr, while jsonlite and xml2 packages can used to parse JSON and XML responses respectively. Note, jsonlite and xml2 does not support the parsing of multiple responses at once.

In order to obtain an OAuth token for a particular API, an application for the API needs to be created. To do so, a test account and a test app has been made for each API. The details are within the report for each API.

The following includes a brief summary of the advantages and disadvantages of each APIs, with links to a more detailed report of the work that has been done on each API.

Facebook:

- Facebook API
- Rfacebook

- **Advantages:**
  - Easy to retrieve data from facebook API using the functions provided by the **Rfacebook**.
  - Easy to access public data of facebook through its **search API**. Part of the functionality is covered by the *Rfacebook* package through the *searchGroup & searchPages* function.
  - Discussions about a post can be easily tracked through extracting comments and replies to those comments.
  - The graph API Explorer tool is useful for testing API calls to understand structure as well as exploring possible fields and edges.

- **Disadvantages:**
  - The facebook API documentation does not provide specific information regarding **Rate Limit**, it specifies that the limit `differs depending on request`.
  - Facebook API V2.0 and above place heavy restrictions on accessing personal information. Details such as profile information, user timeline etc. are not generally publicly available due to privacy settings. It is **not recommended** to scrape the facebook API to search for user related information. This means given an user Id, the only parameters accessible would likely be the user name and no more.
  - To retrieve information regarding nodes, the Id of the node is required. This makes tracking objects/searching for specific nodes very difficult since there is no easy way to obtain the id for specific objects of interest (pages, users, comments, etc.).
  - It was observed that many pages, posts and comments does not include *location* information as part of its content.
  - Custom HTTP requests need to be sent for information not supported by the package.

Facebook Detailed Report

Facebook Report

Twitter:

- REST API
- Stream API
- Search API
- twitteR Package
- streamR Package

- **Advantages:**
  - Most information posted on twitter is public (tweets, users, etc) and can be accessed through the **REST API**.
  - The **search API** for Twitter supports more powerful queries for searching public twitter content.
  - It is possbile to extract the tweets posted any public user.
    - By accessing friendship and friends API endpoints, social network information can be retrieved for most users.
  - The **twitteR Package** provide easy access to the Twitter REST API and handles the sorting of retrieved data through class wrappers that provide convenient functions to access information. Read the documentation for the package for more details.
  - **TwitteR package** retrieves majority of information that is of interest, if not supported by the twitteR, any arbitrary HTTP request can be made to the API to retrieve information.
  - TwitteR establishes connection to data base if requested, which can save searched data easily.
  - Twitter provides a live **stream API** which allows the retrieval of the new updated information on tweeter. This is a unique feature and can be used for real-time tracking for events of interest.
  - **StreamR Package** provides easy access and convenient use of the **stream API**. Tasks such as connecting to the API, reconnecting and saving to File is all handled.

- **Disadvantages:**
  - Location information is available but often users can fill whatever it desired, thus not most reliable.
  - The token produced by **twitteR** is cached internally and does not provide easy access to the token for use outside of **twitteR** functions.
  - Due to the nature of twitter, it is hard to track how users react to tweets posted by other users. This is because the replies are not directly connected to the post.
  - Can not access the feed of other users if application has not been granted access, ie what tweets appear on their account.

  - **It is not possible to public search for old tweets (older than a week) through the official twitter search API.** For access to historical search data, a 3rd party service needs to be paid for. Gnip seems to provide that service.
  - **TwitteR Package** does not support the extraction of **entities** information within a tweet, (ie: media url, hashtags and urls). This is because the responses for entities varies depending on the specific tweet. In order to extract this information, an HTTP request needs to be send and the response needs to be manually parsed.

Twitter Report:

Twitter Report

Google Maps API:

Google Maps API

- **Advantages:**
  - Provides many services including directions between origin and destination, geocoding services, matching points to near roads, elevation, etc.
  - Documentation is clear and very easily understood.
  - Use basic **HTTP request** with simple format, making it easy to use.
- **Disadvantages:**
  - limited free requests per day. The free limit is enough for investigation purposes. If used to process large amounts of data, a fee will need to be payed.

Google Maps Report:

Google Maps Report

Quick Word on OpenStreetMap(OSM):

- Open Street Map Wiki
- Preliminary Report on OSM

Weather Web Data:

An basic script was made to scrape the Environment Canada Ottawa website for hourly updated temperature and 24 hour forecast. The collected data was then explored by making several graphs.

- Weather Web Scraping
- Data Exploration

Instagram:

- From initial searching, the Instagram API is not publicly available. In order to access the API, an Instagram approved application (only applications that promote sharing through Instagram can be approved). Since the API is not available, not further search into Instagram was done. Instagram Permissions Review.

Flickr:

- Flicker is a photo sharing website which allows people to host their photos online.
- Flickr API
- **The API is not maintained or supported by Flickr.**

- **Advantages:**
  - Allows public access with relative ease.
  - Many methods does not require OAuth 1.0 token.
  - Photos usually contain meta data.
  - Photo tags provide an easy way to adjust search query.
- **Disadvantages:**
  - Does not have massive amounts of data available like twitter. This is because many photos are not public.
  - Limited requests.

Flickr Report

Others:

- Exif Tool Exploration: This is a windows application based on the Exif PERL library which allows the extraction of metadata for images.
- Prophet Package Exploration: This is a facebook package that is used to predict trends for time series data, which accounts for yearly variations and seasonal variations.