Social Media Analytics

By Pratap Malladi

Text Data

- Content
- Tweets, Posts, Blogs, messages, comments
- Attributes
- Creator –name, demo graphics, geo
- Date and time
- Hash tags, URLs, references
- Consumers –retweets, likes, shares

Media Data

- Photos, video, live feeds etc.
- Attributes
- Shares / likes

Connections Data

- Following/ followed
- Friends and acquaintances
- Business circles
- Shares / likes
- Attributes of people

“

Applications for Social Media

Customer Management

- Customer opinions are very important to business
- They today express their sentiments through social media
- Products & Services
- Customer Experience
- Contact Centers have evolved to include Social Media experience too.
- Contact unhappy customers to reduce attrition
- Contact interested customers to sell products

Marketing

- Product launches have expanded from Print / Television media to social media
- Social media helps marketers get instant feedback on their messaging.
- Companies identify key persons interested in their products to reach out further

News Media

- Media today understands public sentiment on events from social media
- World happenings
- Elections
- Sports
- This is possible today in real time
- Entertainment media tracks celebrities through social media.

Social Media applications

- Mine social media in real time / historical mode to extract text, connections and media
- Understand sentiments and networks
- Identify key actors/ contacts
- Integrate with other internal applications to create customer 360.

REST APIs

- Almost all Social Media Websites provide REST APIs
- Content is usually JSON
- Authentication and authorization through OAuth
- Developer support and documentation
- Rate limits for free access
- https://apigee.com/console

Challenges with Social Media Analytics.

Unstructured Data

- Most social media is text.
- Text may or may not contain relevant information all the time.
- Filtering data depends upon hashtags and references
- Multi-lingual
- Language used contains a lot of non-standard words/phrases

- **Incomplete & Dirty Data**

- All persons/ companies on the social web do not share all the information
- Information is limited by security/privacy constraints.
- APIs have additional security limits on whose and what data they can access.
- Additional logic required to identify and handle permission issues
- Information available may not confirm to expected formats
- Names
- Dates
- Copyright

Rate Limits

- All public APIs have rate limits
- number of queries
- Size of results
- Developers need to get creative about how to use the available bandwidth in a smart manner
- Pacing queries
- Caching and archiving
- Becomes tough during development phase.
- Use cached / saved data as much as possible
- Data feed simulators

REST API

- **REST**
- Almost all Social media APIs use REST.
- Representational State Transfer. A method of exchanging information between a client and a server
- Uniform interface
- Stateless
- Client server
- Cacheable
- Usually over HTTP
- CRUD Operations (Create, Read, Update, Delete) on resources
- GET, POST, DELETE, PUT as methods
- Resources and actions identified using URIs

`

- **Sample REST Query**
- •Request

- **GET https://api.linkedin.com/v1/people/~?oauth2_access_token=AQVPzkNJsvg1zcsiGGL-b7cCoXgZtBAVPcjBwTlBWxjh_5jOYf_V6LuLxrEkUd0fzasfta8tStSoP5vXPB5GoDywO2BVsg-c1XCB2DkGw9yFZDW3zsK4ZTLKHzL2_T21Sl_CUwoBirI1FllMsEwo-q1GX6pgNvAn2lJunui8RMUsltkTrDk&format=json HTTP/1.1**
- •Response

- **{**
- **"firstName": "Kumaran",**
- **"headline": "Data Science / Analytics Leader",**
- **"id": "xJJOxZaSwN",**
- **"lastName": "Ponnambalam",**
- **"siteStandardProfileRequest": {**
- **"url": "https://www.linkedin.com/profile/view?id=AAoAAADdS1kBecL8JAojUfnbkvm8jGdbj1raItI&authType=name&authToken=Zz7C&trk=api*a3227641*s3301901**

OAuth

Overview

- Open Authorization protocol
- Enables applications to obtain authorized access to other's data
- No need to share passwords with applications /developers
- Supports web, desktop and mobile applications
- Almost all social websites and cloud services use Oauth
- Social media like Twitter, Facebook, LinkedIn, Google, Github
- Cloud apps. like Salesforce, Amazon, Paypal

Roles

- Resource Owner
- Who owns the data (e.g: twitter user)
- Provides authorization for applications to access data
- Authorization Server
- Manages authentication and authorization (e.g. facebook)
- Resource Server
- Provides data (e.g. facebook)
- Can be different from authorization Server
- Client
- Application who needs data
- Get authorization keys from the owner, authenticates through authorization server and access data from resource server.

General workflow

- Owner creates an “application” on the Authorization Server
- Authorization server generates access keys
- Consumer Key/ Consumer Secret /OAuthToken/ OAuthSecret
- API Key
- Owner provides access keys to the client developer
- Developer builds the client to use access keys
- Client authenticates/authorizes with authorization server
- Authorization server issues access token with set timeouts
- Client uses access token to access resources on resource server
- Resource server validates access token with authorization server and provides data to the client.

Linking Data

- Social Media data needs to be linked with other data to obtain between insights
- Customer databases
- Marketing databases
- Other social media
- Requires linking different contact handles
- Twitter handle, Facebook ID, email ID, phone number
- Person API helps provide cross-linking

Twitter Data Mining

- **What is Twitter?**
- A micro-blogging site that allows users to publish their events, comments, likes and dislikes
- 140 character tweets
- Ability to see other's tweets without permissions
- Shares and retweets
- Follow interesting persons and entities
- Asymmetric relationships –you don't need permissions to follow someone

Twitter Data

- Users and timelines
- Tweets
- 140 characters
- User mentions (@)
- Hashtags
- URLs
- Media
- retweets
- Timelines
- Friends and followers
- Direct messages
- Lists and favorites

- Twitter API

Twitter REST API

- https://dev.twitter.com/rest/public
- REST
- Oauth
- JSON
- Searches
- GET, POST and UPDATES
- Rate Limits

Work Flow

- Create an application at https://apps.twitter.com
- Application settings
- Keys and Access tokens
- Consumer Key
- Consumer Secret
- Access Token
- Access Token Secret
- Permissions
- Pretty open
- Access self and other user’s messages and followers

Facebook Data Mining

- **What is Facebook?**
- An online social network service
- Connects people
- Symmetric relationships
- Share messages and media
- Like / unlike / comment
- Create circles like friends, family and acquaintances

Facebook Data

- Users
- Posts
- User mentions (@)
- Hash tags
- URLs
- Media
- Likes, comments and shares
- Timelines
- Friends and groups
- Chat
- Events

- **What is Google+?**
- An online social network service
- Connects people
- Symmetric relationships
- Share messages and media
- Plus One, comments
- Create circles like friends, family and acquaintances

Google+ Data

- People
- Linked to a Google account
- Attributes
- Activities
- Posts and shares
- Plus ones
- Comments
- Associated with activities and people
- Moments

Google+ API

- **Google+ API**
- https ://console.developers.google.com/apis/api/plus/overview
- Oauthand simple API key
- JSON
- Searches
- Privacy –pretty open compared to Facebook
- Rate Limits –less strict
- Documentation : https://developers.google.com/+/web/api/rest/

Work Flow

- Create an application at https://console.developers.google.com/apis
- Enable Google+ API
- Create an API key
- Use the API key in your application

Introduction to Use Cases

Use Cases

- Use cases are based mostly on Twitter and Google+ data
- Privacy and security issue limit who's data we can mine for examples
- Might find it repetitive since the same steps are involved
- Focus on getting data into local data structures.
- Then regular data mining techniques can be used

Machine Learning Overview

- Data contains attributes
- Attributes show relationships (correlation) between entities
- Learning –understanding relationships between entities
- Machine Learning –a computer analyzing the data and learning about relationships
- Machine Learning results in a model built using the data
- Models can be used for grouping and prediction

Data for machine learning

- Machines only understand numbers
- Text Data need to be converted to equivalent numerical representations for ML algorithms to work.
- Number representation
- (Excellent, Good, Bad can be converted to 1,2,3)
- Boolean variables
- 3 new Indicator variables called Rating-Excellent, Rating-Good, Rating-Bad with values 0/1
- Document Term matrix

Unsupervised Learning

- Finding hidden structure / similarity / grouping in data
- Observations grouped based on similarity exhibited by entities
- Similarity between entities could be by
- Distance between values
- Presence / Absence
- Types
- Clustering
- Association Rules Mining
- Collaborative Filtering

Supervised Learning

- Trying to predict unknown data attributes (outcomes) based on known attributes (predictors) for an entity
- Model built based on training data (past data) where outcomes and predictors are known
- Model used to predict future outcomes
- Types
- Regression (continuous outcome values)
- Classification (outcome classes)

Training and Testing Data

- Historical Data contains both predictors and outcomes
- Split as training and testing data
- Training data iused to build the model
- Testing data is used to test the model
- Apply model on testing data
- Predict the outcome
- Compare the outcome with the actual value
- Measure accuracy
- Training and Test fit best practices
- 70-30 split
- Random selection of records. Should maintain data spread in both datasets

Understanding how ML algorithms work

- ML Algorithms work with
- numbers (continuous data)
- classes (discrete/ categorical data)
- ML algorithms don’t work with text.
- All textual data need to be converted into numbers or classes
- This is one of the main responsibilities of data pre-processing

Text Pre-Processing

Text Cleansing

- Remove punctuation
- Remove white space
- Convert to lower case
- Remove numbers
- Remove stop words
- Stemming
- Remove other commonly used words

TF-IDF Overview

- Text Documents are becoming inputs to ML more and more.
- News items for classification
- Email messages for spam detection
- Text search
- Text need to be converted to equivalent numeric representation before ML can be used
- The most popular technique used is Term Frequency – Inverse Document Frequency (TF-IDF)
- TF-IDF output is table where rows represent documents and columns represent words
- Each cell provides a count / value that indicate the “strength“ of the word with respect to the document

TF-IDF formulae

- Text Frequency (given a word w1 and Document d1)
- = (# of times w1 occurs in d1) / (# of words in d1)
- Inverse Document Frequency (given a word w1)
- = log e (Total # of docs / Total docs with w1)
- TF-IDF = TF * IDF

TF-IDF steps

- 1.Original documents

- Doc 1 = “ This is a sampling of good words”
- Doc 2 = “ He said again and again the same word after word”
- Doc 3 = “ words can really hurt”

- 2.After cleansing

- Doc 1 = “sample good word”
- Doc 2 = “again again same word word”
- Doc 3 = “ word real hurt”

TF-IDF (contd.)

- Creating the count table

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
Doc 1	1	1	1				
Doc 2			2	2	1		
Doc 3			1			1	1

- Finding Text Frequency

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
Doc 1	.33	.33	.33				
Doc 2			.4	.4	.2		
Doc 3			.33			.33	.33

TF-IDF (contd.)

- Finding Inverse Document Frequency
  - Log e (Total docs / docs with the word)

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
IDF	1.098	1.098	0	1.098	1.098	1.098	1.098

- Finding TF-IDF (TF * IDF)

Document	*sample*	*good*	*word*	*again*	*same*	*real*	*hurt*
Doc 1	.36	.36	0				
Doc 2			0	.44	.22		
Doc 3			0			.36	.36

Linking Data

- Social media data by itself has limited use
- Linking required with CRM / customer /marketing databases to obtain bigger value
  - Unsatisfied customer – who products he brought? When? Are there any open tickets?
  - Prospective customer – have we reached out to him before?
- Link twitter/ Facebook handles with email / phone number
  - FullContact Person API