Leveraging ML for Analysis

Outline

- Classifying Input
- Features, feature extraction
- Training
- Evaluation

Types of ML

- Machine Learning (ML) is a computational approach to classifying or labeling types of input
- Two broad approaches
  - Supervised
    - The learning is based on a training set of data that has been labeled in advance (often my hand)
  - Unsupervised
    - Learning is inferred from unlabeled data

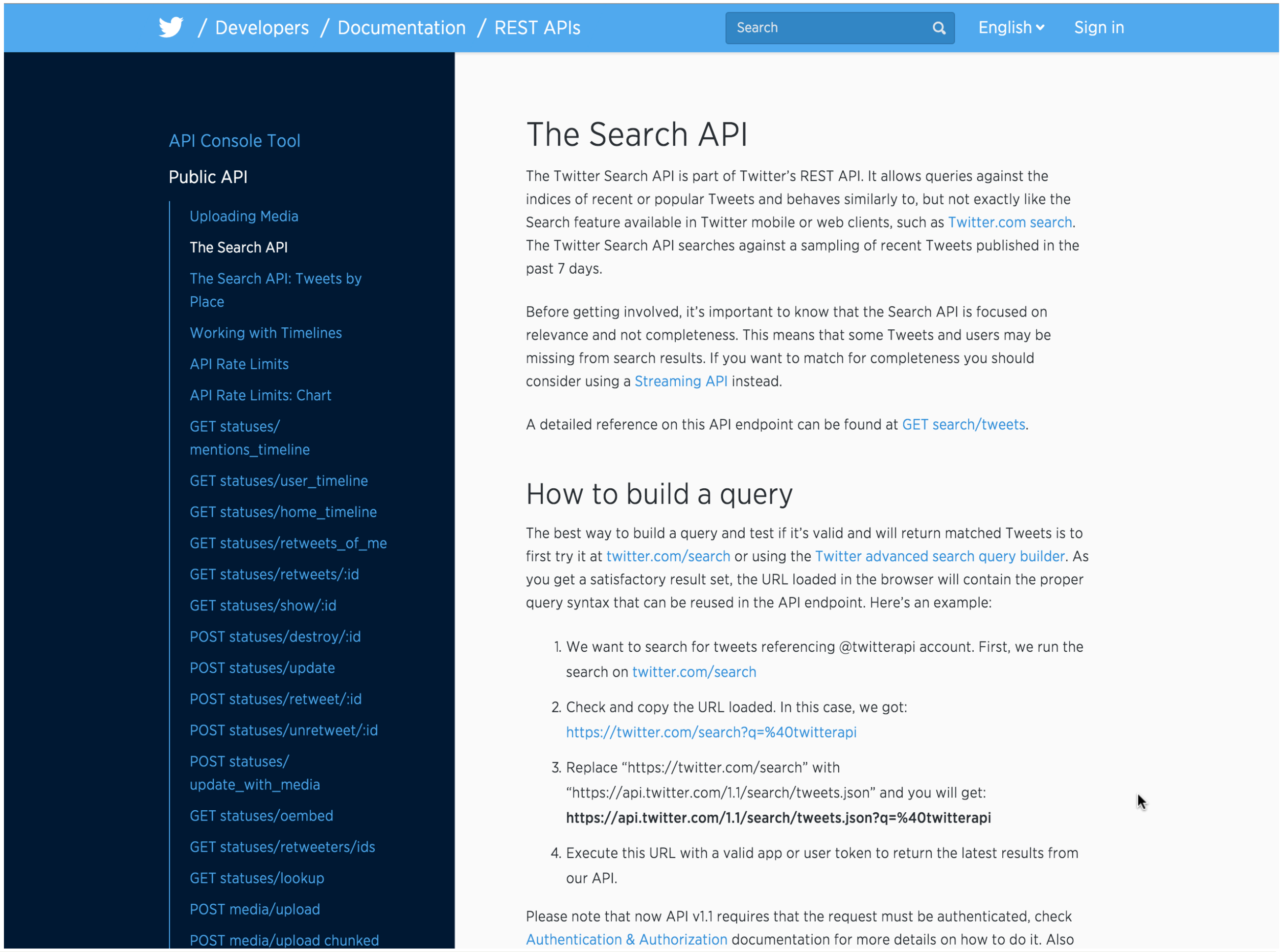
Types of Classification/Labeling

- Binary classification
  - Answers the question does this label/classification apply?
    - Yes or No
    - Assume dichotomous labels (classes)
- Multiple classification
  - Answers the question does this input belong to one of several different categories?

Binary Classifications

- Simple sentiment analysis
  - Is this tweet "happy" or "sad"?
- Generalize to any binary valence
  - Positive to Negative
  - Bright to Dark
  - Introverted to extroverted
- How might this fail?

Sentiment in Twitter – a Query Operator


/ Developers / Documentation / REST APIs
Search
English
Sign in
API Console Tool
Public API
Uploading Media
The Search API
The Search API: Tweets by Place
Working with Timelines
API Rate Limits
API Rate Limits: Chart
GET statuses/mentions_timeline
GET statuses/user_timeline
GET statuses/home_timeline
GET statuses/retweets_of_me
GET statuses/retweets/:id
GET statuses/show/:id
POST statuses/destroy/:id
POST statuses/update
POST statuses/retweet/:id
POST statuses/unretweet/:id
POST statuses/update_with_media
GET statuses/oembed
GET statuses/retweeters/ids
GET statuses/lookup
POST media/upload
POST media/upload chunked

The Search API

The Twitter Search API is part of Twitter's REST API. It allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days.

Before getting involved, it's important to know that the Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results. If you want to match for completeness you should consider using a Streaming API instead.

A detailed reference on this API endpoint can be found at GET search/tweets.

How to build a query

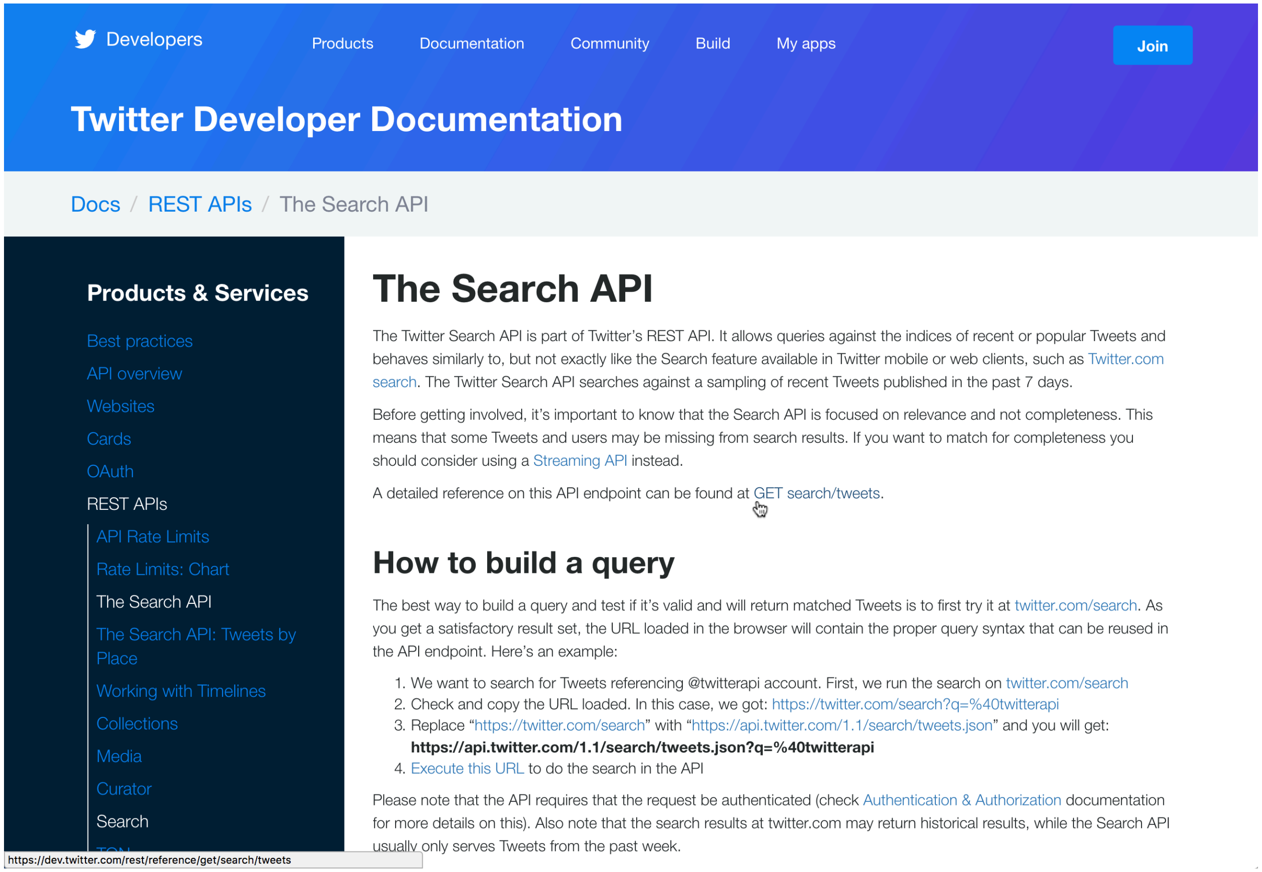
The best way to build a query and test if it's valid and will return matched Tweets is to first try it at twitter.com/search or using the Twitter advanced search query builder. As you get a satisfactory result set, the URL loaded in the browser will contain the proper query syntax that can be reused in the API endpoint. Here's an example:

1. We want to search for tweets referencing @twitterapi account. First, we run the search on twitter.com/search
2. Check and copy the URL loaded. In this case, we got: https://twitter.com/search?q=%40twitterapi
3. Replace "https://twitter.com/search" with "https://api.twitter.com/1.1/search/tweets.json" and you will get: **https://api.twitter.com/1.1/search/tweets.json?q=%40twitterapi**
4. Execute this URL with a valid app or user token to return the latest results from our API.

Please note that now API v1.1 requires that the request must be authenticated, check Authentication & Authorization documentation for more details on how to do it. Also

Sentiment in Twitter
A Query Operator

- REST API
  - Search



The Search API

The Twitter Search API is part of Twitter's REST API. It allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days.

Before getting involved, it's important to know that the Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results. If you want to match for completeness you should consider using a Streaming API instead.

A detailed reference on this API endpoint can be found at GET search/tweets.

How to build a query

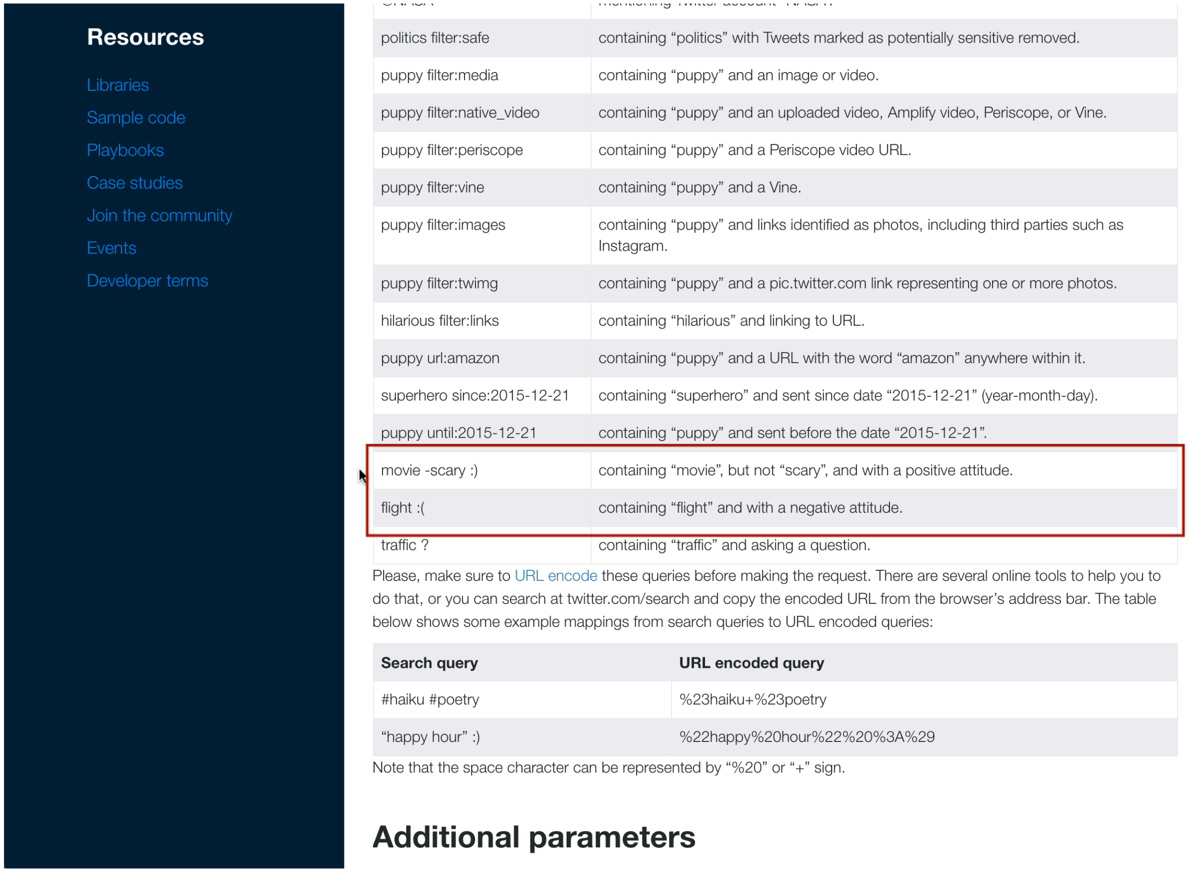
The best way to build a query and test if it's valid and will return matched Tweets is to first try it at twitter.com/search. As you get a satisfactory result set, the URL loaded in the browser will contain the proper query syntax that can be reused in the API endpoint. Here's an example:

1. We want to search for Tweets referencing @twitterapi account. First, we run the search on twitter.com/search
2. Check and copy the URL loaded. In this case, we got: https://twitter.com/search?q=%40twitterapi
3. Replace "https://twitter.com/search" with "https://api.twitter.com/1.1/search/tweets.json" and you will get: **https://api.twitter.com/1.1/search/tweets.json?q=%40twitterapi**
4. Execute this URL to do the search in the API

Please note that the API requires that the request be authenticated (check Authentication & Authorization documentation for more details on this). Also note that the search results at twitter.com may return historical results, while the Search API usually only serves Tweets from the past week.

Sentiment in Twitter
A Query Operator

- Scroll



Resources

- Libraries
- Sample code
- Playbooks
- Case studies
- Join the community
- Events
- Developer terms

@NASA	mentioning Twitter account "NASA"
politics filter:safe	containing "politics" with Tweets marked as potentially sensitive removed.
puppy filter:media	containing "puppy" and an image or video.
puppy filter:native_video	containing "puppy" and an uploaded video, Amplify video, Periscope, or Vine.
puppy filter:periscope	containing "puppy" and a Periscope video URL.
puppy filter:vine	containing "puppy" and a Vine.
puppy filter:images	containing "puppy" and links identified as photos, including third parties such as Instagram.
puppy filter:twimg	containing "puppy" and a pic.twitter.com link representing one or more photos.
hilarious filter:links	containing "hilarious" and linking to URL.
puppy url:amazon	containing "puppy" and a URL with the word "amazon" anywhere within it.
superhero since:2015-12-21	containing "superhero" and sent since date "2015-12-21" (year-month-day).
puppy until:2015-12-21	containing "puppy" and sent before the date "2015-12-21".
movie -scary :)	containing "movie", but not "scary", and with a positive attitude.
flight :(	containing "flight" and with a negative attitude.
traffic ?	containing "traffic" and asking a question.

Please, make sure to URL encode these queries before making the request. There are several online tools to help you to do that, or you can search at twitter.com/search and copy the encoded URL from the browser's address bar. The table below shows some example mappings from search queries to URL encoded queries:

Search query	URL encoded query
#haiku #poetry	%23haiku+%23poetry
"happy hour" :)	%22happy%20hour%22%20%3A%29

Note that the space character can be represented by "%20" or "+" sign.

Additional parameters

Twitter Query Operators

Operator	**Finds tweets…**
watching now	containing both “watching” and “now”. This is the default operator.
“happy hour”	containing the exact phrase “happy hour”.
love OR hate	containing either “love” or “hate” (or both).
beer -root	containing “beer” but not “root”.
#haiku	containing the hashtag “haiku”.
from:interior	sent from Twitter account “interior”.
to:NASA	tweets authored in reply to Twitter account “NASA”.
@NASA	mentioning Twitter account “NASA”.
politics filter:safe	containing “politics” with Tweets marked as potentially sensitive removed.

Twitter Query Operators

Operator	**Finds tweets…**
puppy filter:media	containing “puppy” and an image or video.
puppy filter:images	containing “puppy” and an image.
hilarious filter:links	containing “hilarious” and linking to URL.
superhero since:2015-12-21	containing “superhero” and sent since date “2015-12-21” (year-month-day).
puppy until:2015-12-21	containing “puppy” and sent before the date “2015-12-21”.
movie -scary :)	containing “movie”, but not “scary”, and with a positive attitude.
flight :(	containing “flight” and with a negative attitude.
traffic ?	containing “traffic” and asking a question.

Demo

- Try out Twitter Sentiment operators
- How could we try this?

Other Classification Problems

- Suppose you wanted to classify data using other categories?
- How would you build a classifier?

Process for Creating a Classifier

- Collect Data
- Create a sub-sample
- Pick one (or several) classification algorithms to try
- Select key features
- Score the sub-sample, positive/negative examples
- Train Classifier
- Validate Classifier
- Apply Classifier

Process for Creating a Classifier

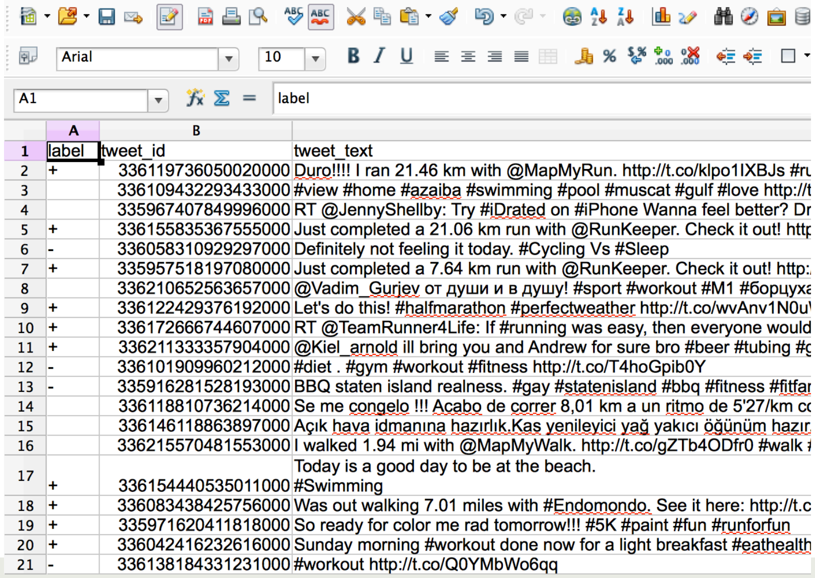
- Collect Data
- Create a sub-sample
- Pick one (or several) classification algorithms to try
- Select key features
- Score the sub-sample, positive/negative examples
- Train Classifier
- Validate Classifier
- Apply Classifier

Samples to Explore

- In hcde user module, ml directory
  - Classification.py – a basic object
  - ClassifyTweet.py – a subclass of Classification
- Sample code
  - explore_feature_selection.py
  - explore_classification.py

Labeled CSV Tweet data

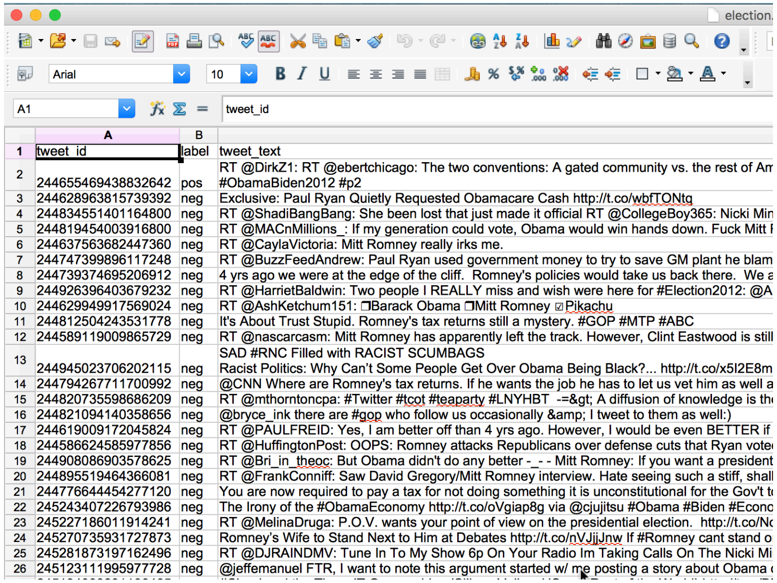
- fitness_label_data1.csv
  - Dump – based on simple_sample.py (using the file output option)
  - Labeled – positive and negative labeling
  - Must have
    - 'label'
    - 'tweet_text'



	A	B	
1	label	tweet_id	tweet_text
2	+	336119736050020000	Duro!!!! I ran 21.46 km with @MapMyRun. http://t.co/klpo1IXBJs #ru
3		336109432293433000	#view #home #azaiba #swimming #pool #muscat #gulf #love http://t
4		335967407849996000	RT @JennyShellby: Try #iDrated on #iPhone Wanna feel better? Dr
5	+	336155835367555000	Just completed a 21.06 km run with @RunKeeper. Check it out! http
6	-	336058310929297000	Definitely not feeling it today. #Cycling Vs #Sleep
7	+	335957518197080000	Just completed a 7.64 km run with @RunKeeper. Check it out! http:
8		336210652563657000	@Vadim_Gurjev от души и в душу! #sport #workout #M1 #борцуха
9	+	336122429376192000	Let's do this! #halfmarathon #perfectweather http://t.co/wvAnv1N0u
10	+	336172666744607000	RT @TeamRunner4Life: If #running was easy, then everyone would
11	+	336211333357904000	@Kiel_arnold ill bring you and Andrew for sure bro #beer #tubing #g
12	-	336101909960212000	#diet . #gym #workout #fitness http://t.co/T4hoGpib0Y
13	-	335916281528193000	BBQ staten island realness. #gay #statenisland #bbq #fitness #fitfar
14		336118810736214000	Se me congelo !!! Acabo de correr 8,01 km a un ritmo de 5'27/km co
15		336146118863897000	Açık hava idmanına hazırlık.Kas yenileyici yağ yakıcı öğünüm hazır
16		336215570481553000	I walked 1.94 mi with @MapMyWalk. http://t.co/gZTb4ODfr0 #walk #
17	+	336154440535011000	Today is a good day to be at the beach. #Swimming
18	+	336083438425756000	Was out walking 7.01 miles with #Endomondo. See it here: http://t.c
19	+	335971620411818000	So ready for color me rad tomorrow!!! #5K #paint #fun #runforfun
20	+	336042416232616000	Sunday morning #workout done now for a light breakfast #eathealth
21	-	336138184331231000	#workout http://t.co/Q0YMbWo6qq

Labeled CSV Tweet data

- Two samples for the fitness data
  - fitness_label_data1.csv
  - fitness_label_data2.csv



A1 tweet_id

	A	B	
1	tweet_id	label	tweet_text
2	244655469438832642	pos	RT @DirkZ1: RT @ebertchicago: The two conventions: A gated community vs. the rest of Am #ObamaBiden2012 #p2
3	244628963815739392	neg	Exclusive: Paul Ryan Quietly Requested Obamacare Cash http://t.co/wbfTONtq
4	244834551401164800	neg	RT @ShadiBangBang: She been lost that just made it official RT @CollegeBoy365: Nicki Min
5	244819454003916800	neg	RT @MACnMillions_: If my generation could vote, Obama would win hands down. Fuck Mitt
6	244637563682447360	neg	RT @CaylaVictoria: Mitt Romney really irks me.
7	244747399896117248	neg	RT @BuzzFeedAndrew: Paul Ryan used government money to try to save GM plant he blam
8	244739374695206912	neg	4 yrs ago we were at the edge of the cliff. Romney's policies would take us back there. We a
9	244926396403679232	neg	RT @HarrietBaldwin: Two people I REALLY miss and wish were here for #Election2012: @A
10	244629949917569024	neg	RT @AshKetchum151: ❒Barack Obama ❒Mitt Romney ☑ Pikachu
11	244812504243531778	neg	It's About Trust Stupid. Romney's tax returns still a mystery. #GOP #MTP #ABC
12	244589119009865729	neg	RT @nascarcasm: Mitt Romney has apparently left the track. However, Clint Eastwood is still
13	244945023706202115	neg	SAD #RNC Filled with RACIST SCUMBAGS Racist Politics: Why Can’t Some People Get Over Obama Being Black?... http://t.co/x5I2E8m
14	244794267711700992	neg	@CNN Where are Romney's tax returns. If he wants the job he has to let us vet him as well a
15	244820735598686209	neg	RT @mthorntoncpa: #Twitter #tcot #teaparty #LNYHBT -=&gt; A diffusion of knowledge is th
16	244821094140358656	neg	@bryce_ink there are #gop who follow us occasionally &amp; I tweet to them as well:)
17	244619009172045824	neg	RT @PAULFREID: Yes, I am better off than 4 yrs ago. However, I would be even BETTER if
18	244586624585977856	neg	RT @HuffingtonPost: OOPS: Romney attacks Republicans over defense cuts that Ryan vote
19	244908086903578625	neg	RT @Bri_in_theoc: But Obama didn't do any better -_- - Mitt Romney: If you want a president
20	244895519464366081	neg	RT @FrankConniff: Saw David Gregory/Mitt Romney interview. Hate seeing such a stiff, shall
21	244776644454277120	neg	You are now required to pay a tax for not doing something it is unconstitutional for the Gov't t
22	245243407226793986	neg	The Irony of the #ObamaEconomy http://t.co/oVgiap8g via @cjujitsu #Obama #Biden #Econo
23	245227186011914241	neg	RT @MelinaDruga: P.O.V. wants your point of view on the presidential election. http://t.co/No
24	245270735931727873	neg	Romney’s Wife to Stand Next to Him at Debates http://t.co/nVJjjJnw If #Romney cant stand o
25	245281873197162496	neg	RT @DJRAINDMV: Tune In To My Show 6p On Your Radio Im Taking Calls On The Nicki Mi
26	245123111995977728	neg	@jeffemanuel FTR, I want to note this argument started w/ me posting a story about Obama

Process for Creating a Classifier

- Collect Data
- Create a sub-sample
- Pick a Classifier
- Select key features
- Score the sub-sample, positive/negative examples
- Train Classifier
- Validate Classifier
- Apply Classifier

Feature Selection

- What are the 'features' of tweets?

- How could you decide which features are important?

Demo Feature Selection

Demo Classification

Interpreting Top Features

```
Most Informative Features
  #Swimming = True          negati : positi =      4.7 : 1.0
  #gym = True               negati : positi =      4.7 : 1.0
  #fitness = True           negati : positi =      3.9 : 1.0
  #RunKeeper = True         positi : negati =      3.4 : 1.0
  completed = True          positi : negati =      3.2 : 1.0
  today. = True             negati : positi =      2.8 : 1.0
  #Workout = True           negati : positi =      2.8 : 1.0
  bring = True              negati : positi =      2.8 : 1.0
```

Reminder

- Week 9
  - Project "Studio" class session
    - Ray and I will wander from group to group
    - Location TBD