Rogelio Zuniga R.

Travel Reviews and Airline Revenue

March 2016 | DAT-30

Background - RASM

MIT Airline Data Project

Massachusetts Institute of Technology

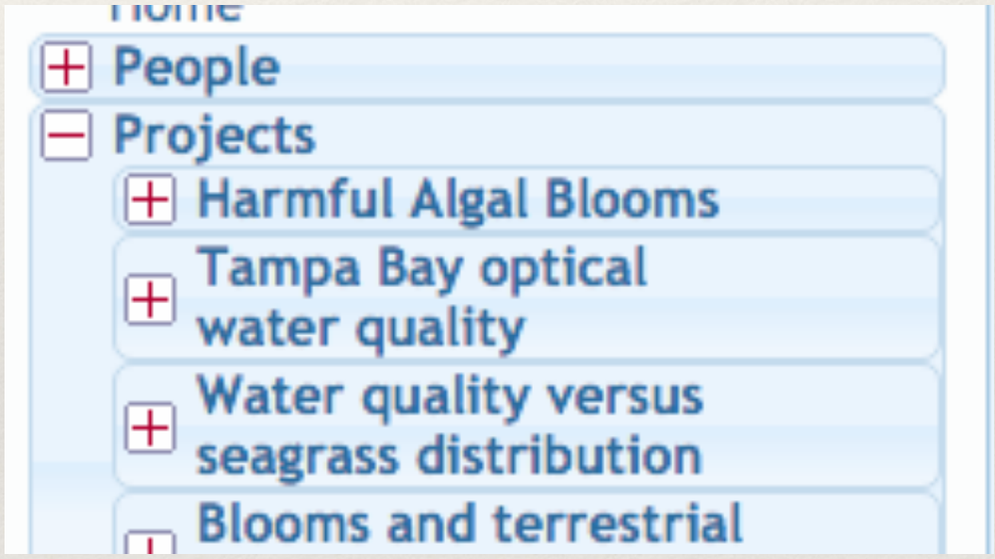
Passenger Revenue per Available Seat Mile (PRASM)

(cents per Available Seat Mile)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
American	10.17	9.19	8.37	8.62	8.61	9.41	10.25	10.65	11.14	9.90	10.93	11.62	12.27	12.67
Continental	9.62	8.81	8.53	8.62	8.75	9.27	9.94	10.45	11.05	9.39	10.62	11.75	-	-
Delta	9.60	8.40	8.15	8.53	8.33	8.48	9.30	9.99	10.67	9.14	10.57	11.50	12.16	12.63
Northwest	9.20	8.34	8.28	8.55	9.21	9.62	10.56	10.77	11.11	9.09	-	-	-	-
United	9.46	8.17	7.74	7.78	8.22	8.88	9.71	10.46	10.84	9.18	11.03	11.90	11.95	12.14
US Airways	11.34	9.85	9.26	9.55	9.35	9.44	16.37	14.57	10.75	9.27	10.44	11.52	11.91	12.21
America West	8.02	7.28	7.11	7.56	7.27	8.23	-	-	-	-	-	-	-	-
--sub Network	**9.73**	**8.64**	**8.23**	**8.45**	**8.57**	**9.09**	**9.95**	**10.46**	**10.94**	**9.37**	**10.73**	**11.64**	**12.09**	**12.43**
Southwest	9.00	8.07	7.58	7.80	7.94	8.31	9.21	9.19	9.88	9.74	11.28	12.11	12.05	12.81
jetBlue	7.41	7.29	7.47	7.05	6.44	6.83	7.78	8.19	9.66	9.18	9.81	10.95	11.35	11.60
AirTran	10.22	9.83	8.49	8.78	8.35	9.04	9.52	9.66	10.06	8.92	9.75	10.54	-	-
Frontier	10.72	9.16	8.58	8.27	7.82	8.77	9.07	8.84	9.91	8.75	9.86	12.32	9.81	10.48
Virgin America	-	-	-	-	-	-	-	1.69	7.59	7.35	8.60	9.67	9.76	10.49
-- sub LCC	**9.17**	**8.24**	**7.71**	**7.83**	**7.73**	**8.16**	**8.97**	**9.00**	**9.81**	**9.37**	**10.58**	**11.56**	**11.61**	**12.27**
Alaska	9.11	8.73	8.39	8.57	8.80	9.50	10.20	10.24	10.52	9.98	10.64	11.36	11.67	11.50
Hawaiian	6.39	6.16	8.30	8.69	9.69	9.67	9.90	9.64	11.62	10.72	11.38	12.31	12.03	11.53
Allegiant	-	-	-	-	3.94	5.69	6.46	6.94	7.73	7.10	7.59	8.64	8.65	8.85
-- sub Other	**8.25**	**7.93**	**8.31**	**8.43**	**8.83**	**9.35**	**9.83**	**9.76**	**10.48**	**9.77**	**10.37**	**11.24**	**11.34**	**11.13**

Sargassum (brown macroalga)

"In 2015, many beaches in the Caribbean and Mexico suffered from Sargassum landing, with numerous news reports on their local impacts. To date, however, no one knows what caused the dramatic increases in Sargassum landing in 2015."



The Optical Oceanography Lab (http://optics.marine.usf.edu) has been tracking Sargassum since 2010 using satellite imagery and numerical models, and has provided near real-time daily imagery on the Web

Motivation

- Connection between
  - E**nvironmental indicators** - beach debris (trash, seaweed), shoreline integrity (coral reef health, presence of mangroves), water quality (clarity)
  - A**irline industry revenue** (RASM)
- Data sets are very low quality - small datasets, inconsistent research
- Place a **Monetary value** on environmental factors in order to prioritize issues and locations for program decisions

My Question

- “Is there a relationship between perceived destination cleanliness and travel to that location?”
- “Can social media reviews be trusted to represent true sentiment?”
- Is there a relationship between perceived environmental factors and revenue?

The Data: Twitter

- **Twitter API search**: Geolocation coordinates and key words determined to return relevant comments :

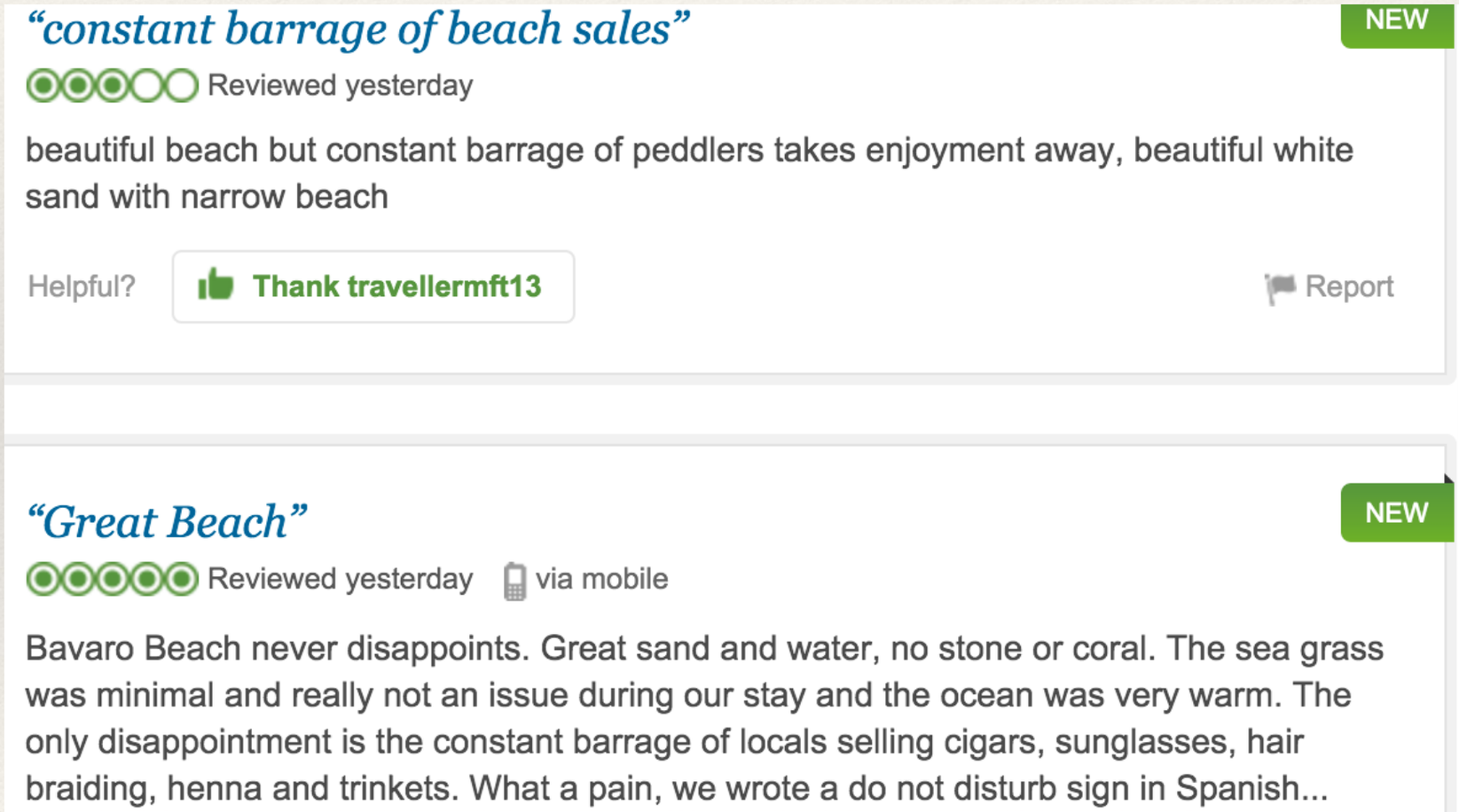
"https://twitter.com/search?q=beach%20trash%20jamaica&src=typd"

Dr. Claire Nelson @DrClaireNelson · 12 Nov 2015

Now that would be nice.. I am sick of seeing people leave their **trash** on the **beach** in **Jamaica**. We mustban Styrofoam

The Data: TripAdvisor

- **TripAdvisor**: TA "place" codes



Exploring Social Media is Hard!

Twitter: search API using TwitterSearch

- Geocodes don't help
- Keywords lead to irrelevant returns
- Very hard to search anything except new tweets

Exploring Social Media is Hard

TripAdvisor: direct from the website using BeautifulSoup

- Keywords lead to irrelevant returns
- Messy results!
- Need to organize a lot of results
- Not allowed to use API for mining data

Yikes!

- Without being able to search Tweets far into the past I can't compare sentiment scores over time
- On the other hand TripAdvisor reviews are far better qualitatively

Better Quality but More Work

8,842 Reviews from our TripAdvisor Community

Read reviews that mention:

Search reviews

All reviews | can walk for miles | long walks | plenty of palm trees | seaweed problem | white powder | walk forever | great sand | cleaned every day | washed ashore | great walking | lots of water sports | seaweed floating | cleaned everyday | gentle surf | great resorts | lovely ocean | washing up | plenty of sun beds | nice temperature | clean rooms

Traveler rating

- ☐ Excellent (25)
- ☐ Very good (15)
- ☐ Average (4)
- ☐ Poor (4)
- ☐ Terrible (0)

Traveler type

- ☐ Families (12)
- ☐ Couples (24)
- ☐ Solo (1)
- ☐ Business (0)
- ☐ Friends (4)

Time of year

- ☐ Mar-May (7)
- ☐ Jun-Aug (15)
- ☐ Sep-Nov (11)
- ☐ Dec-Feb (15)

Language

- ○ All languages
- ◉ English (48)
- ○ Spanish
- ○ Portuguese

```
▼<div class="entry">
  ▼<p id="review_301296294">
      "
      This is a lovely beach when the "
      <span class="searchHit">seaweed</span>
      " isn't bad. For most days of our trip, it w
      the last day, it was so bad the hotels just
      one of the resorts next to ours was using a
      entire time I was ankle deep in it and had t
      always so bad, but due to it, I would not ch
```

Decisions…

- Use only TripAdvisor?
  - Decided to use only TripAdvisor reviews
- How to deal with reviews that include keywords but are not relevant? Naive Bayes?
  - Use LDA and k-means to see if there are relevant topics
- How long will it take to gather enough data, scrape it and process it?
  - It takes much longer than expected to setup a clean data set

Data Cleaning

- **Download** of html files to local disk via wget:

```
#!/bin/bash
URL_start="https://www.tripadvisor.com/Attraction_Review-"
PLACE_CODE="g666621-d4053401"
SEP="-Reviews-or"
URL_end="-Dover_Beach-St_Lawrence_Gap_Christ_Church_Parish_Barbados.html"
COUNTER=10
    while [  $COUNTER -lt 4000 ]; do
    echo `wget --convert-links --domains www.tripadvisor.com --no-parent $URL_start
$PLACE_CODE$SEP$COUNTER$URL_end`
       let COUNTER+=10
   done
```

Data Cleaning

- Parse with **BeautifulSoup**:
  - ratingDateResults = nextTag.find(‘span','relativeDate')
  - nextReviewTag = nextTag.find(‘p’,'partial_entry')

	place_code	destination_code	date	review
0	g1006573	d1102010	February 27 2016	
1	g1006573	d1102010	February 27 2016	
2	g1006573	d1102010	February 13 2016	This is an experience that should not be misse...
3	g1006573	d1102010	February 13 2016	
4	g1006573	d1102010	February 13 2016	The horses are well cared for and well behaved...
5	g1006573	d1102010	February 13 2016	I had the most wonderful time riding with Pamp...
6	g1006573	d1102010	February 13 2016	We have horses at home and wanted to ride on a...

Data Cleaning

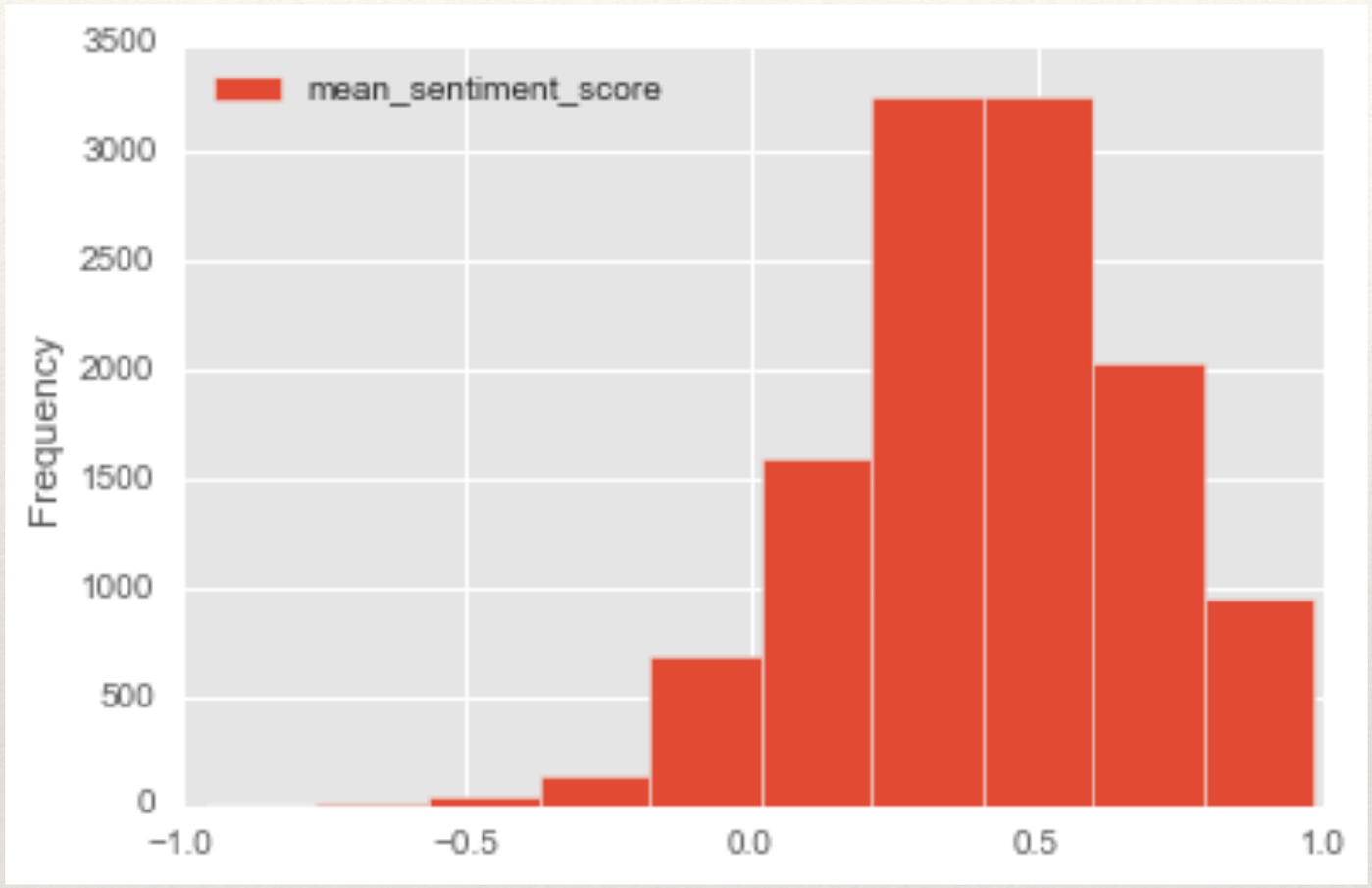
- The resulting data set is made up of:
- **A place code**: g1006573
- **A destination code**: d1102010
- **A date**: February 27 2016
- **A review text**: "The guide and ride was great. Communication was not the best. They use public transportation and didn't make then clear to use (how are we supposed to know what bus to get on). Make sure that they make it very clear to you."

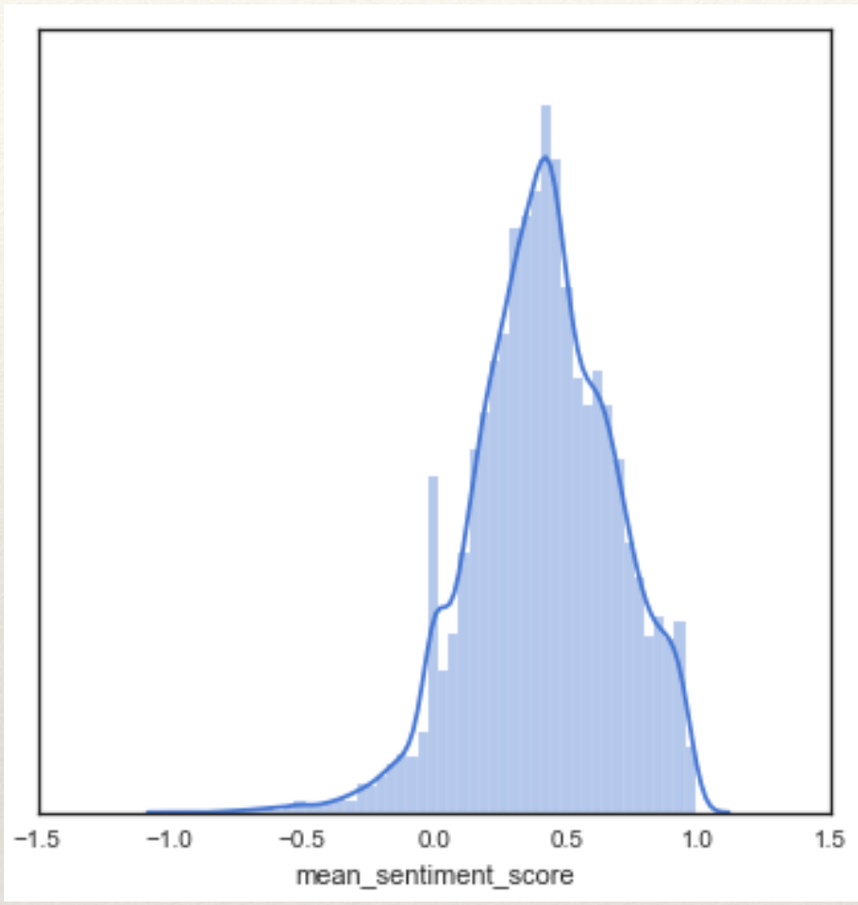
2 Problems With Dataset

Destination	Positive	Negative	Neutral	Total
Enterprise Beach	**264**	7	23	294
Bavaro Beach	**8,158**	240	670	9,068
Seven Mile Beach	**3,652**	116	271	4,039
Palm Beach	**3,685**	80	275	4,040

Over a window of **only ~ 35 days**

Sentiment Distribution





	place_code	destination_code	date	review	mean_sentiment_score
1	g147249	d148915	2016-02-28	Great beach very nice clear water. The fact th...	0.621067
2	g147249	d148915	2016-02-27	I love this beach because the water was the wa...	0.399975
3	g147249	d148915	2016-02-25	This was my favorite beach in Aruba. The waves...	0.194900
4	g147249	d148915	2016-02-25	much better beach then where the hotels are. W...	0.118033

Sentiment Scoring with Vader

Stopped here a couple of times.
{'neg': 0.322, 'neu': 0.678, 'pos': 0.0, 'compound': -0.2263}

The beach is a bit rocky but I did not think it was too bad.
{'neg': 0.284, 'neu': 0.716, 'pos': 0.0, 'compound': -0.6956}

The kids were able to find some pretty sea shells here.
{'neg': 0.0, 'neu': 0.758, 'pos': 0.242, 'compound': 0.4939}

It's Aruba.
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

How can you find anything wrong with the beaches?
{'neg': 0.279, 'neu': 0.721, 'pos': 0.0, 'compound': -0.4767}

Sentiment Hits & Misses

Great snorkel beach.
{'neg': 0.0, 'neu': 0.328, 'pos': 0.672, 'compound': 0.6249}
Not too many tourist around when you avoid the times that the tourist boats get there.
{'neg': 0.128, 'neu': 0.872, 'pos': 0.0, 'compound': -0.296}

Is there a bad beach in Aruba?
{'neg': 0.412, 'neu': 0.588, 'pos': 0.0, 'compound': -0.5423}

Enjoyed this beach immensely!
{'neg': 0.0, 'neu': 0.455, 'pos': 0.545, 'compound': 0.5562}

Google where the Snorkeling entry point is because it is not obvious (hint- go towards the left as you look from the Parking lot).
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

Nice little beach which was not at all crowded!
{'neg': 0.0, 'neu': 0.721, 'pos': 0.279, 'compound': 0.4753}

LDA Topics

max_df=1, stop_words=None

```
[(0,
  [(u'aninal', 0.0020600404719852213),
   (u'habitats', 0.0020600404719852213),
   (u'acquire', 0.0020600404719852213),
   (u'drenched', 0.0020384282085782156),
   (u'waving', 0.0020384282085782156),
   (u'availale', 0.0020383832326325304),
   (u'wate', 0.0020383832326325304),
   (u'tom', 0.0020383677865344124),
   (u'shauna', 0.0020383677865344124),
   (u'lenny', 0.0020382854242067426)]),
 (1,
  [(u'mufongo', 0.0023579345510437925),
   (u'terrified', 0.0012156351635225519),
   (u'garlic', 0.0012156351635225519),
   (u'plantain', 0.0012156351635225519),
   (u'spelled', 0.0012156351635225519),
   (u'herb', 0.0012156351635225519),
   (u'cracklings', 0.0012156351635225519),
   (u'restaurantssome', 0.0012133187267820273),
   (u'shopschair', 0.0012133187267820273),
   (u'gamesafter', 0.0012133187267820273)]),
```

```
(2,
 [(u'ear', 0.0016873252422783432),
  (u'occupy', 0.0016647159991512424),
  (u'4x4', 0.0016646887520366741),
  (u'5usd', 0.0016645741196287696),
  (u'sportparasalingscubajet', 0.0009063482059023786),
  (u'doyou', 0.0009063482059023786),
  (u'amazingwater', 0.0009063482059023786),
  (u'fron', 0.0009063482059023786),
  (u'resortshotel', 0.0009063482059023786),
  (u'prity', 0.0009063482059023786)]),
(3,
 [(u'mondays', 0.0020140821218010223),
  (u'aswell', 0.0019953261164041284),
  (u'peppers', 0.0010646412544167084),
  (u'agaumizukraft', 0.0010646412544167084),
  (u'caymana', 0.0010646412544167084),
  (u'spice', 0.0010646412544167084),
  (u'curry', 0.0010646412544167084),
  (u'crepes', 0.0010646411202959829),
  (u'tony', 0.0010646411202959829),
  (u'rene', 0.0010646411202959829)])]
```

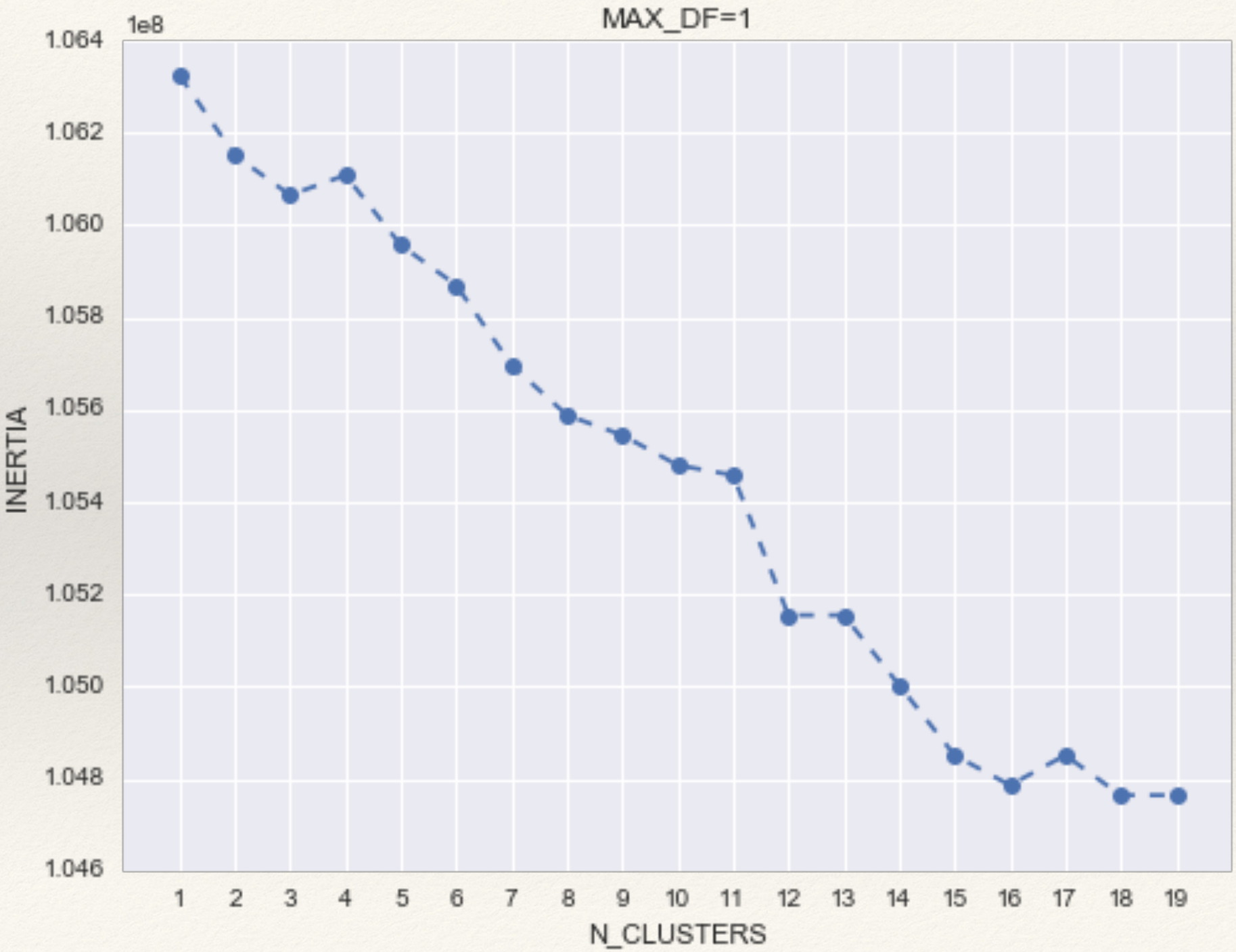
LDA Topics

max_df=.7, stop_words="english"

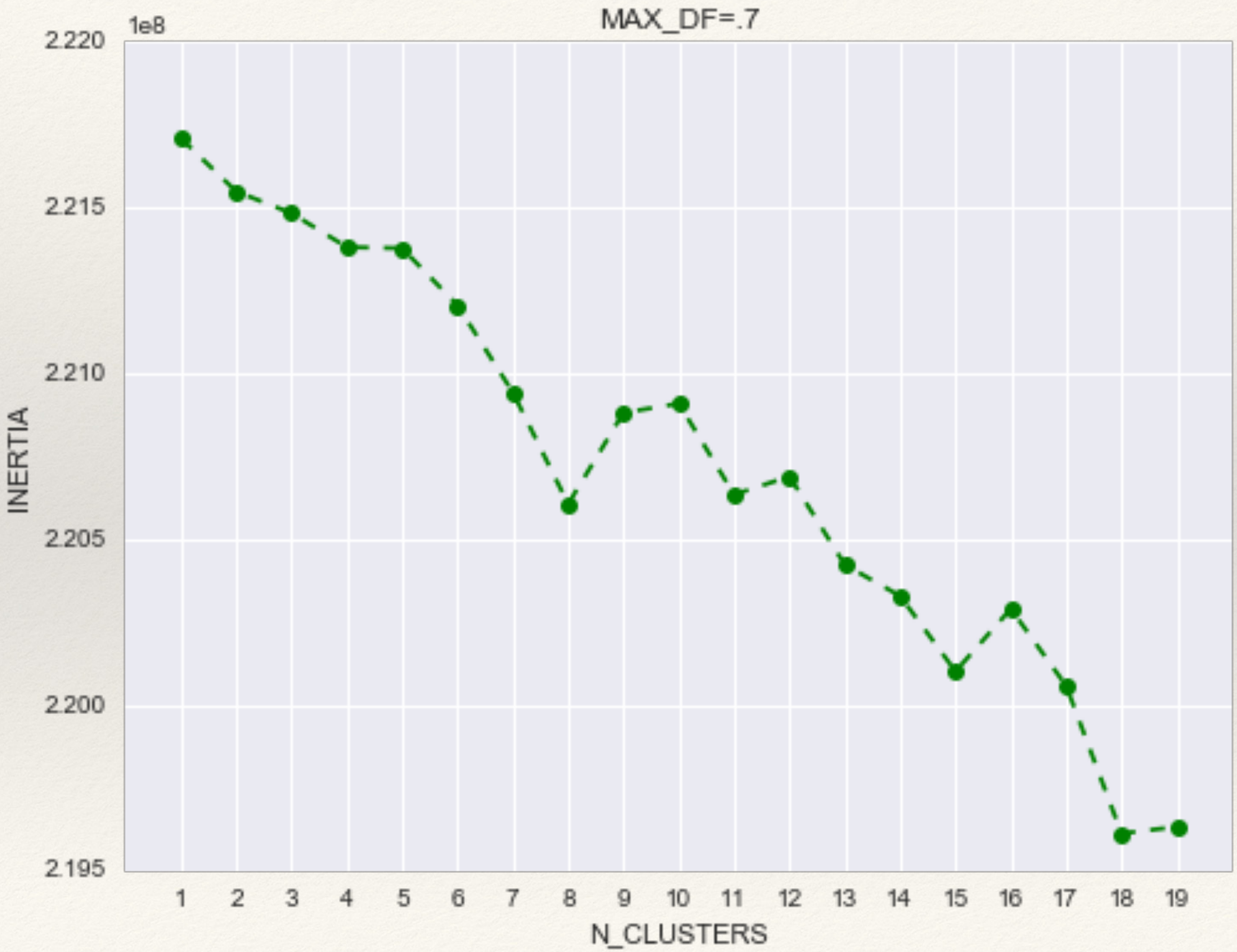
```
[(0,
  [(u'beach', 0.0703167907057414),
   (u'animals', 0.0264001989130390O5),
   (u'great', 0.0184421387268907 32),
   (u'mile', 0.0155898626559750 43),
   (u'seven', 0.0138246486182101 72),
   (u'best', 0.0126194627409889 14),
   (u'beautiful', 0.0114050998409280 27),
   (u'nice', 0.0107958519486959 02),
   (u'place', 0.0104521726514928 59),
   (u'recommend', 0.0100649599900008 83)]),
 (1,
  [(u'beach', 0.034572778453555 22),
   (u'great', 0.0261845414964749 19),
   (u'cayman', 0.0222135069489492 88),
   (u'water', 0.0146756121646208 18),
   (u'just', 0.0134667491901641 05),
   (u'time', 0.0119859405104613 68),
   (u'beautiful', 0.0111075113828811 65),
   (u'people', 0.0109387697679076 27),
   (u'place', 0.0098436167377300 9),
   (u'cruise', 0.00948882410557339 54)]),
```

```
(2,
 [(u'beach', 0.0258552956684646 97),
  (u'feed', 0.0216818010202478 18),
  (u'chairs', 0.0132645347009903 8),
  (u'city', 0.0117494150285910 16),
  (u'staff', 0.0114835533803174 05),
  (u'great', 0.0105410781051324 94),
  (u'beautiful', 0.0098275179858099 502),
  (u'water', 0.0095563305827567 786),
  (u'friendly', 0.0093701350299330 334),
  (u'umbrellas', 0.0088307145237483 675)]),
 (3,
  [(u'tour', 0.0208010888821621 89),
   (u'stingrays', 0.0204018091103902 11),
   (u'trip', 0.0140145283587803 75),
   (u'great', 0.0137816753908499 67),
   (u'beach', 0.0109283796344243 87),
   (u'good', 0.0103020477855715 3),
   (u'stingray', 0.0101051804876805 35),
   (u'sting', 0.0098509415835358 154),
   (u'experience', 0.0095380560553601 952),
   (u'friendly', 0.0093559476388164 441)]),
```

K-MESS!


MAX_DF=1
1e8
1.064
1.062
1.060
1.058
1.056
1.054
1.052
1.050
1.048
1.046
INERTIA
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
N_CLUSTERS

K-MESS!


MAX_DF=.7
1e8
2.220
2.215
2.210
2.205
2.200
2.195
INERTIA
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
N_CLUSTERS

Conclusions and Learnings

- NLP is really FUN!
- Very hard to use a “free” social media text corpus
- A lot of data engineering involved
- Should setup experiment well ahead of time
- Use alternative sources for the corpus