How to derive data-driven insights ... from user-generated content
Text analysis - automatically, regularly & based on large amount of data.

1. Automatically gather relevant content
2. Cleaning & Linguistics
3. Relevance ranking of gathered content
4. Data-driven calculation of main insights
5. Visualization in dashboard & reports
What do people care about in NYC?

What is a trending hotel and what about bars?

Are there unmet needs I could address?

Products for new target groups
Marketing campaigns in new places
Data-driven category management

...
Founded in 2017

Big Data text analytics experts

Located in Munich
Our Big Data market research approach explained on a common situation:

Visiting foreign places.
The problem:
Too little time, too much going on.
Opinion of 1 author
Too many opinions

Restaurants in New York City

Your Selections
- Restaurants
- Japanese

Establishment Type
- Restaurants
- Dessert
- Coffee & Tea

Reservations
- Online Reservations

Cuisines & Dishes

Sort by: Relevance

| Name | Rating | Reviews | Type | Price | Location | Rating
|------|--------|---------|------|-------|----------|--------|
| Fuji Hibachi Japanese Steak House sushi bar | 4.5/5 | 38 Reviews | Japanese, Asian, Grill, Steakhouse, Sushi, Vegetarian Friendly, Vegan Op... | $$$ | New York City | "AMAZING SUSHI AND HIBACHI!!!" 04/24/2018
| Momosan Ramen & Sake | 4.5/5 | 400 Reviews | Japanese, Asian, Soups, Vegetarian Friendly, Gluten Free Options | $$$ | New York City | "The gyozas are perfection." 05/31/2017
| Izakaya Mew | 4.5/5 | 859 Reviews | Japanese, Sushi, Asian, Vegetarian Friendly, Vegan Options, Gluten Free Options | $$$-$$$$ | New York City | "Husband and son both had tonkatsu ramen..." 02/19/2018

Order Online
Making sense of ...

12,5 years

of NYC TripAdvisor Forum
Jan 2005 – Jun 2018
Much content!

1.6 mio posts

89.815 users

Nobody can read that.
There is one problem with user-generated content:

Too much.
Solution?
Text processing pipeline

1. Automatically gather relevant content
   - Spidering
   - Extraction of text and metadata
   - Normalizing

2. Cleaning & Linguistics
   - Natural language processing
   - Language detection
   - Synonyms

3. Relevance ranking of gathered content
   - Uniqueness
   - Domain similarity
   - Inverted domain frequency

4. Data-driven calculation of main insights
   - Topic modeling
   - Word embeddings
   - Regression
   - Outlier detection
   - Feature extraction
   - Clustering

5. Display in dashboard & reports
   - Overall structure
   - Timelines
   - Semantics
   - Word combinations
   - Categories
   - Word frequencies
Create data-driven personas of NYC-travelers

Data-driven focus points for digital marketing, category management, product design, personalization
How to create data-driven personas

Topic 1

Kim
User #1

Friday night visit to NOHRA
May 2, 2015, 12:49 PM

I need to get tickets for UNCG Five Friday Night at the NOHRA. What time should we arrive to get a ticket? Any tips for the visit since I understand it can be very crowded?

Thanks for your help.

Ferry from Red Hook to Pier 11/10 Wall st.
Jan 2, 2017, 4:44 AM

Kim: I'm going to take a ferry from Red Hook to Manhattan.
Kim: There are several options, but I'm not sure which one to choose.
Kim: I've heard about the Green Line and the White Line, but I'm not sure which one is the best.
Kim: Do you have any suggestions or recommendations for me?

Hank
User #2

Getting around New York
May 2, 2015, 12:49 AM

I'm coming to NYC for a week to meet with some colleagues. I am staying in a hotel in Midtown, and I need to get around the city. I'm not sure which transportation option is the best. Can you recommend any options?

Marty
User #3

Elle Island
May 10, 2015, 8:23 PM

The last time we took our kids to NYC, we did the statue tour but Ellis was closed. This time we will go to Ellis only. How long should we account for ferry travel time and then to Ellis?

Thanks in advance.

How do I get from JFK to Jersey City?
Dec 10, 2015, 8:21 AM

Good morning, can someone help me get from JFK to the New Jersey Transit train station? Would this be inexpensive?

Thanks.

Olivia
#68,641

Where to watch EPL near Yankee Stadium?
May 5, 2016, 5:44 AM

I'm due to attend the Yankee v. Blue Jays game on the 19th. I'm in town on Thursday evening which is the closest I'll be able to watch an English Premier League soccer game. I'm a bit lost for options. Any recommendations?

Ferry from Red Hook to Pier 11/10 Wall St.
Jan 2, 2017, 4:44 AM

Olivia: I'm trying to decide if I should take the ferry from Red Hook to Pier 11/10 Wall St. or take the train from the airport. Can you provide any advice or recommendations?

Topic 2

Topic 3

Topic 4

Topic 5

Topic 6
Result: 6 data-driven Personas

Topic 1: Transportation seeker
jfk train airport flight car taxi bus way fly arrive

Topic 2: NYC Newbie
thank nyc hi look help good suggestion appreciate advice visit

Topic 3: Accommodation seeker
hotel stay room night look area square bed price times

Topic 4: Hedonist
just great good time like place think really people food

Topic 5: Organized traveler
day tour walk museum brooklyn square island central park park plan

Topic 6: Discount hunter
ticket buy game seat book purchase discount just sell website
Summary topic modeling

1st take-away business value of data-driven personas

- Detect hidden customer segments based on interest
- Detect distinct topics customer segments talk about
- Decisions backed by data - for digital marketing, category management, product design, personalization, ...
Detect what people are interested in when talking about NYC.

Align your messages to what people actually like about you. Detect changing interests in real-time.
Search result for “airport” in the TripAdvisor forum
Search result for “airport” after training word embeddings

There are three airports in New York:

- JFK (John F. Kennedy)
- EWR (Newark Liberty International Airport)
- LGA (LaGuardia)
Search result for “cocktail” with the same trained word embeddings

Three NYC special drinks:

Nightcap  Mimosa  Brooklyn Lager
Summary word embeddings

2\textsuperscript{nd} take-away business value of semantic context

- Detect relevant context of a topic
- Detect changing interests in real-time
  - Benefit from changing trends
  - Create semantically aware search results
Detect trend changes: NYC Sightseeing Trends.
Detect trend changes: NYC Sightseeing Trends.
Detect trend changes: NYC Sightseeing Trends. Losers

Regressions
<table>
<thead>
<tr>
<th>Trending sights</th>
<th>long term</th>
<th>Short term</th>
<th>Absolute</th>
</tr>
</thead>
<tbody>
<tr>
<td>times square</td>
<td>0.00%</td>
<td>4.69%</td>
<td>4.69%</td>
</tr>
<tr>
<td>one world observatory</td>
<td>-0.30%</td>
<td>2.32%</td>
<td>2.61%</td>
</tr>
<tr>
<td>path</td>
<td>0.23%</td>
<td>2.16%</td>
<td>1.93%</td>
</tr>
<tr>
<td>chinatown</td>
<td>-0.26%</td>
<td>1.63%</td>
<td>1.89%</td>
</tr>
<tr>
<td>soho</td>
<td>-0.31%</td>
<td>1.36%</td>
<td>1.67%</td>
</tr>
<tr>
<td>chelsea</td>
<td>0.50%</td>
<td>1.85%</td>
<td>1.35%</td>
</tr>
<tr>
<td>greenwich village</td>
<td>-0.06%</td>
<td>1.28%</td>
<td>1.34%</td>
</tr>
<tr>
<td>empire state building</td>
<td>-0.06%</td>
<td>1.21%</td>
<td>1.27%</td>
</tr>
<tr>
<td>roosevelt island</td>
<td>0.18%</td>
<td>1.15%</td>
<td>0.97%</td>
</tr>
<tr>
<td>the high line</td>
<td>0.55%</td>
<td>1.49%</td>
<td>0.94%</td>
</tr>
<tr>
<td>upper east side</td>
<td>-0.14%</td>
<td>0.69%</td>
<td>0.83%</td>
</tr>
<tr>
<td>gershwin theater</td>
<td>0.03%</td>
<td>0.83%</td>
<td>0.80%</td>
</tr>
<tr>
<td>madison square garden</td>
<td>0.02%</td>
<td>0.69%</td>
<td>0.68%</td>
</tr>
<tr>
<td>the oculus</td>
<td>0.25%</td>
<td>0.92%</td>
<td>0.66%</td>
</tr>
<tr>
<td>century 21</td>
<td>-0.45%</td>
<td>0.16%</td>
<td>0.61%</td>
</tr>
<tr>
<td>the mall</td>
<td>0.02%</td>
<td>0.60%</td>
<td>0.58%</td>
</tr>
</tbody>
</table>
#20 Tribeca (Triangle Below Canal)

New York’s most expensive postal code

Textile buildings
Summary regressions

3rd take-away: business value of trend analysis

- Find changing trends
- Focus on short-term hypes & biggest outliers
  - Adjust service offering to trend
  - Benchmark with competitors
To wrap it up
insights & business value from UGC analysis

Insights
- Data-driven personas
- Semantic context
- Changing interests

Business value
- Decisions backed by relevant data for marketing, category management, product design, ...
- Align your messages to what people really like about you & adjust over time
Looking beyond UGC

derive business value from other text sources

Data-driven approach to deriving diverse, un-biased insight from ...

- Technical documentation
- Change requests
- Company wikis
- Scientific publications
- Technical debts
- Future cost-drivers
- Knowledge bottlenecks
- Emerging competing technologies
- ...

Emerging competing technologies
Data-driven travel recommendation from 1.6 mio NYC TripAdvisor posts

Nightcap, Tribeca, Café Gitane, Circle line
Mimosa, Unisphere, Gallagher's Steak House, HOHO bus
Brooklyn Lager, Flea theatre, Yefsi Estiatorio, Free tours 😊