Update

**Restriction Levels:**

1. Initial :

Retrieved 21473 Manhattan listings

Retrieved 433320 Manhattan reviews

Retrieved 18633 Manhattan hosts

Retrieved only 14839 (79.64%) Manhattan hosts with reviews

Retrieved 11345 Manhattan host trips

Retrieved 442380 reviews from hosts who have hosted Manhattan Guests

Total Number of 397693 guests that have reviewed Manhattan listings

Retrieved 192768 (48.47%) guests that have reviewed Manhattan listings

1. Restrict to only Single Listing :

Revised number of Hosts: 15522 (decreased 16.70 %)

Revised number of Listings: 15522 (decreased 27.71 %)

Revised number of Reviews: 271729 (decreased 37.29 %)

Revised number of Overall Guests: 255027 (decreased 35.87 %)

Revised number of Retrieved Guests: 121901 (decreased 36.76 %)

1. Remove Hosts/Listings With No Reviews:

Remove Listings with no reviews

Revised number of Listings: 12025 (decreased 22.53 %)

Remove Hosts with no reviews

Revised number of Hosts: 12025 (decreased 22.53 %)

1. Remove Reviews that are just cancellation notifications:

Removing cancellation notifications from reviews

Revised number of Reviews: 264935 (decreased 2.50 %)

Revised number of Hosts: 11587 (decreased 3.64 %)

Revised number of Listings: 11587 (decreased 3.64 %)

Revised number of Overall Guests: 249495 (decreased 2.17 %)

Revised number of Retrieved Guests: 119609 (decreased 1.88 %)

1. Remove empty /short reviews

Removing empty(, None, nan) reviews

Revised number of Reviews: 264675 (decreased 0.10 %)

Removing reviews less than 5 words

Revised number of Reviews: 257166 (decreased 2.84 %)

Revised number of Hosts: 11510 (decreased 0.66 %)

Revised number of Listings: 11510 (decreased 0.66 %)

Revised number of Overall Guests: 242507 (decreased 2.80 %)

Revised number of Retrieved Guests: 117542 (decreased 1.73 %)

1. Restrict to English Reviews:

Removing reviews based on Language Restrictions

-Reviews that have 'English' language detection agreements between 'langdetect' and googletrans = 224173

-Reviews that contained emoticons in them preventing 'googletrans' from correct detection = 917

-Revised number of Reviews: 225090 (decreased 12.47 %)

Revised number of Hosts: 11333 (decreased 1.54 %)

Revised number of Listings: 11333 (decreased 1.54 %)

Revised number of Overall Guests: 211141 (decreased 12.93 %)

Revised number of Retrieved Guests: 102684 (decreased 12.64 %)

1. Restrict to only hosts with 5 or more reviews – Engaged guests

Revised number of Hosts: 6924 (decreased 38.90 %)

Revised number of Listings: 6924 (decreased 38.90 %)

Revised number of Reviews: 215960 (decreased 4.06 %)

Revised number of Overall Guests: 202989 (decreased 3.86 %)

Revised number of Retrieved Guests: 101288 (decreased 1.36 %)

1. For Picture Comparison: Restrict hosts to only those with one person in their profile picture – 1-to-1 relationship

Revised number of Hosts: 3504 (decreased 49.39 %)

Revised number of Listings: 3504 (decreased 49.39 %)

Revised number of Reviews: 104554 (decreased 51.59 %)

Revised number of Overall Guests: 100549 (decreased 50.47 %)

Revised number of Retrieved Guests: 50751 (decreased 49.89 %)

**Sentiment Analysis** – Satisfaction

5 Tools tested on: AFINN, VADER, SO\_CAL, OPINION LEXICON, SENTIMENT\_140

INTERRATER – 52 Examples of Reviews based on Upper Quarter Length reviews

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Extremely Unsatisfied  (=<0.2) | Quite Unsatisfied  (>0.2 and <=0.4) | Average  (>0.4 and <=0.6) | Quite Satisfied  (>0.6 and <=0.8) | Extremely Satisfied  (0.8< ) |
| Count | 2 | 8 | 7 | 11 | 24 |

Tried Sentence level and using overall Review

BEST PERFORMING ON WHOLE REVIEW: OPINION LEXICON and VADER

BEST PERFORMING ON SENTENCE LEVEL: VADER,

OPINION LEXICON – did not perform well on sentence level as many of the sentences averaged each other out

177703361 : Vader-sentence-level

**LISTING DESCRIPTION The title of the listing is "Renovated Manhattan Apartment" don't be fooled by that! It is nowhere close to that.For one it is on the border of Manhattan and East Harlem but has more of the Harlem feel. The address we were given was incomplete and incorrect so I had to call several times and the owner's neighbor had to speak to my uber driver to get the proper address. PHOTOS In the photos, the property is showed as being in a revived area with beautiful artwork on the street walls. This wasn't our reality when we finally made it to the address and were met by the cleaning lady. The apartment is on top of shops and we missed the entrance because it is sandwiched between two store fronts. We then had to go through a long flight of stairs to get to the 4th floor which I will not recommend for the elderly. Although I cannot say the pictures are incorrect, there was some "photoshopping" during the production of the pictures. NOISE Because it is on a busy street it is very noisy both during the day and at night and there is no way of shutting the noise out at night. AMENITIES Inside, was clean and it was stocked with towels and bedsheets. However, the pictures in the bathroom was misrepresented, there is a broken soap stand inside the bath tub with a somewhat sharp edge sticking out which needs to be fixed and can be dangerous if you have kids with you. We were surprised to know there was no microwave in the apartment as well which was a bit of an inconvenience especially with kids. Overall this is an apartment I will never rent again because the pictures are a misrepresentation of the actual location and look of the place. I must say communication was great and his neighbor who i had to deal with was helpful. Again expect NO "renovated Manhattan apartment and be prepared for the long climb to the 4th floor!**

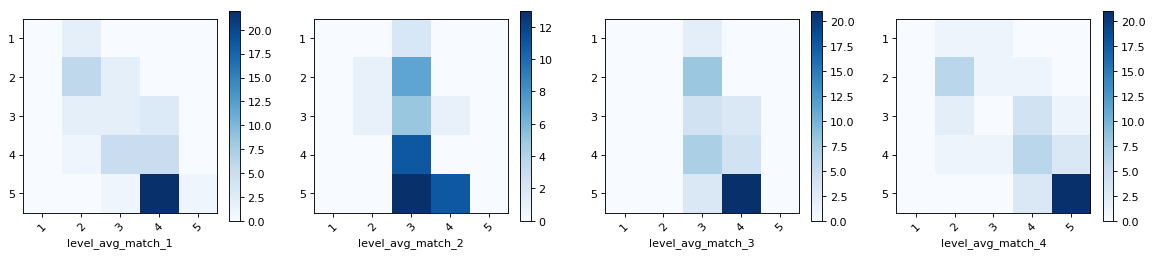
177703361 : opinion\_bow-sentence-level

**LISTING DESCRIPTION The title of the listing is "Renovated Manhattan Apartment" don't be fooled by that! It is nowhere close to that.For one it is on the border of Manhattan and East Harlem but has more of the Harlem feel. The address we were given was incomplete and incorrect so I had to call several times and the owner's neighbor had to speak to my uber driver to get the proper address. PHOTOS In the photos, the property is showed as being in a revived area with beautiful artwork on the street walls. This wasn't our reality when we finally made it to the address and were met by the cleaning lady. The apartment is on top of shops and we missed the entrance because it is sandwiched between two store fronts. We then had to go through a long flight of stairs to get to the 4th floor which I will not recommend for the elderly. Although I cannot say the pictures are incorrect, there was some "photoshopping" during the production of the pictures. NOISE Because it is on a busy street it is very noisy both during the day and at night and there is no way of shutting the noise out at night. AMENITIES Inside, was clean and it was stocked with towels and bedsheets. However, the pictures in the bathroom was misrepresented, there is a broken soap stand inside the bath tub with a somewhat sharp edge sticking out which needs to be fixed and can be dangerous if you have kids with you. We were surprised to know there was no microwave in the apartment as well which was a bit of an inconvenience especially with kids. Overall this is an apartment I will never rent again because the pictures are a misrepresentation of the actual location and look of the place. I must say communication was great and his neighbor who i had to deal with was helpful. Again expect NO "renovated Manhattan apartment and be prepared for the long climb to the 4th floor!**

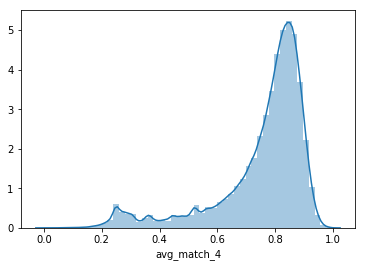
Improvements on conjunctions and negation

Trying to use mood as well to accompany sentiment

After testing on a range of ensembles, an ensemble of the Vader (Whole review) and Vader(sentence level) gave the least number of classification errors, but needs improvement on really bad reviews( 1 rating)



Applying this to all reviews gave this distribution



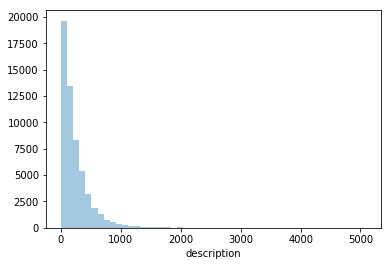
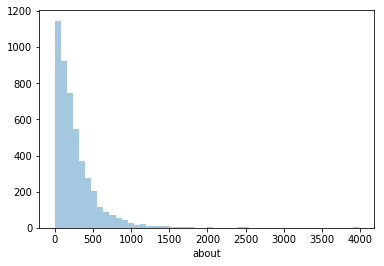
**Host and Guest Profiles** - using Dataset 7

**31.8 %** of Host profiles were blank

**44.9%** of Guest profiles were blank

**77% of Host-Guest relationships** where either **one** or **both** has an empty profile.

From those that were not Blank, here are the distributions of the length:

Hosts Guests

**Vocabulary Length**: 62287

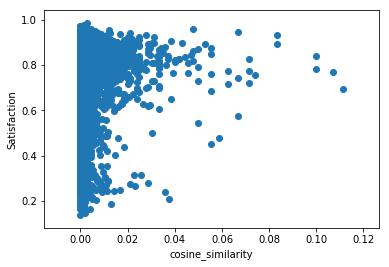
**Top words include:** travel, love, enjoy, work, food, music, meeting, family - hobby like words, and people

Have extracted topics but still improving them.

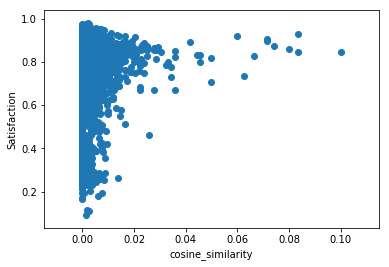
Using profiles Description and vocabulary

**Cosine similarity (BOW) vs satisfaction**

Full

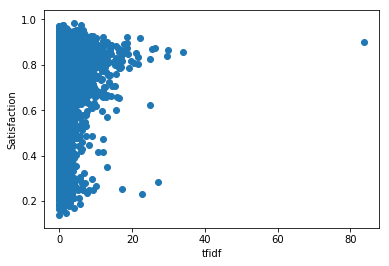


Shared

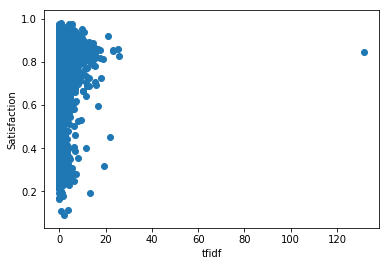


**Tfidf vs satisfaction**

Full



Shared



IN PROGRESS: Experimenting with Using sentiment and Mood analysis to form lists of likes and dislikes of each profile.

Other Updates/concerns:

* When comparing diversity/traveller metric (eg. For country), I can only get a number for the host i.e. the number of diverse interactions he/she has had with guests/hosts. Don’t have any information about the interactions of guests.
* Similarity based on country/location – I have used the longitude/latitude of the guests’ country versus long/lat of USA/NYC – probably be careful as this may just be a listing they have and not actually stay there.
* For those in America – I have also used (a) if it is different from the hosts, and then (b) distances (using long and lat)
* 50% of hosts, have been hosted by other hosts
* 160 hosts, were guests in Manhattan too.
* Host to guest reviews – capped at first 10 – Airbnb site. – Didn’t notice before
* Have done a mini test with demographics where there seems to be a relationship with age . main Demographic Data script – has been running -
* Demographic Data – how to combine into one metric – should I?
* Using AirBnb vocabulary – on what is important to him based on reviews he has given other guests.
* Analysis Deadline 14 August.