Project Summary: Classifying Airbnb Reviews Using NLP

I recently finished Project Fletcher, a project where we were tasked with using both NLP and unsupervised learning in a project. My data source was InsideAirbnb.com, a website that scrapes Airbnb data an posts it directly on their own website. It's a great data set, and I one I was excited to work with.

Data

The data set included 4,911 listed rental properties in the Portland, Oregon area. Those rentals had amassed an impressive total of 318,262 reviews by people who had stayed at the properties. Each of those reviews required the user to provide a text review, rather than just a star rating.

Other than the text ratings, the data included:

- Description and summary of the property provided by hosts
- A summary section about the host
- Type of listing (entire home, room in a home, mother in law suite, etc.)
- The overall rating of the property (from 0-100%), and several other ratings regarding the property's value, the accuracy of the listing, etc. that were on a scale from 1-10
- Information about the guest who provided the review

Methodology

Although the data set in general was quite good, I still did quite a bit of data munging and preprocessing. I created features for each review including, but not limited to:

- Length of the review, in characters
- Number of sentences
- Number of words
- Average length of each sentence
- Most frequent words in each review
- Polarity of each review (positive/negative)
- Subjectivity of each review (objective/subjective)

I then used LDA and spent a lot of time trying to get that right: tuning inputs and outputs, experimenting with different combinations, and trying to get a list of working, viable groups that I could understand and explain. I ended up using bi-grams, and stemming words to get more accurate counts.

I ultimately landed upon 10 topics, which I was moderately happy with. At first I had run into the problem of getting one single topic that was way too big (i.e., it encompassed over 80% of the documents.) I was able to take care of that by tuning my hyperparameters, and I learned a lot in the process.

Results

My topic modeling provided some interesting results. Several of my topics were pretty good, but the issue I kept running into is that topic modeling *on all Airbnb reviews* meant that reviewers were mostly talking about the same "topic" - the Airbnb they had just stayed at! There wasn't enough diversity in the reviews to glom onto. Still, after looking at my topics, and using LDA to get probabilities for each document, I felt confident that my efforts had not been all for naught. I used K-Means clustering to get an idea of which reviews fit best together, and I had a pretty even distribution. Given more time, I'd like to look further into the factors that defined each group and try to validate them otherwise.

I then used the results of my topic modeling with LDA, and the K-Means clusters I had predicted, and fed those in to regression models in order to attempt to predict the price that each listing would go for. My thought was that there might be some properties which were very, very well reviewed, but weren't charging enough given the fact that everyone loved them - and vice-versa.

Sadly, I ran into trouble with the regression. I wasn't able to come up with a statistically significant result, and the NLP features and K-means did not seem to make the analysis any more robust.

Lessons Learned/Next Steps

Looking back, this was a tough project to tackle. I was attracted to it in large part because of my previous work with real estate (which I thought had gone well), and what I perceived as an excellent data source in thousands of Airbnb reviews.

I think the difficulty I had with creating a viable regression model can be summed up in a couple of points. For one, Airbnb reviews seem to carry verry little actual information! They're generally "feel-good", rosy reviews that could be driven by "quid pro quo" (i.e., the hosts rate guests on their stay too, and no one wants to get a bad review in return). In addition, no one is really that "original" with their reviews - each one tends to sound like the next, minus a few details and flourishes. The other point I'd make is that the Portland market seems a bit strange, as well. The price range is generally not that great - and that makes predicting price a very muddy process.

Going forward, I'd be more interested in looking at a corpus/data that is more diverse - from a language perspective, and from a dependent variable perspective. It became more and more obvious to me that the reviews on the site are the "best", most curated experiences that people have had. Whether that's due to people not wanting to get a bad rating, or Airbnb ferreting out bad reviews, or some other factor - I don't know. Still, I learned a lot about NLP, and I'm excited to take what I've learned and apply it to some new data sets - with hopefully less homogenous texts!