TOPIC MODELING: AIRBNB REVIEWS

Submitted By: Tushar Deswal

Preeyank Pable

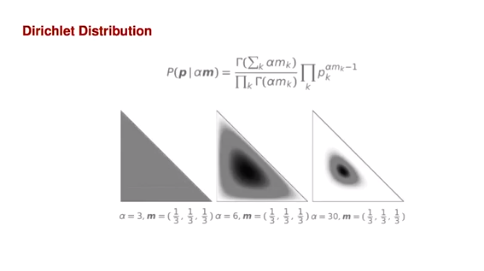
**Business Objective:** The goal is to make the listing process smooth for hosts at Airbnb by considering other attributes through mining the user reviews.

**Current scenario at Airbnb**: Airbnb uses the below listed parameters to upgrade a host to a superhost

1. Overall rating should be 4.8+
2. Response rate should be higher than 90 percent.
3. Should host at least 10 stays in a year.
4. Zero cancellations. Honor every reservation unless there is an extenuating circumstance.

Although, this is a fairly good criterion, however there are other factors as well which should be considered to identify a normal host and a super host i.e. cleanliness of the place, crime rate of the neighborhood .

**Method Used:** We have done topic modeling of user reviews through Latent Dirichlet Distribution (LDA). LDA is a “generative probabilistic model” of a collection of composites made up of parts. In terms of topic modeling, the composites are documents and the parts are words and/or phrases (n-grams). It uses the Dirichlet Distribution which is defined in the range [0,1].



**Dataset:** We did it for Boston. We got two tables from Airbnb website. The listing table contains information about the host and the listing. The reviews table has all the reviews against each listing ID. We used Inner Join in SQL to merge both of these tables together on the common parameter Listing ID.

**Pre-processing the data:** There are three important steps to perform on the data before we pass it through the model. They are:

1. Tokenization- It is the process of breaking a steam of text into tokens.
2. Lemmatization- Removes inflectional endings only and returns the base form of word known as lemma
3. Stop-words removal- Removal of commonly occurring words – a, an, and, as, for, in, of. We don’t want our model to pick these words as they will not provide any information and because of a large number of occurrences of these words, there is a very high probability that our model might pick them.

**Building the model:**

Document – Word Matrix**:** A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents.

The sparsity of the matrix = 0.71%**,** quite good. We use Perplexity and Log-likelihood to come up with optimal number of Topics via Cross - validation. We have considered 10 topics for this model.

Document-Topic Matrix:



**Results and Interpretations:**  Out of all the 10 topics, most significant topics are:

Topic 2(Good experience). Dominant words are: place, stay, great, clean, recommend, etc.

Topic 5(Commute/travel). Dominant words: walk, station, close, minute, subway, etc.

Topic 7(Communicative host). Dominant words: Quick, question, respond, answer, early, etc.