Why use Topic Model?

In order to learn from the reviews and create response accurately, it's necessary to capture the gist of the review descriptions rather than getting lost in the noisy words distributions. By applying topic modelling we expect to come up with a set of coherent words to describe the corpus in general and each review in specific. This very specific word sets would be further used in generating refined responses by the chatbot.

Challenges:

While few topics seemed consistent many seemed incongruous and difficult to interpret.

Analyzing we see the following reasons:

1: The corpus being reviews of restaurants and cuisines; the vocabs are very distinct due to large usage of words like cuisine name ( eg. Ramen, kaya toast etc ) and very specific slangs causing a very flat distribution of words.

2: Some of the words are used in different contexts.

eg. One of the review talking about vegetarian food near Buddha temple and Museum. Here we were expecting topics to contain words like vegetarian, soup, lunch but the topic shows words like entrance, street, museum. This is because the word 'museum' is more associated with "museum Restaurants", "Art science museum" and so on.

3: Spelling errors was another major challenge causing flat distribution of the words as every incorrect spelling was treated as a new word distorting relative distributions.

4: Even though we filtered the stopwords using nltk corpus, we see significant number of words having little or no information i.e. stopwords and therefore a more comprehensive stop words lexicon was needed and used. This brought some improvement over the initial topics but still not enough to consistently summarize each review.

What we could do further to improve the topics.

1: Some of the research on the topic extraction show that “Noun” words and phrases are more indicative of the underlying topic than the other POS. Therefore it would be interesting to see if applying topic modelling techniques over this curtailed corpus brings any improvement.

2: Even though we used only LDA due to lack of enough time, there are range of topic modelling techniques that could be tried and evaluated.