## 6. With reference to the data (\*i.e.\* using numbers, figures, maps, and descriptive statistics), what does an analysis of Hosts and Listing types suggest about the nature of Airbnb lets in London?

### 6.1 Why should we choose the textual information?

Many studies[@xiao:2016] have analyzed various aspects of Airbnb listings, including price, spatial distribution, room type, etc. However, the "textual description", with more impressive potential than numeric fields, also plays a crucial role in shaping renters' first impressions of the listings, contributing to facilitating successful rental transactions. Additionally, hosts would also focus on the feedback of market demands to adjust their descriptions in response to policy requirements or economic trends.

Therefore, we have the reasonable motivation to scrutinize the textual features/characteristics from the data, aiming to generalize, classify and summarize some insightful conclusion. After correlating the insights with rental potential value, we hope to obtain valuable information about the short-term rental industry based on boroughs or wards as the basic geographical units.

### 6.2 What can we dig from the textual information?

Datasets consists of two textual fields: ‘Description’ and ‘Amenities’ both from the host’s subjective statement as self-promotion. ‘description’ column is some sentences describing listing’s advantages xxxxxx. ‘Amenities’ column is bunch of facilities and amenities inside of affiliated with the listing.

After some cleaning and preprocessing of the dataset, there are two set of questions corresponding to the two columns respectively.

1. Which topics would host like to focus on when promoting their properties?

We could use the LDA model to generalize topics and get the most frequent keywords in those topics. Firstly, we need to calculate iteratively the coherence of the LDA model with the number of topics ranging from 1 to 40, in order to determine the most appropriate number of topics for summarizing the hosts' textual descriptions (Figure 1).

Then, word cloud shows that among 16 topics, there are xxxxxxxxxxxxx.

2. Do the listings in the same neighbourhood, or with the same spatial location, share the similar amenities?

Amenities 有着高度的分类特性，这意味着许多的amenities有着相似性，例如“xxxxxxxxxxxxx”，因此我们需要从海量的词汇中找出不同amenities的相似性，并将他们进行合适的分类。

We use the Word2Vec model to classify 打量打量的words and phrases. Each of them would be presented as a multi-dimensional vector. 之后通过UMAP方法对他们进行降维至二维(Figure 2)。在图中，每个点都代表着某个房源的amenities特征，而颜色相似的点意味着这两个房源的ameinites特征高度相似。Afterwards, we 重新将这些点按照他们的实际地理位置放回map中。这样我们就可以知道，对于某个特定的区域或社区，区域内的房源是有着高度的同质化特征（颜色高度相似）还是异质化特征（颜色较为杂乱）。

### 6.3 Which indicator guide the branding?

Branding and recommendation system of Airbnb platform aims to xxxxxxx to make more money. Yet in terms of community and regulation, Airbnb should xxxxxxxx. Therefore comes the value question: what indicator could represent the potential opportunities for listing’s branding or promotion?

We multiple the price of each listing by its total nights in the last year, total nights was calculated by minimum nights, maximum nights and number of reviews. 虽然technically this is an approximate number, but it aligns with the data from the Inside Airbnb. Afterwards, we compare ‘sum\_income’ with the average property value to get an integrated index to indicate this listing’s ‘cost-benefit ratio’.

### 6.4 How does the indicator correlate with textual information?

"How are these textual features correlated with the composite index X? What kind of textual features positively contribute to enhancing the composite index X?"

"What are the key words in the existing property descriptions? Which textual features are beneficial in improving the composite index X?"