Interim Report

Supervisor: Giacomo Livan

Fazaan Hassan

06/02/2019

Project Title

The Use of Statistical Natural Language Processing Techniques to Rationalise the Evolution of Airbnb Reviews

Progress to Date

Literature Review:

Summarising papers to understand current NLP techniques, methodologies and preprocessing techniques to conduct linguistic evaluation.

General Preprocessing:

Removal of Non-English words, Removal of stop words, creation of unigrams on a per review basis, tokenisation and duplicate word removal.

Unsupervised Learning:

Implementation of Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorisation (NMF) for topic modelling and extracting words which relate to each corresponding topic.

Statistical Metrics:

Implemented the following; Type Token Ratio (TTR), Average Review Length and Word Frequency (WF)

Work Iterations:

- Iteration 1
    - TTR, WF, LDA and NMF for all data, with no preprocessing and sorting
- Iteration 2
    - Reviews sorted by date, 2011-2019, We have cumulative TTR, Unique Word Count and WF
- Iteration 3
    - TTR and NMF for 8 Boroughs
- Iteration 4
    - Average Review Length for all Dataset sorted by date
    - Review Length for 4 main boroughs, followed Mean, Standard Deviation, Max and Min length
- Iteration 5
    - Looking at 3 specific properties in the Kensington and Chelsea Borough.

To Do:

Segment reviews based on prices of properties and select top and bottom 25% of reviews. Precompute prior metrics and see if we can deduce whether there is a shift in language.