Airbnb Reviews

NLP/Unsupervised Learning

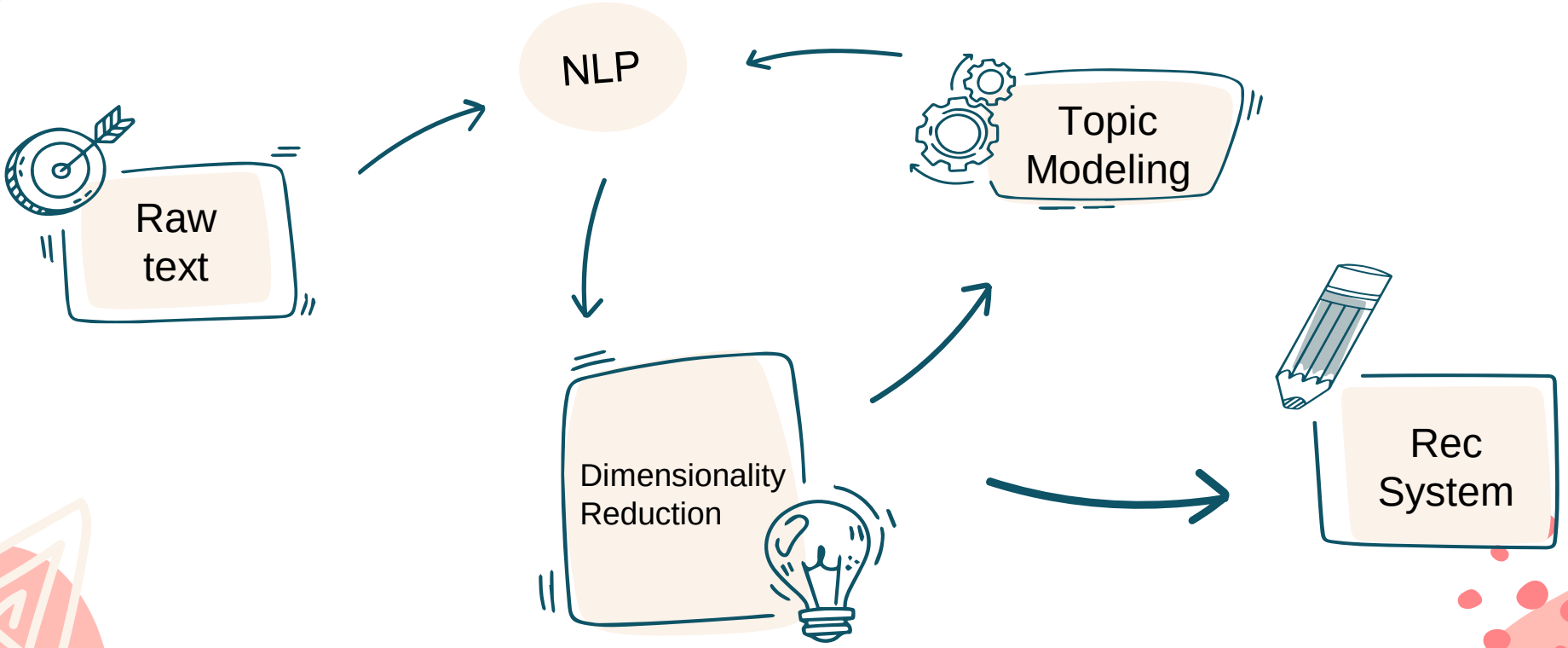
Objective

Applications:

1. Topic modeling with reviews can help Airbnb improve customer experience

2. The reviews recommendation system let's Airbnb find listings that have the most similar reviews

Workflow


NLP
Raw text
Topic Modeling
Dimensionality Reduction
Rec System

Methodology

Data

- Inside Airbnb
- NYC 2009-2021
- 700k data points

Tools

- NLTK
- Gensim
- SVD
- pyLDAvis
- sklearn

Models

- CoreX
- LDA
- NMF

Recommendation System

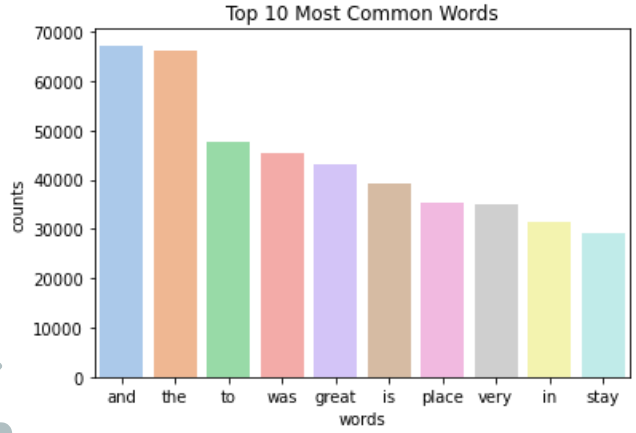
- Content-based

Text Preprocessing

Before

48

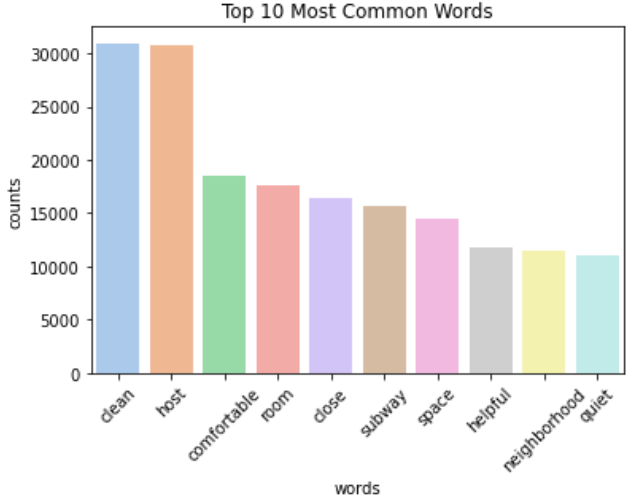
Average length of document



After

16

Average length of document



Topic Modeling

- NMF/LDA – 3 topics
- **CoreX** – 5 topics

Anchor words: interior, distance, issues, host, clean

Location		subway, restaurants, distance, walk
Booking Logistics		host, automated, cancel, posting
Host Communications		new, questions, comfortable, available
Rooms		**kitchen, bathroom, bed, room**
Cleanliness		**clean, sparkling, tidy, spotlessly**

Recommendation System

- Tf-IDF → SVD → Rec System

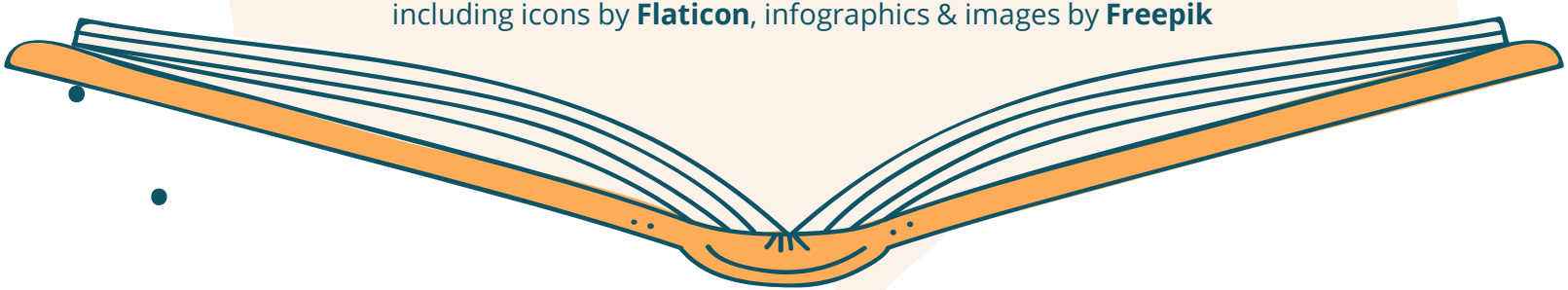
Test review: I really like it she was nice and the places was organize and clean

Future Work

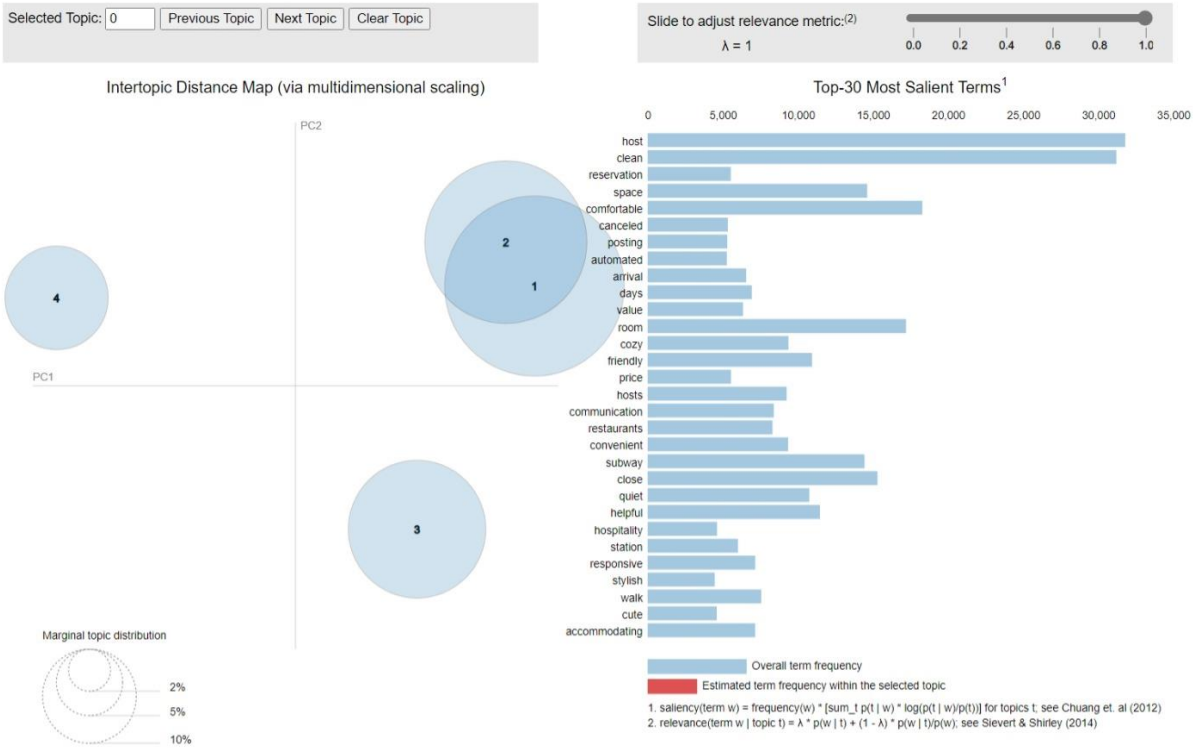
- Obtain more information from Airbnb for further analysis
- Build out recommendation system to include more components

THANK YOU!

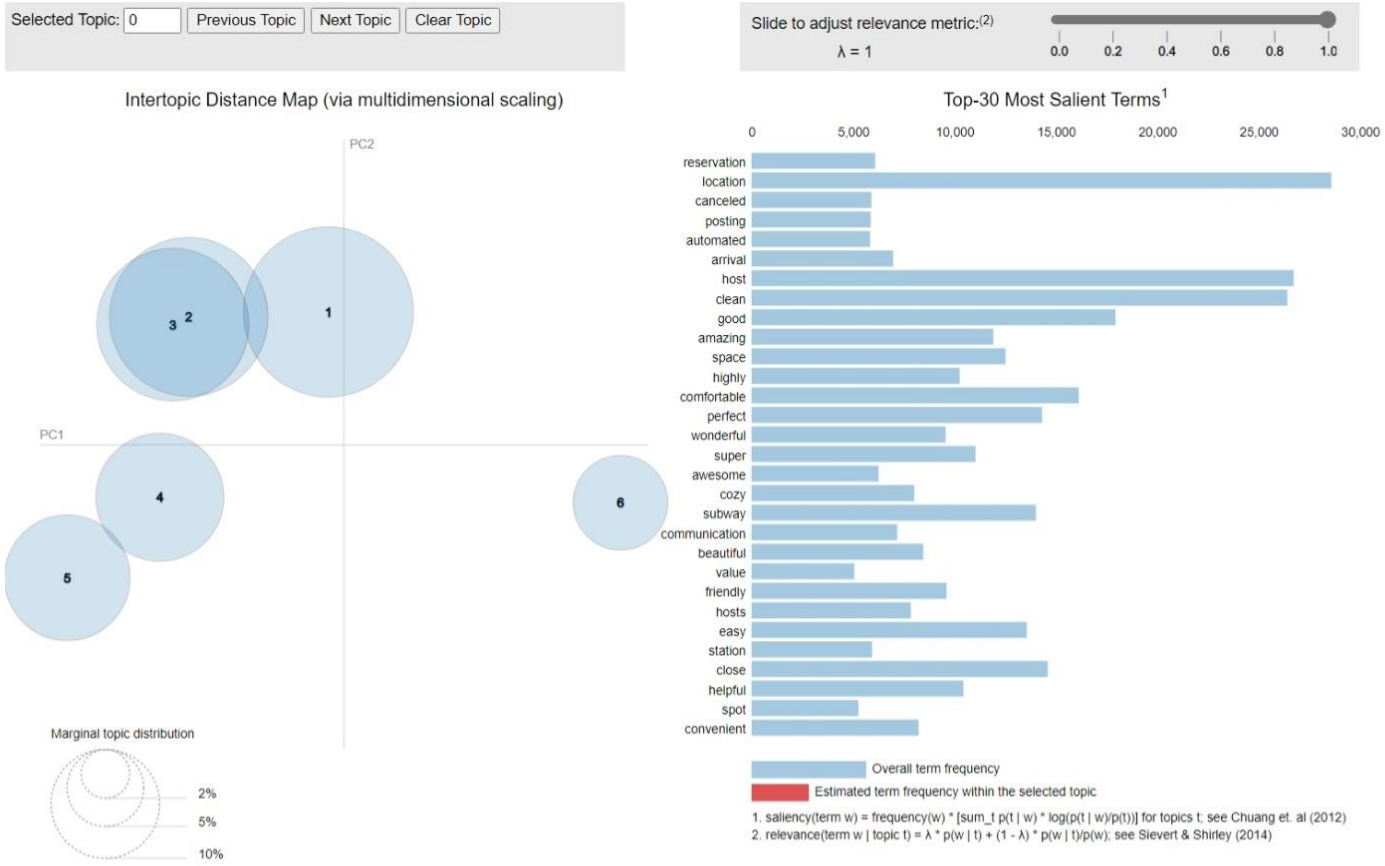
Does anyone have any questions?



Appendix



LDA 4 topics


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0 5,000 10,000 15,000 20,000 25,000 30,000
reservation
location
canceled
posting
automated
arrival
host
clean
good
amazing
space
highly
comfortable
perfect
wonderful
super
awesome
cozy
subway
communication
beautiful
value
friendly
hosts
easy
station
close
helpful
spot
convenient
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

LDA 6 topics