Airbnb Reviews

NLP/Unsupervised Learning

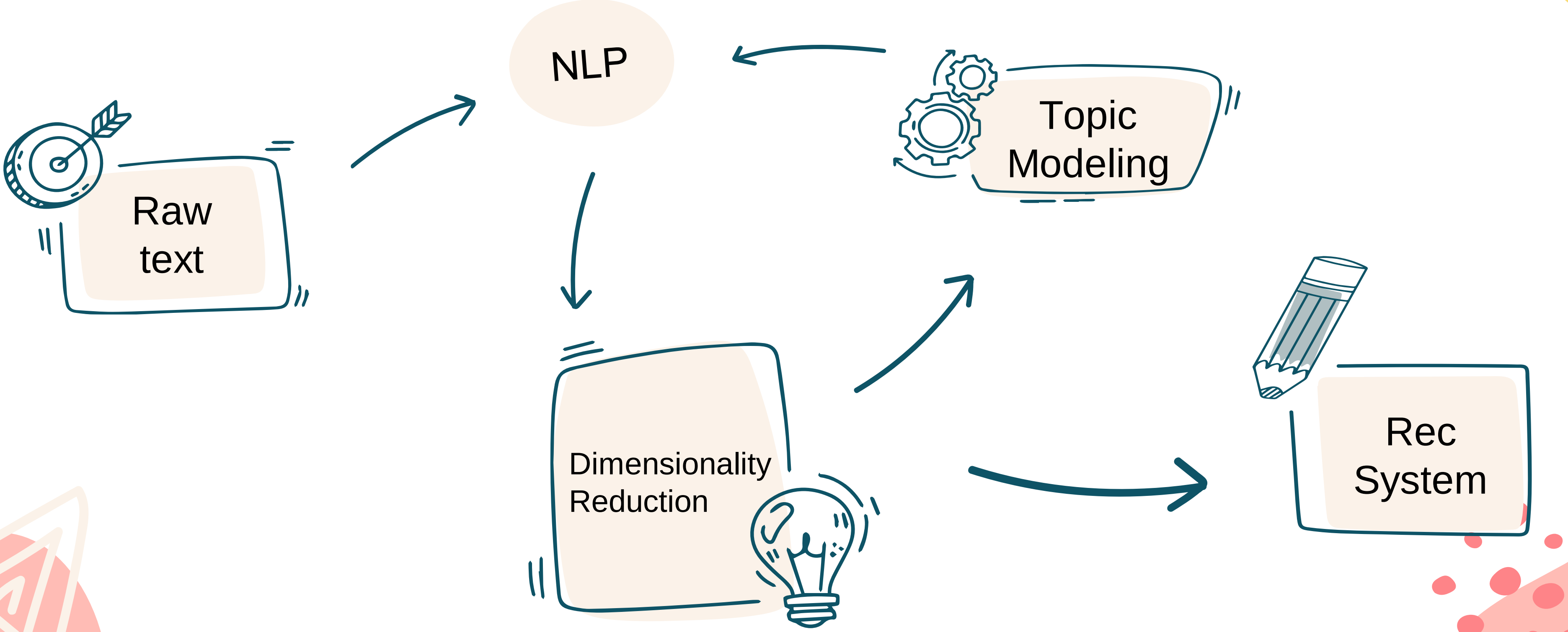
Objective

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



Raw text

NLP

Dimensionality Reduction

Topic Modeling

Rec System

Methodology

Data

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

Tools

- NLTK
- Gensim
- SVD
- pyLDAvis
- sklearn

Models

- CoreX
- LDA
- NMF

Recommendation System

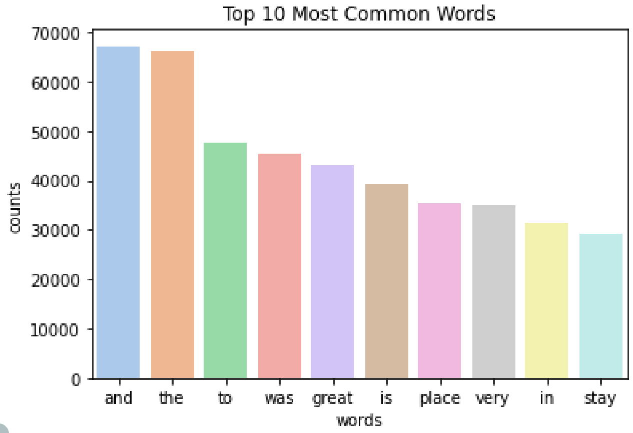
- Content-based

Text Preprocessing

Before

48

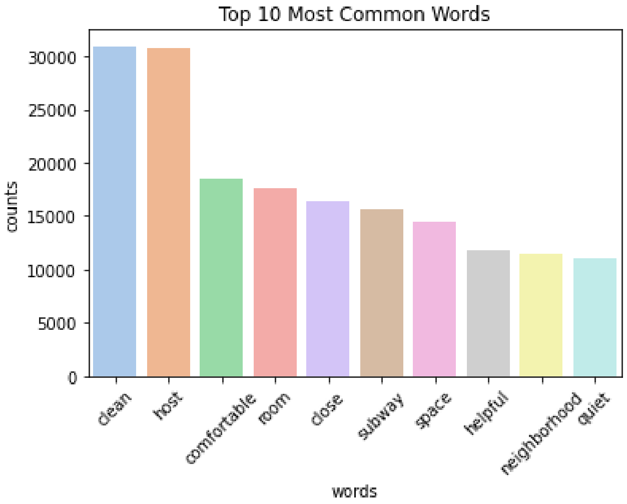
Average length of document



After

16

Average length of document



Topic Modeling

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

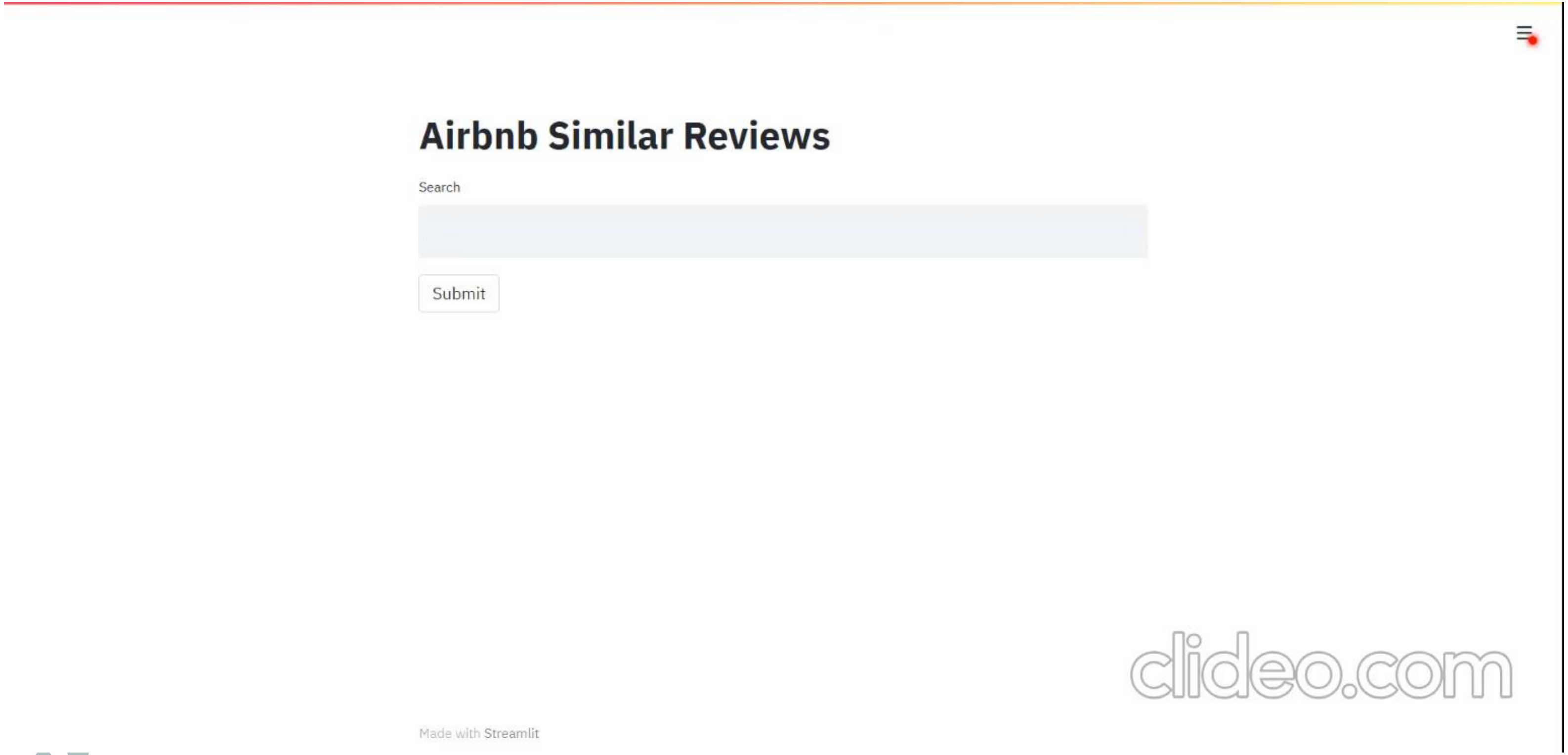
Anchor words: interior, distance, issues, host, clean

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



Airbnb Similar Reviews

Future Work

○ Obtain more information from Airbnb for further analysis

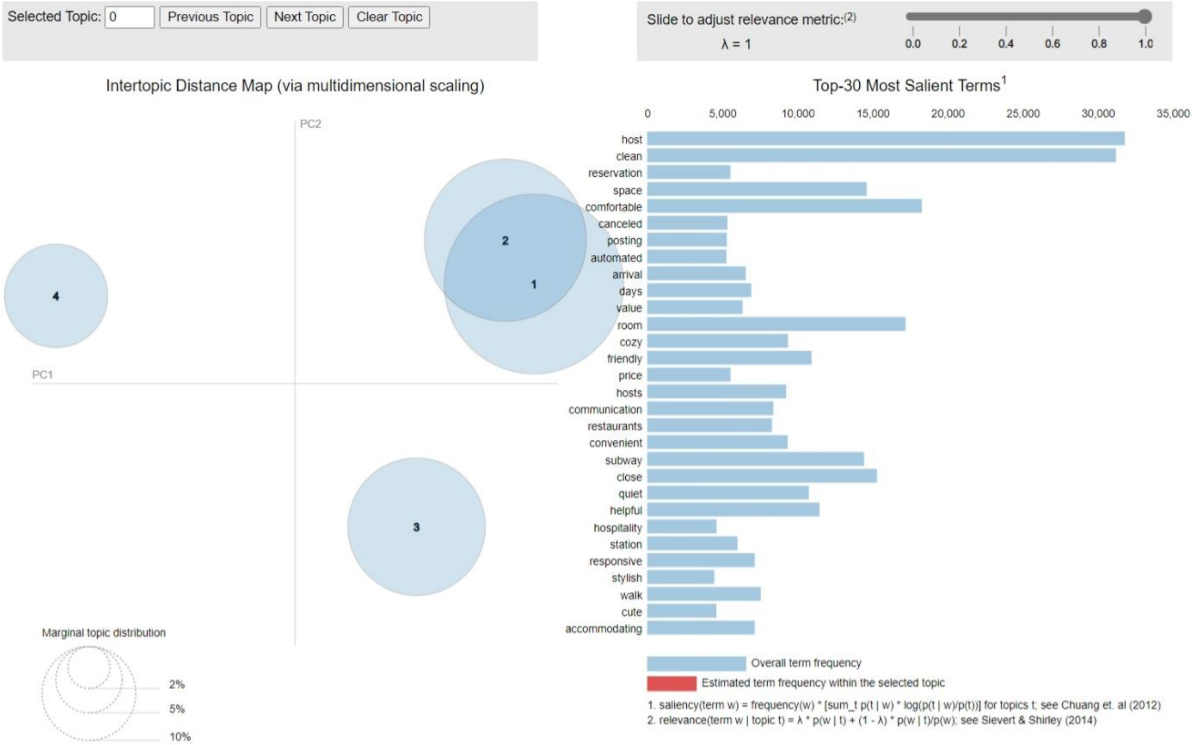
○ Build out recommendation system to include more components

THANK YOU!

Does anyone have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

Appendix

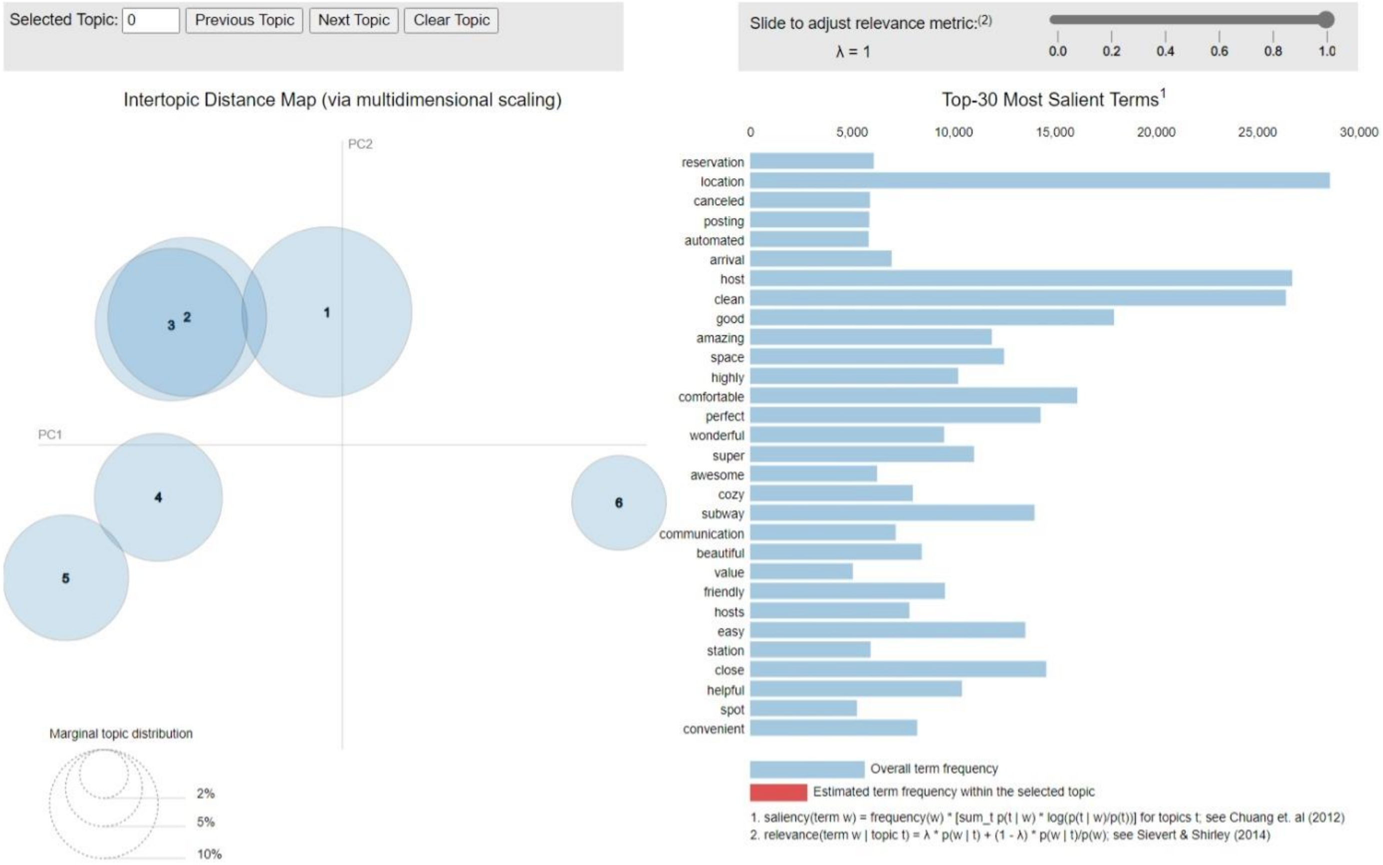


Selected Topic: 0

Slide to adjust relevance metric:(2)
λ = 1

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 4 topics



Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 1

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