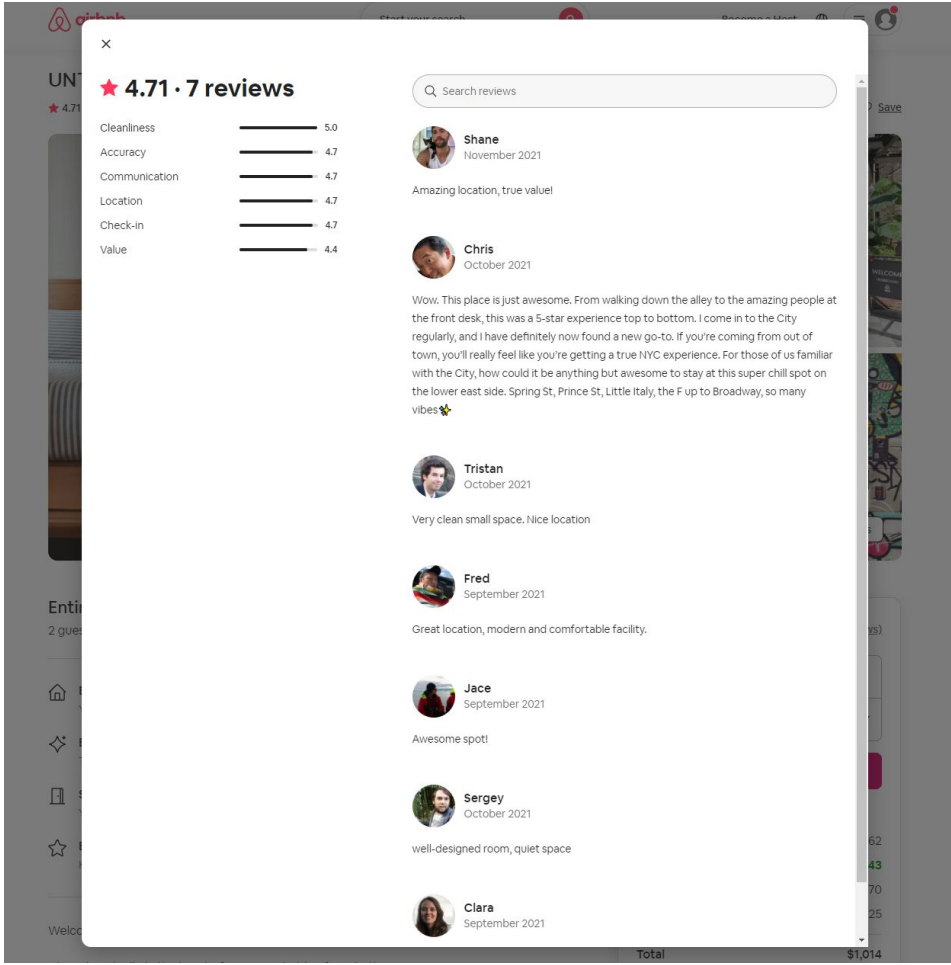
Topic Modelling for Reviews, Sources Affecting Guests' Positive and Negative Experiences

Gulay Samatli Pac

Why reviews matter?

- Reviews **help guests** choose their travel plans wisely and enable hosts to open their homes with confidence and **attract guests**.



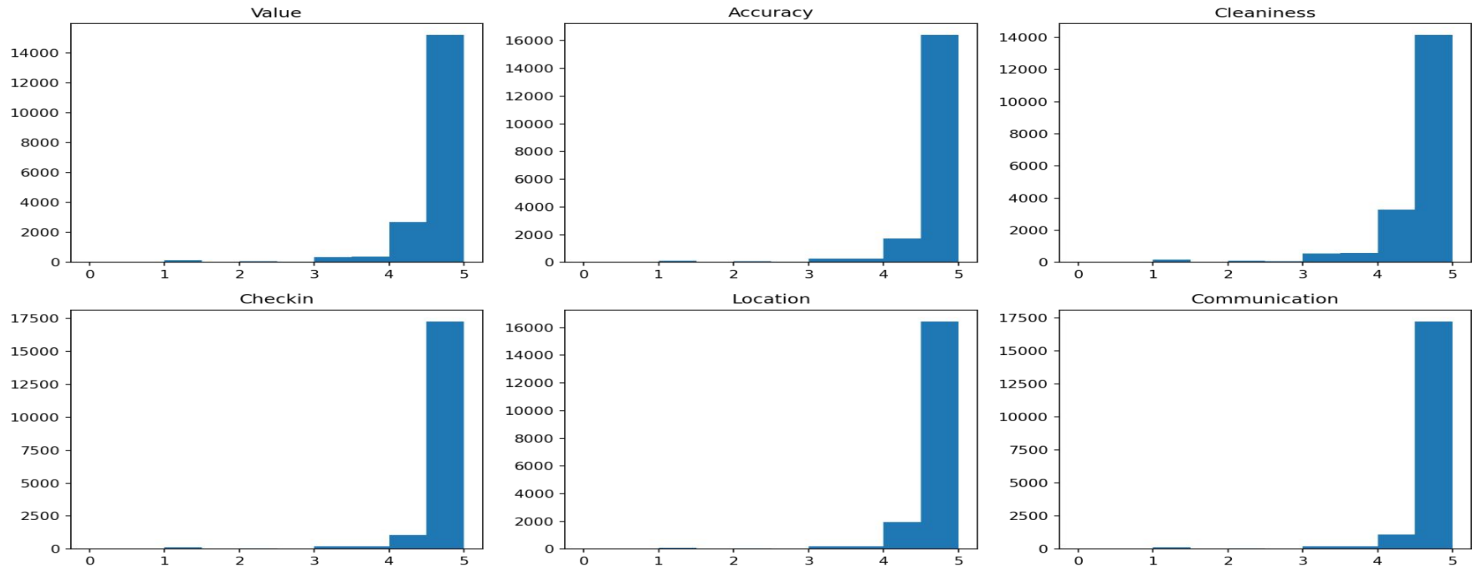
Why reviews matter?

- The first thing to avoid when booking an Airbnb:

  Little or No Reviews

What about ratings?

- Guests can provide star ratings with 1–5 stars on specific aspects of their experience.
- The problem:
  - All review scores are highly positive scores.
  - There is no scores less than 4.5 out of 5(see below graphs).



Methodology

Data

- NYC Listings Dataset, InsideAirbnb

- More than 80K Reviews

Data Cleaning & EDA

- Remove numbers, capital letters and punctuations

- Eliminate non-English reviews

- Lemmatize

- Tools: Pandas, Numpy, langdetect, NLTK, Matplotlib, Seaborn

Topic Modeling

- TF-IDF Vectorizer
- CountVectorizer
- Models :
    - NMF
    - SVD
    - LDA
    - CorEx

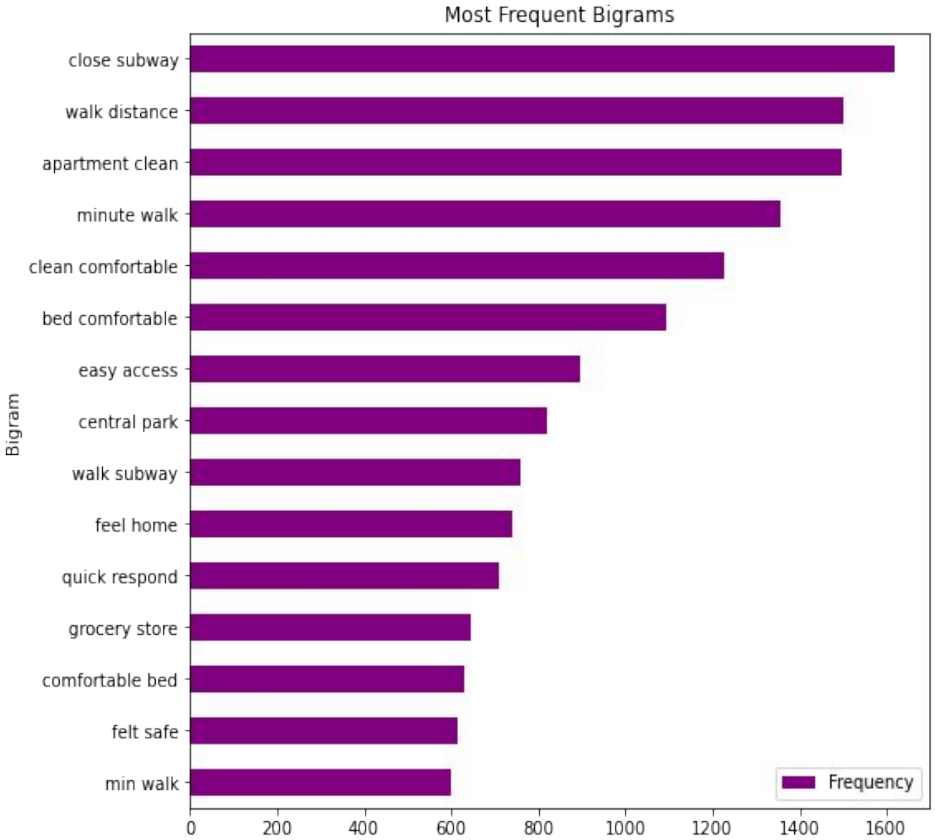
- Tools : sklearn, pyLDAvis

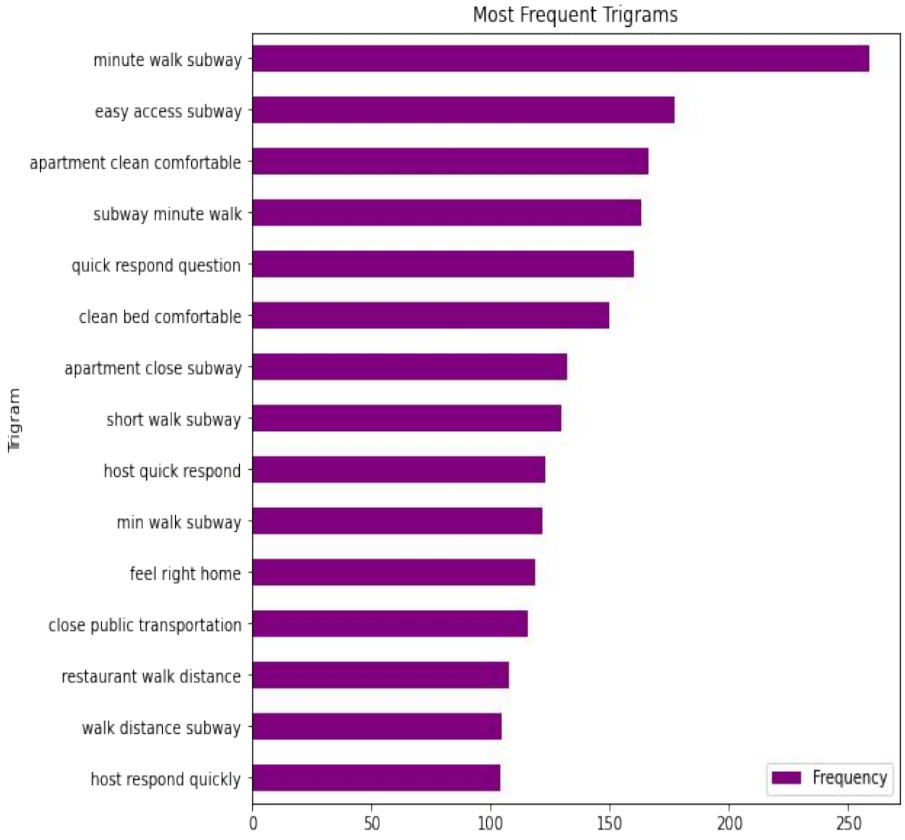
Final Model

- Count Vectorizer (ngram_range=(1,2), max_df = 0.7, min_df = 10)
- LDA
- VaderSentimentAnalysis

- Tools : sklearn, pyLDAvis

Reviews: Most common words in reviews





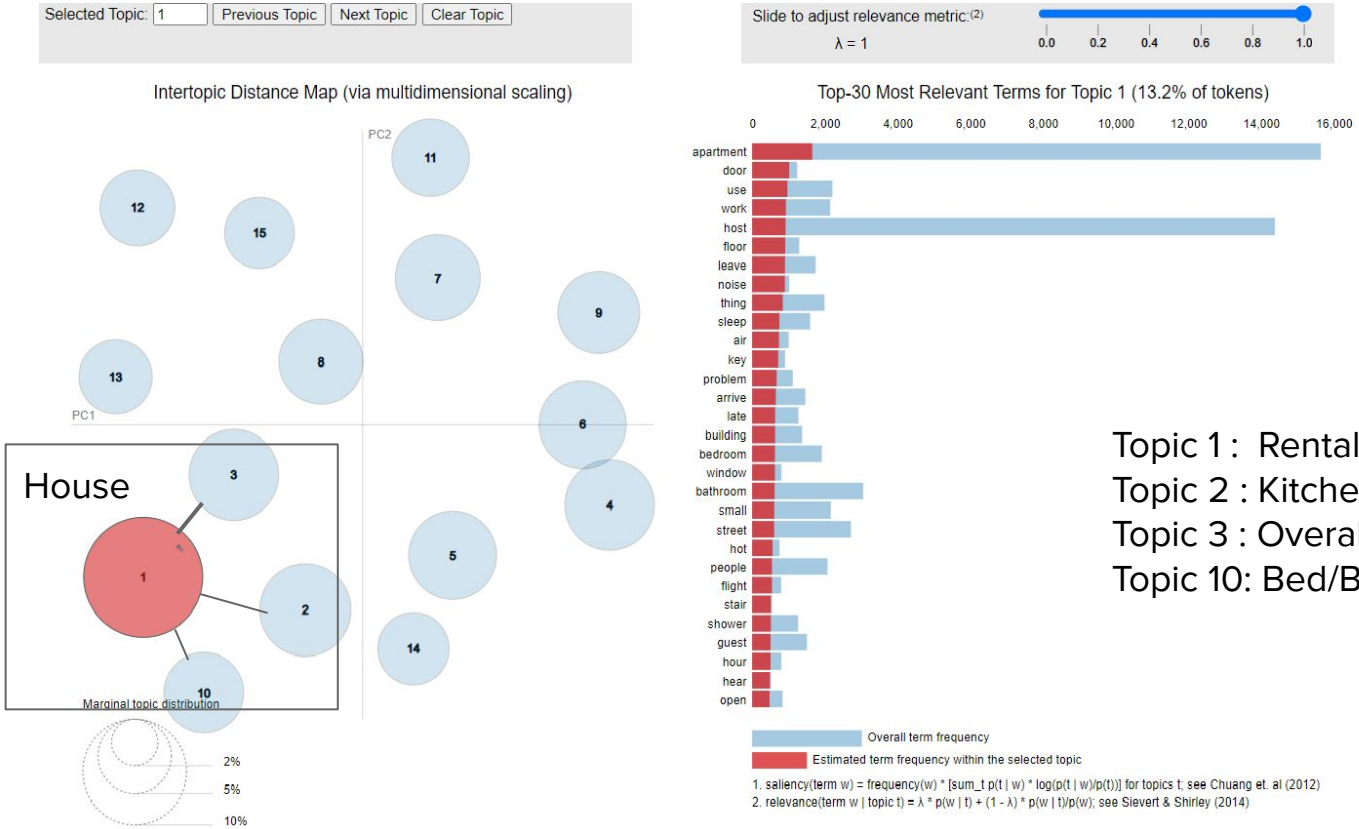
Topic Modelling

- Vectorizer :
    - CountVectorizer
- Topic Modeler :
    - LDA
- Number of Topic 15

Topics

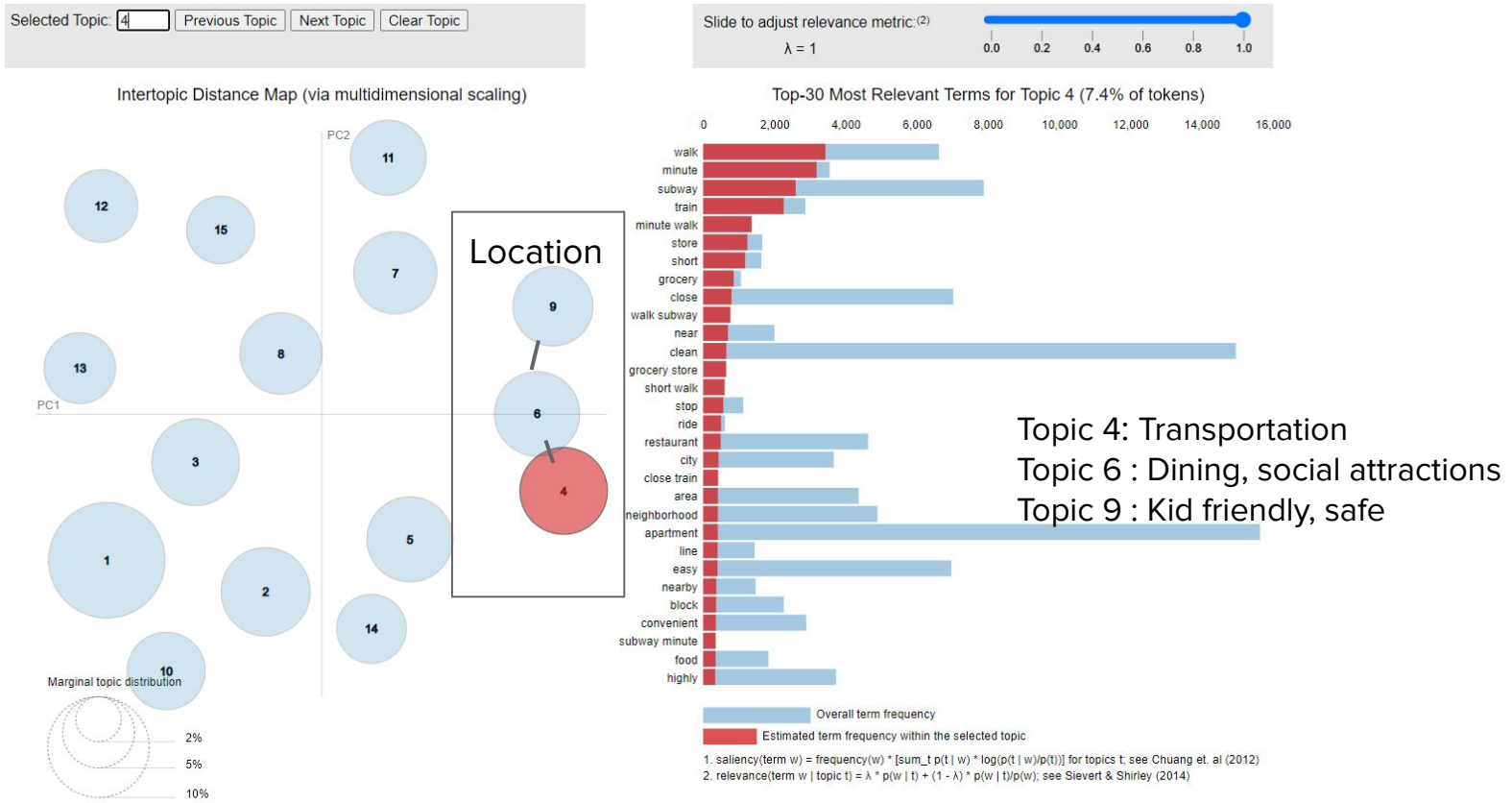
🏠 Rental interior
🍳 Kitchen
Overall Airbnb Experience
🚇 Neighborhood- accessibility to transportation
👫 Home-like comfort/experience
🍷 Neighborhood/ accessibility to dining, social attractions
🧼 Cleanliness
🛎️ Host-hospitality
🚸 Location- safe/family friendly
🛏️🛁 Bed/Bathroom
😄 Overall trip experience
📲 Host-responsiveness
🔑 Convenience (check in/out, comfort, hotel like)
💰 Comfort/Value
🏡 Listing Accuracy

Topic Visualization



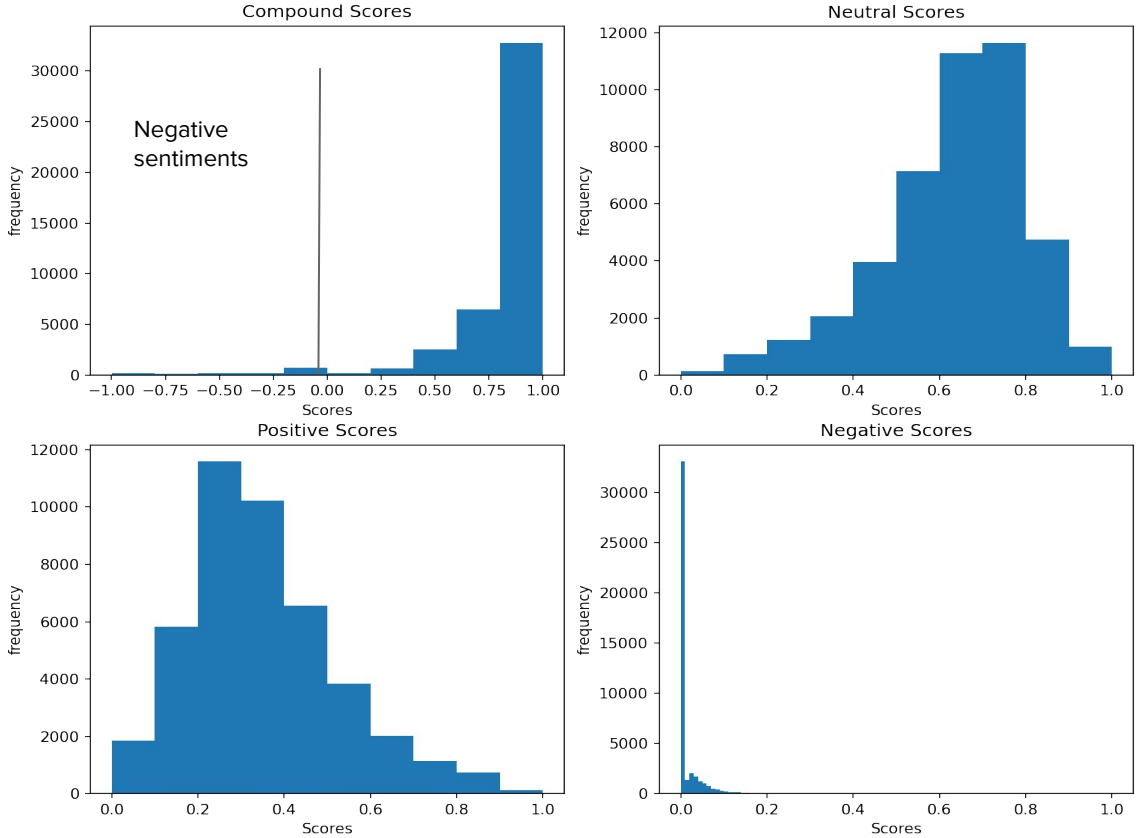
Topic 1 : Rental interior
Topic 2 : Kitchen
Topic 3 : Overall experience
Topic 10: Bed/Bathroom

Topic Visualization



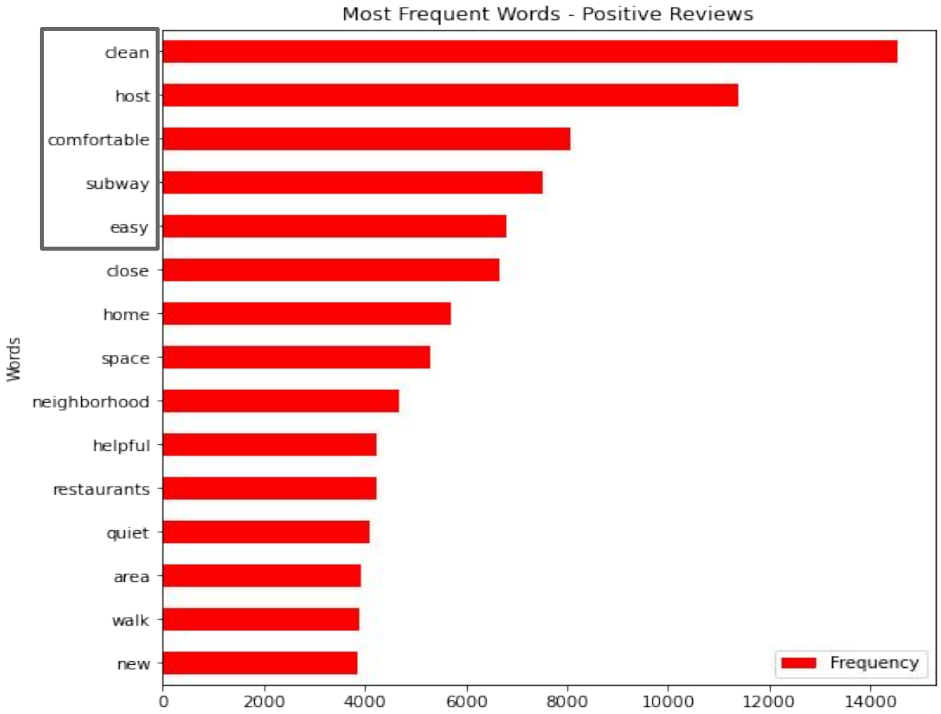
Topic 4: Transportation
Topic 6 : Dining, social attractions
Topic 9 : Kid friendly, safe

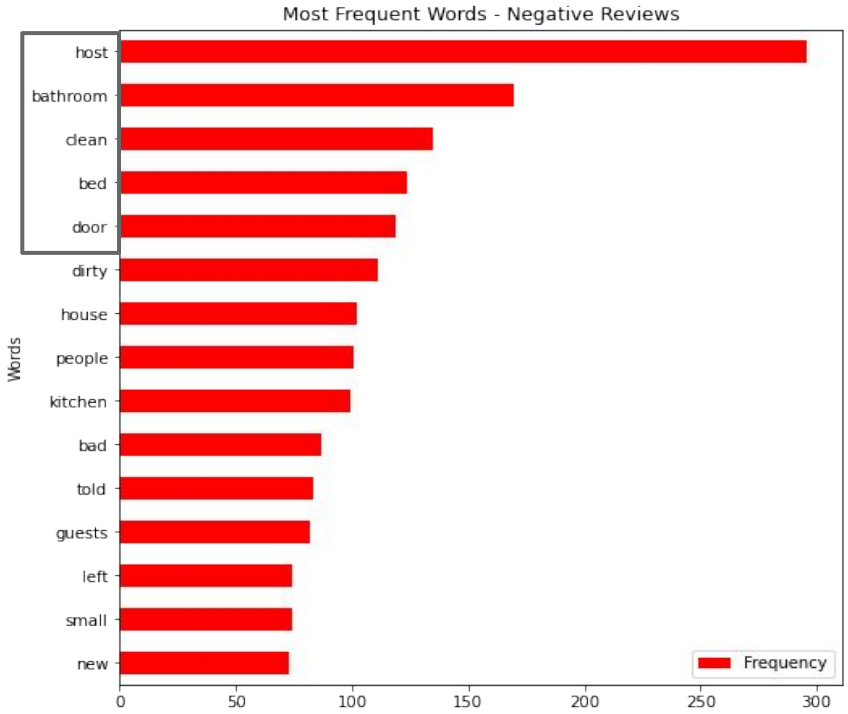
Reviews: Sentiment Analysis/ VaderSentiment



- Most guest had pleasant experiences during their Airbnb stays.

Positive vs Negative Reviews





Host interaction, Cleanliness ➜ Overall experience

Insights for Hosts

Common topics in reviews

- The condition of the house
  - Kitchen, bedrooms, and bathrooms
- Cleanliness
- Location
- Host responsiveness and hospitality

Next Steps

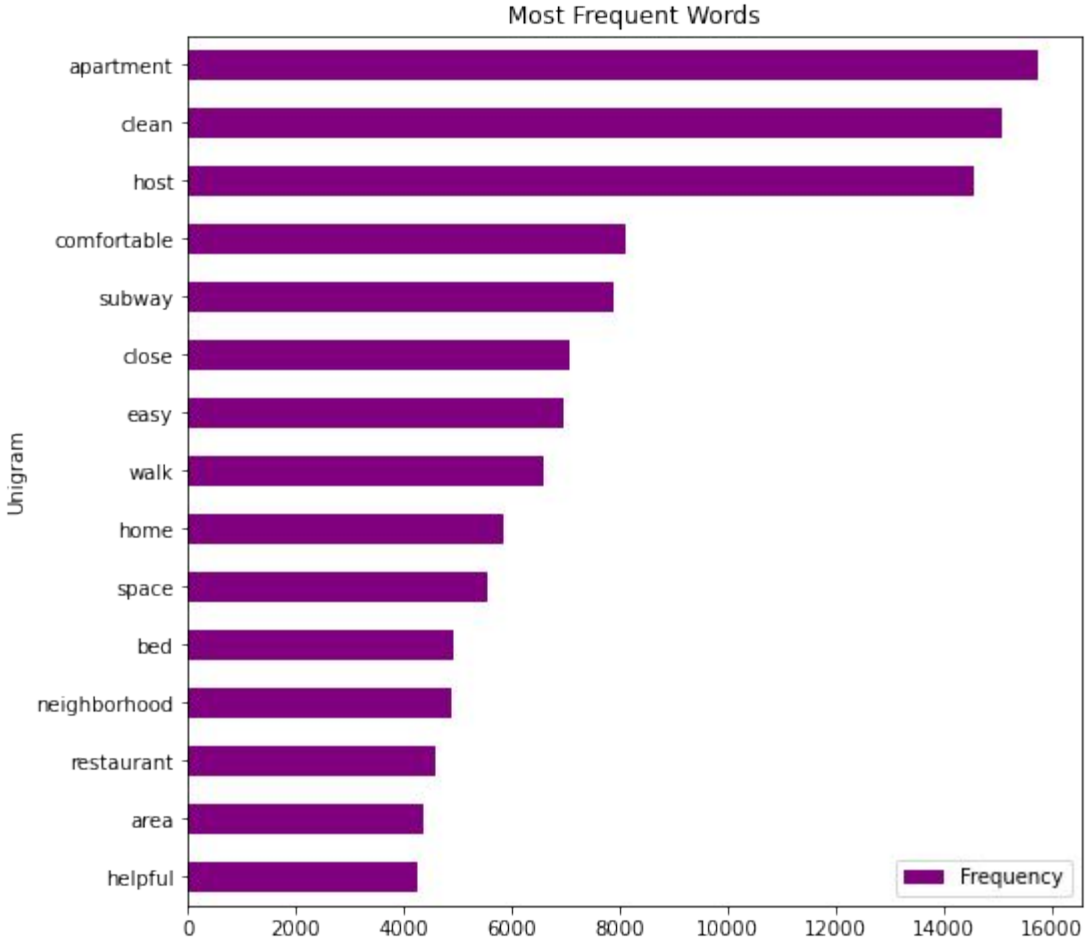
- Cluster analysis to group topics
- Use keywords as a filtering options for the search tool
- Incorporate reviews as a feature to predict occupancy rates

Questions

Thank you!

Appendix

Reviews: Most common words in reviews



Review Topics

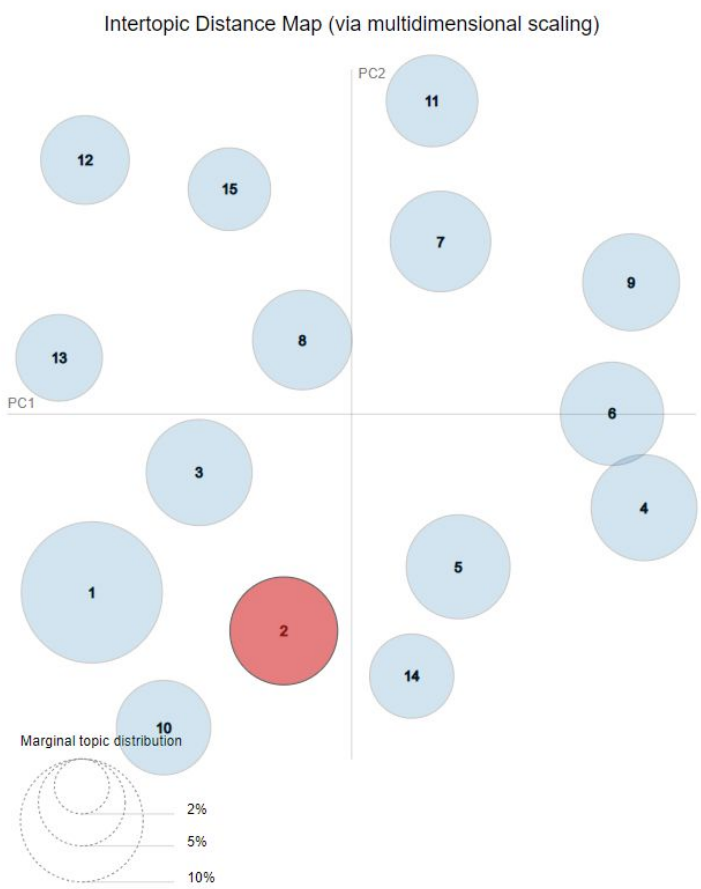
1 Apartment interior issues
2 Kitchen Experience
3 Airbnb Experience
4 Neighborhood- accessibility to transportation
5 Home-like comfort/experience
6 Neighborhood/ accessibility to social attractions
7 Cleanliness
8 Host-hospitality
9 Location- safety/family friendly
10 Bed/Bathroom
11 Overall trip experience
12 Host-responsiveness
13 Convenience (check in/out, comfort, hotel like)
14 Comfort/Value
15 Listing Accuracy

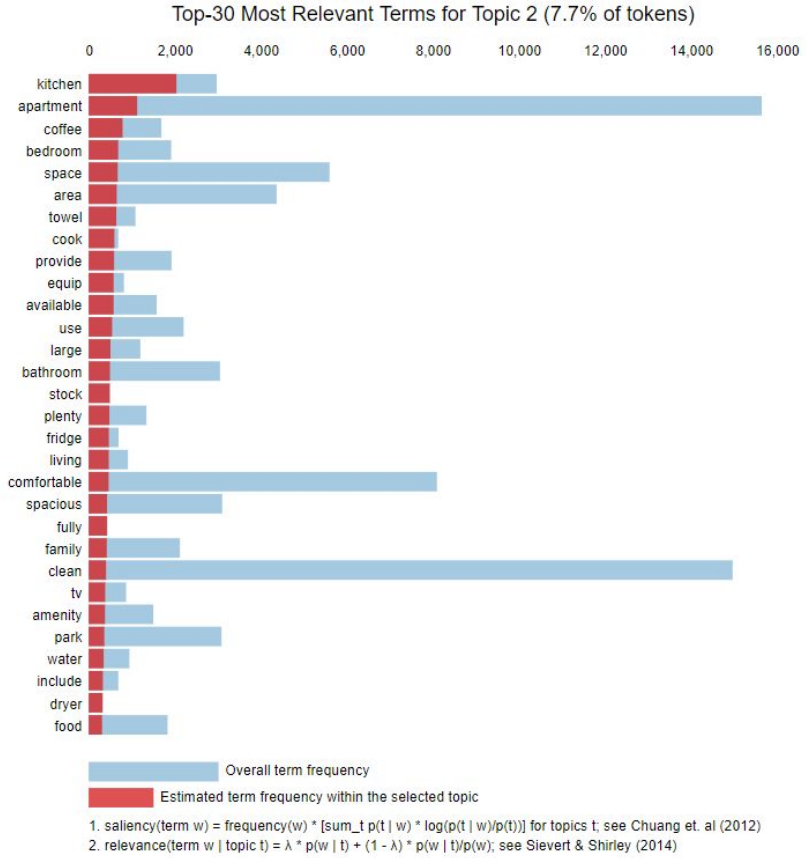
Selected Topic: 2 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

λ = 1

0.0 0.2 0.4 0.6 0.8 1.0





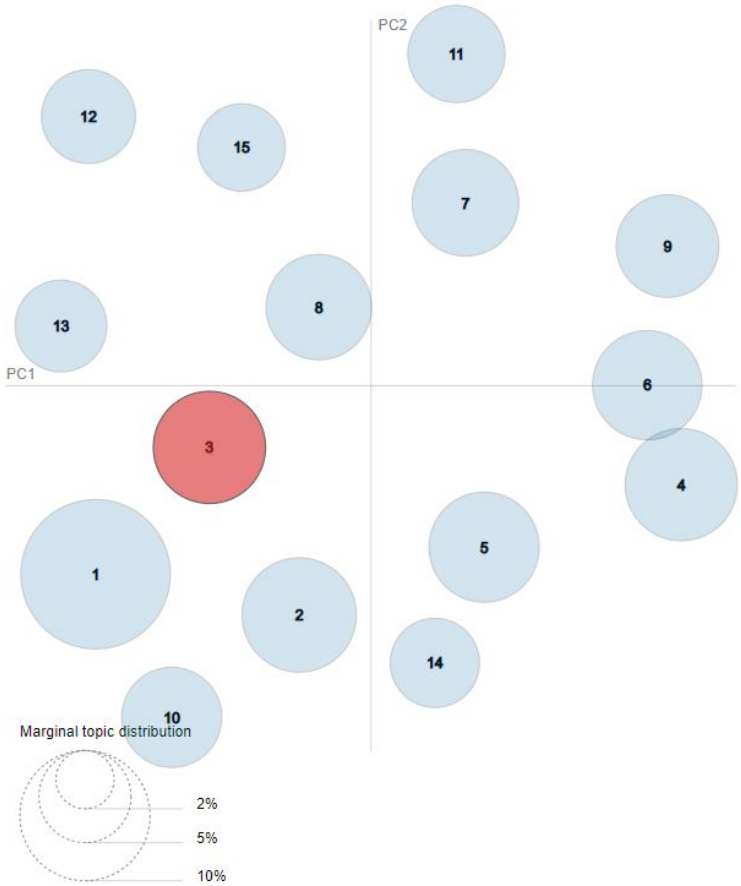
Selected Topic: 3 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

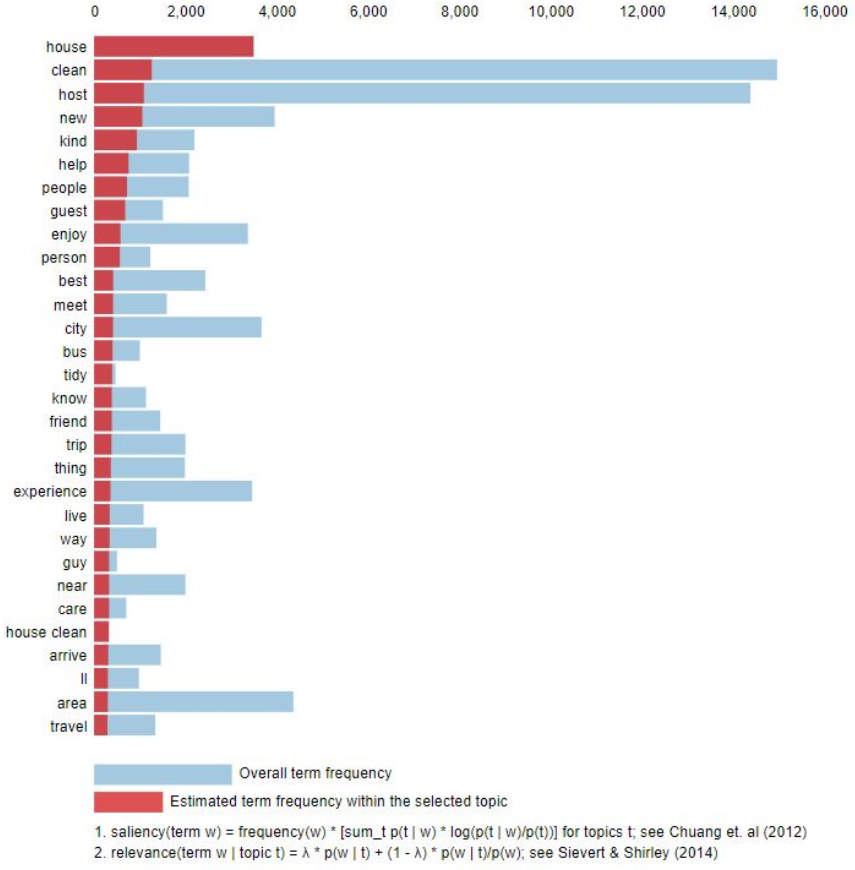
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 3 (7.4% of tokens)



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)

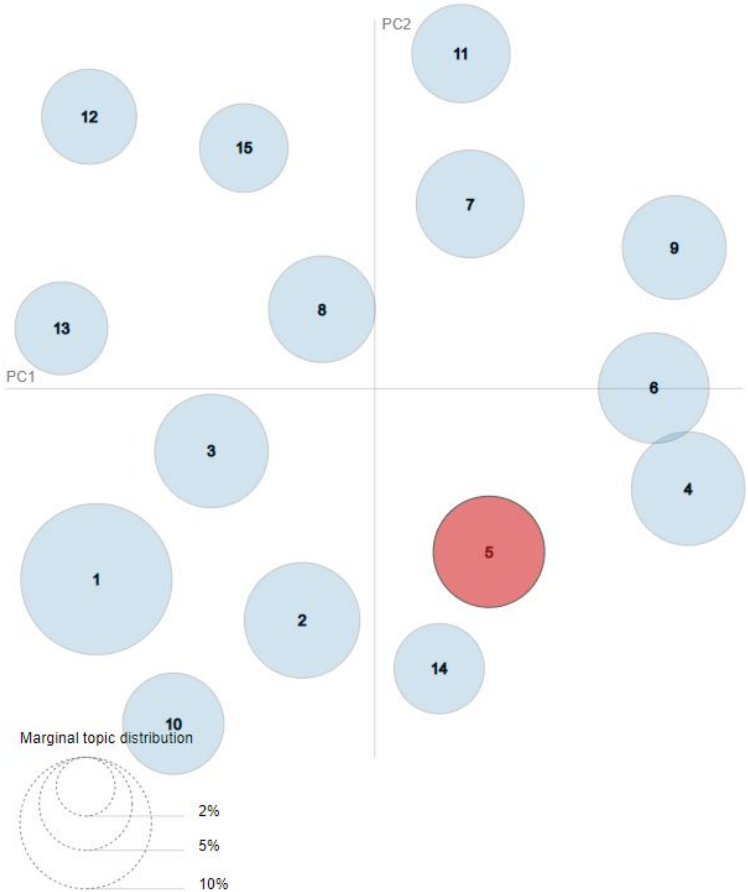
Selected Topic: 5 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

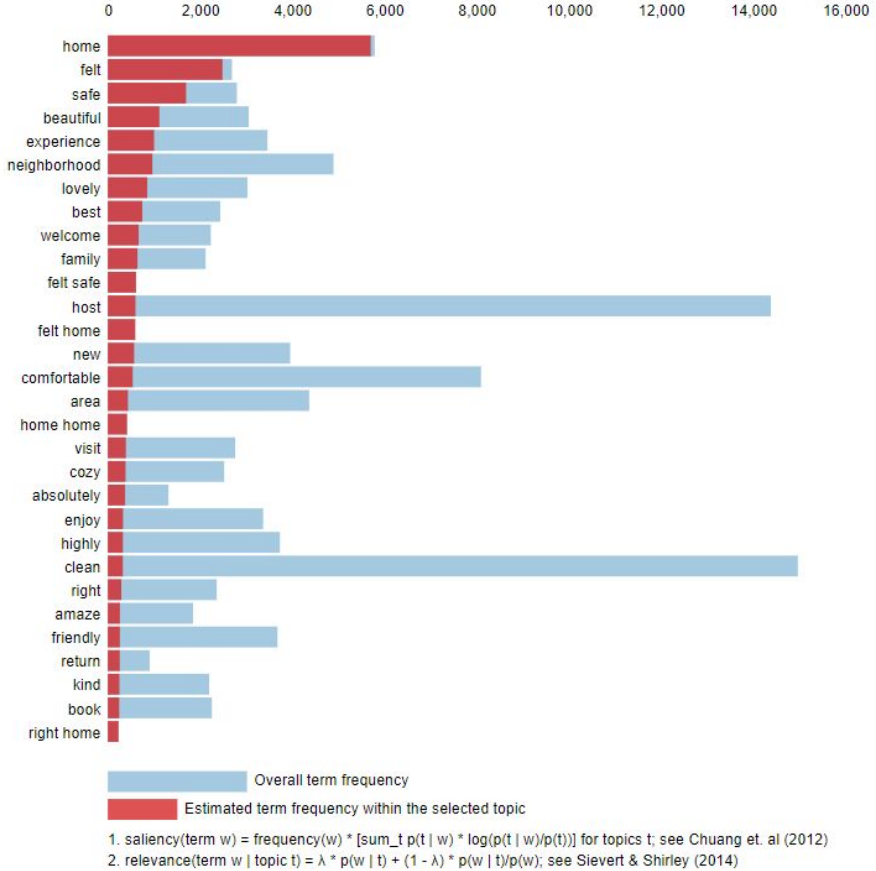
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

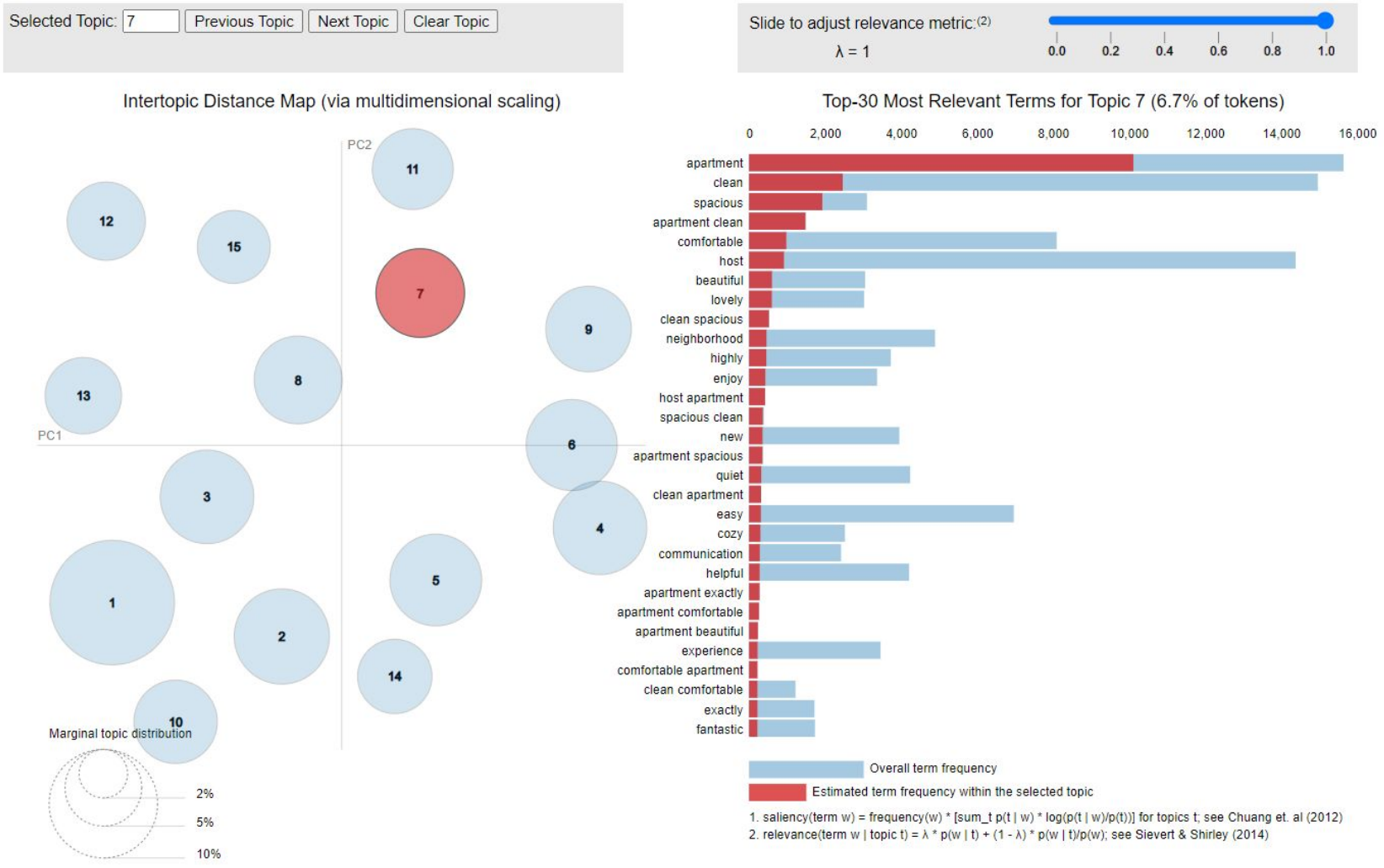
Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 5 (7.1% of tokens)



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)


Selected Topic: 7
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
11
12
15
7
9
8
13
6
3
4
5
1
2
14
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (6.7% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
apartment
clean
spacious
apartment clean
comfortable
host
beautiful
lovely
clean spacious
neighborhood
highly
enjoy
host apartment
spacious clean
new
apartment spacious
quiet
clean apartment
easy
cozy
communication
helpful
apartment exactly
apartment comfortable
apartment beautiful
experience
comfortable apartment
clean comfortable
exactly
fantastic
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)

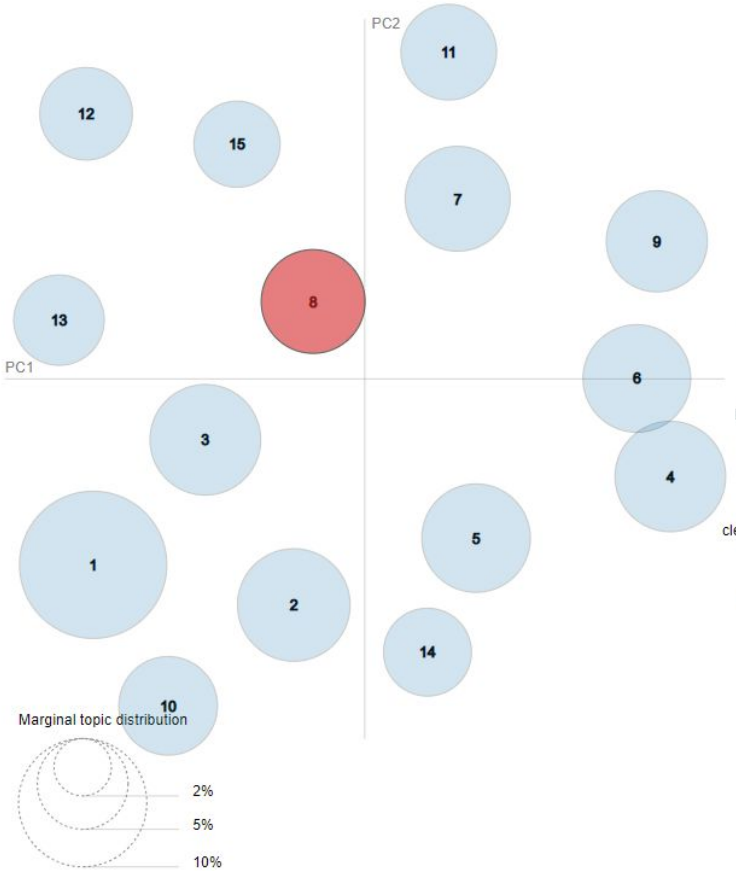
Selected Topic: 8 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

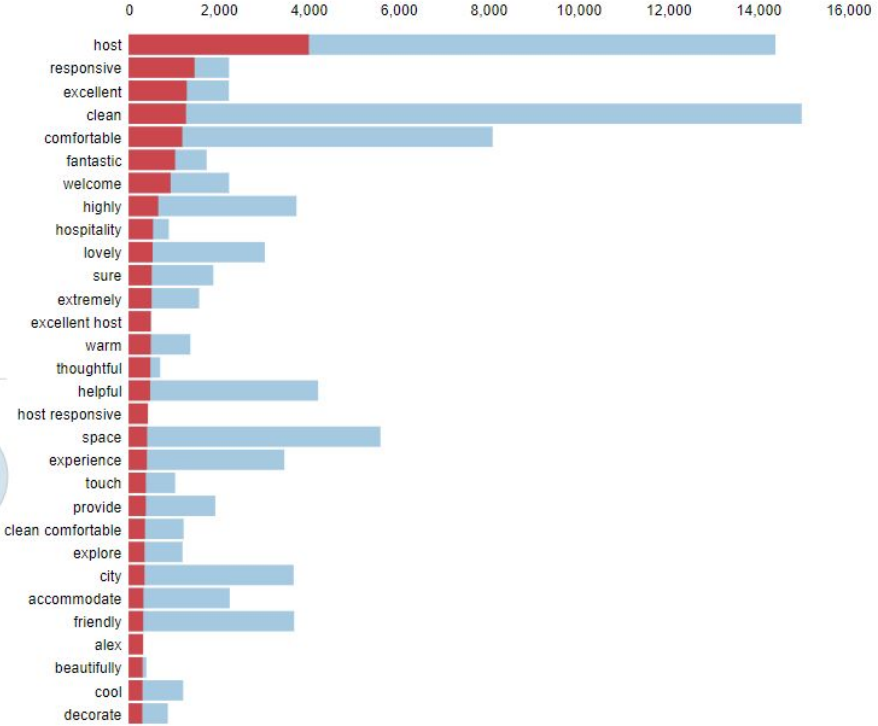
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 8 (6.5% of tokens)



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)

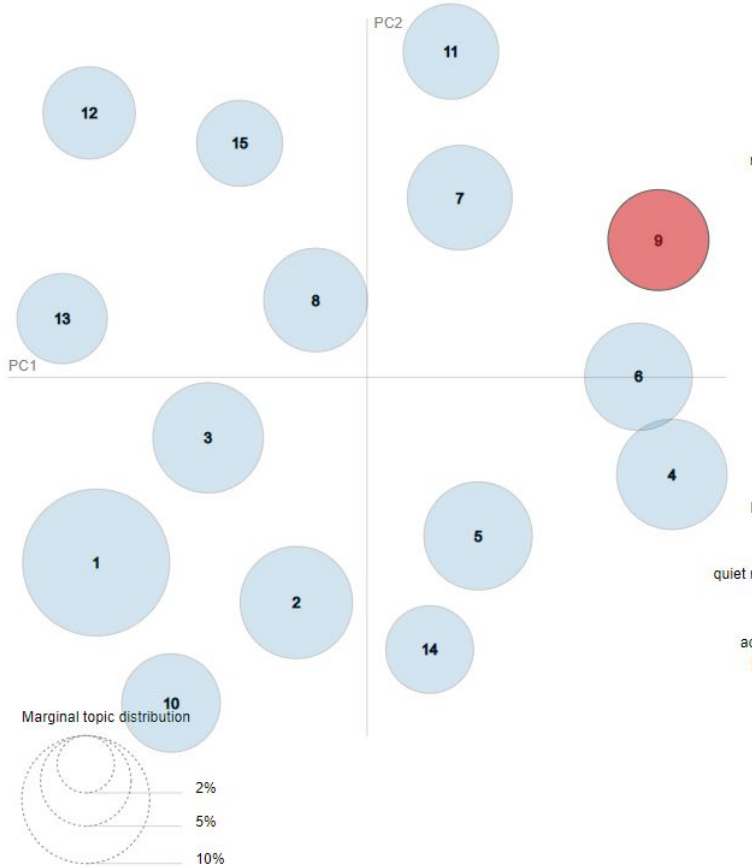
Selected Topic: 9 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

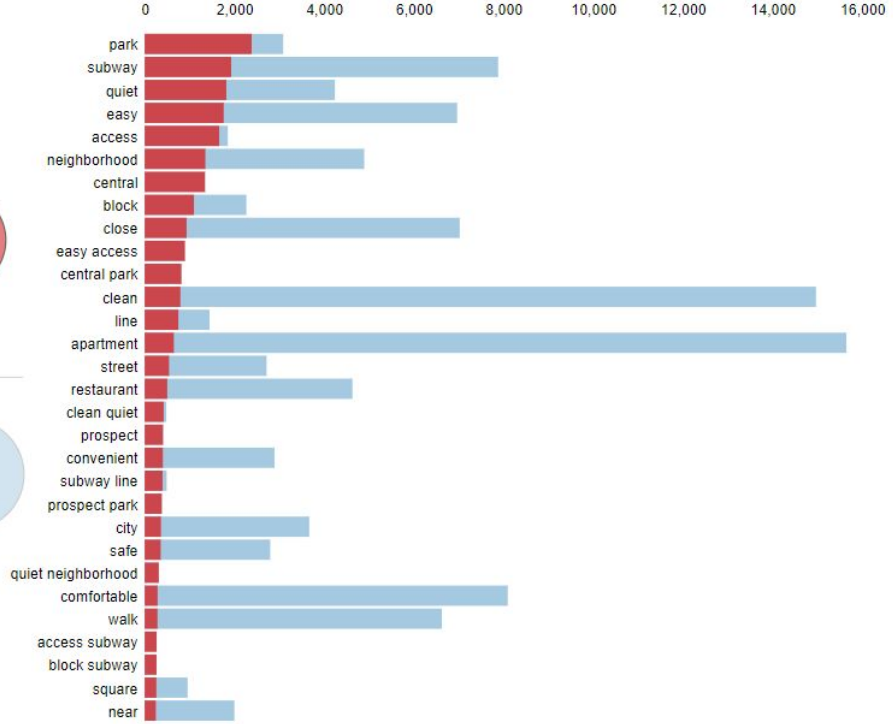
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 9 (6.2% of tokens)



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)

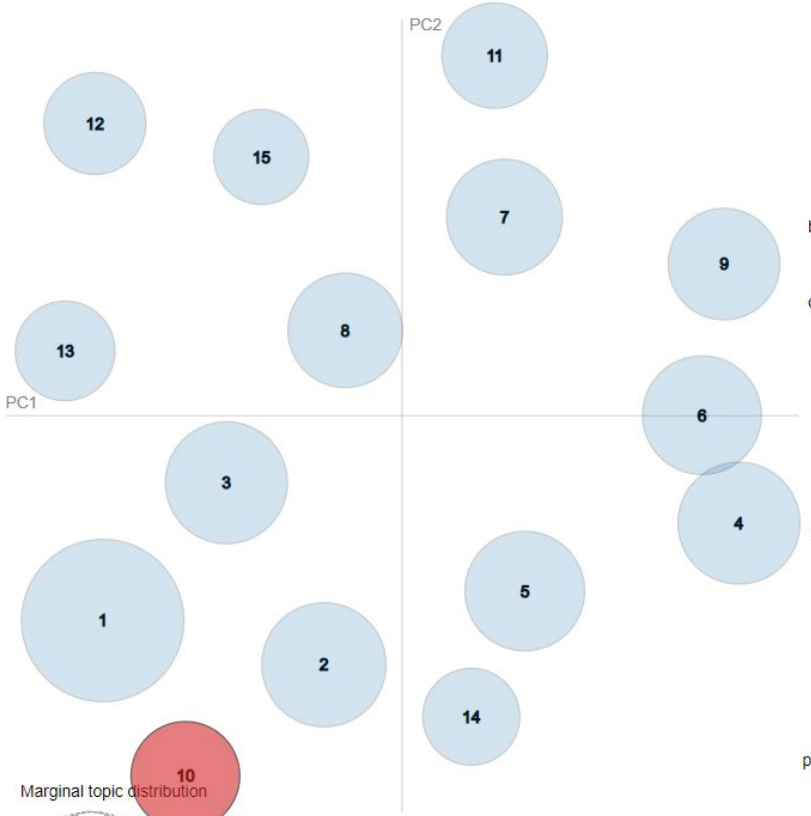
Selected Topic: 10 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)

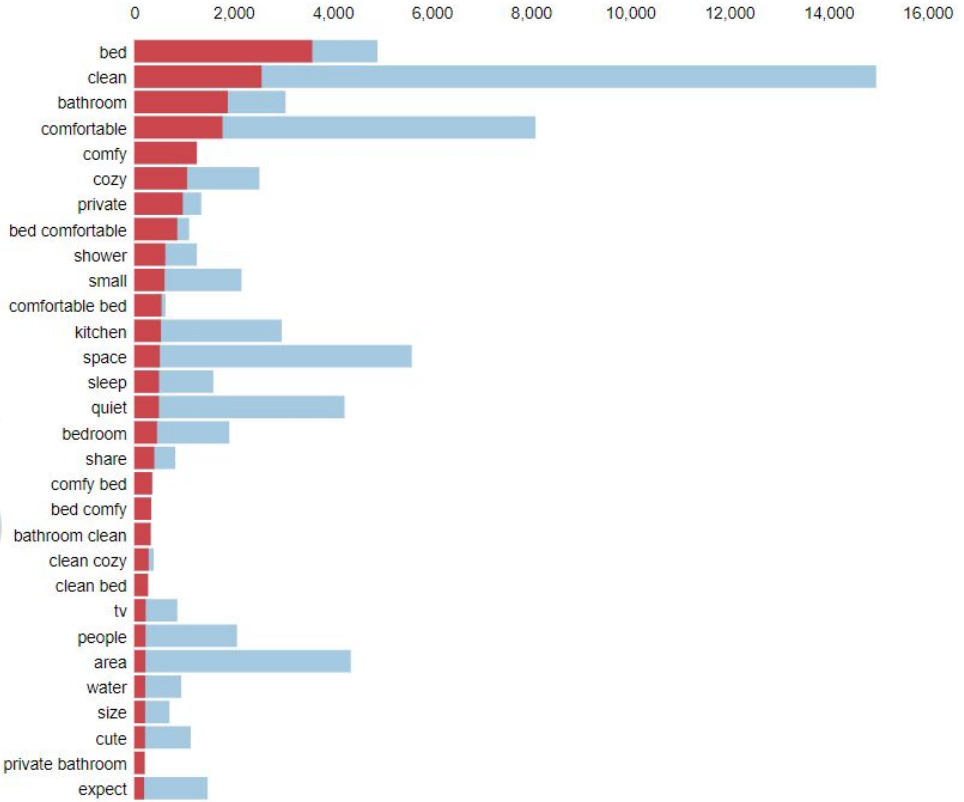
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 10 (5.9% of tokens)



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)

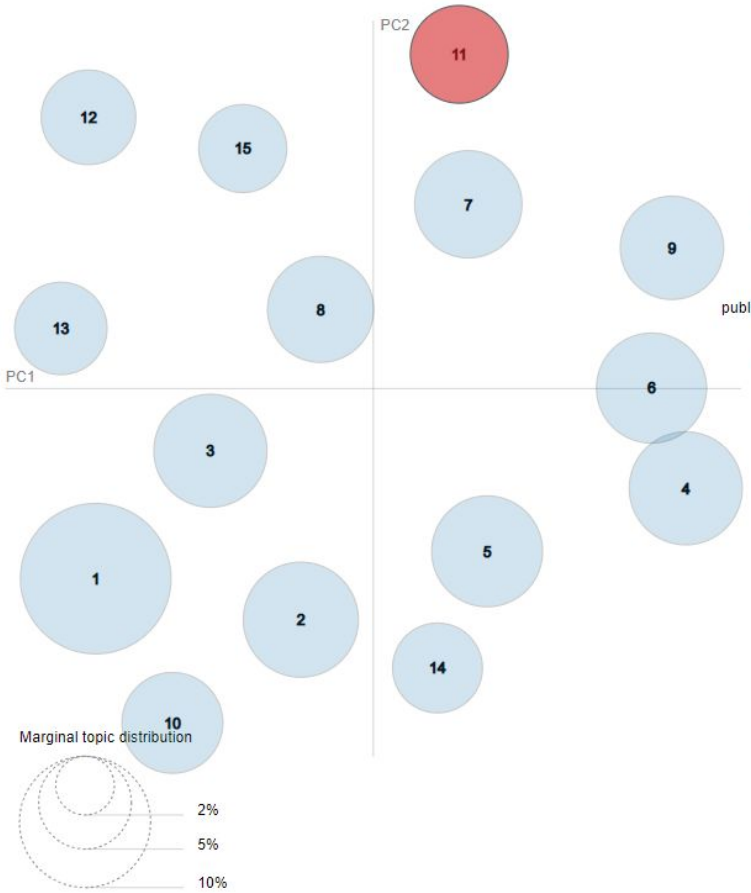
Selected Topic: 10 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

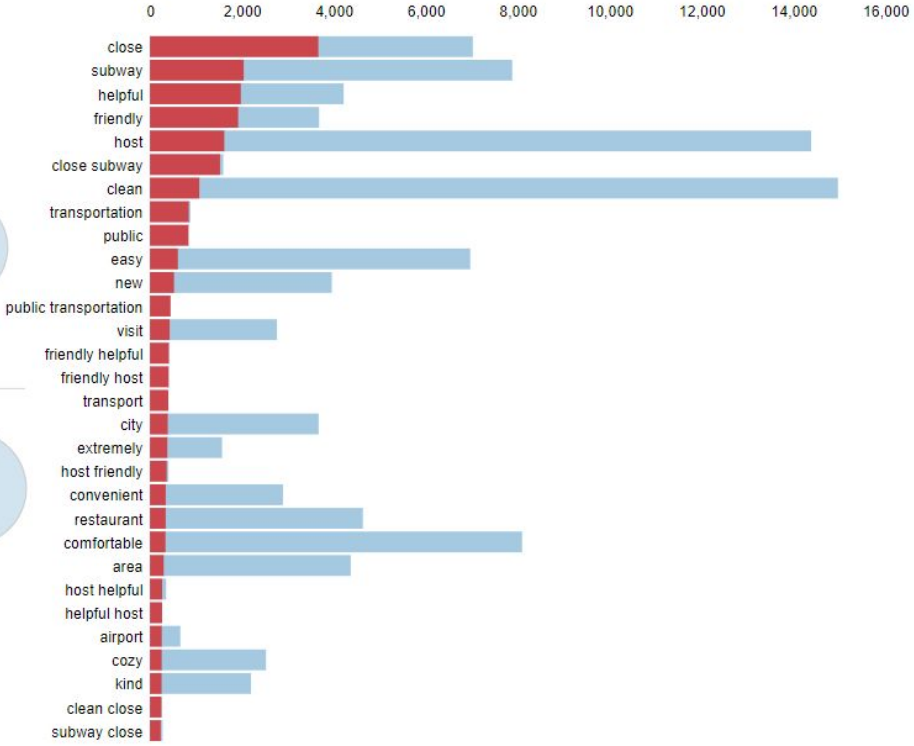
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 11 (5.5% of tokens)



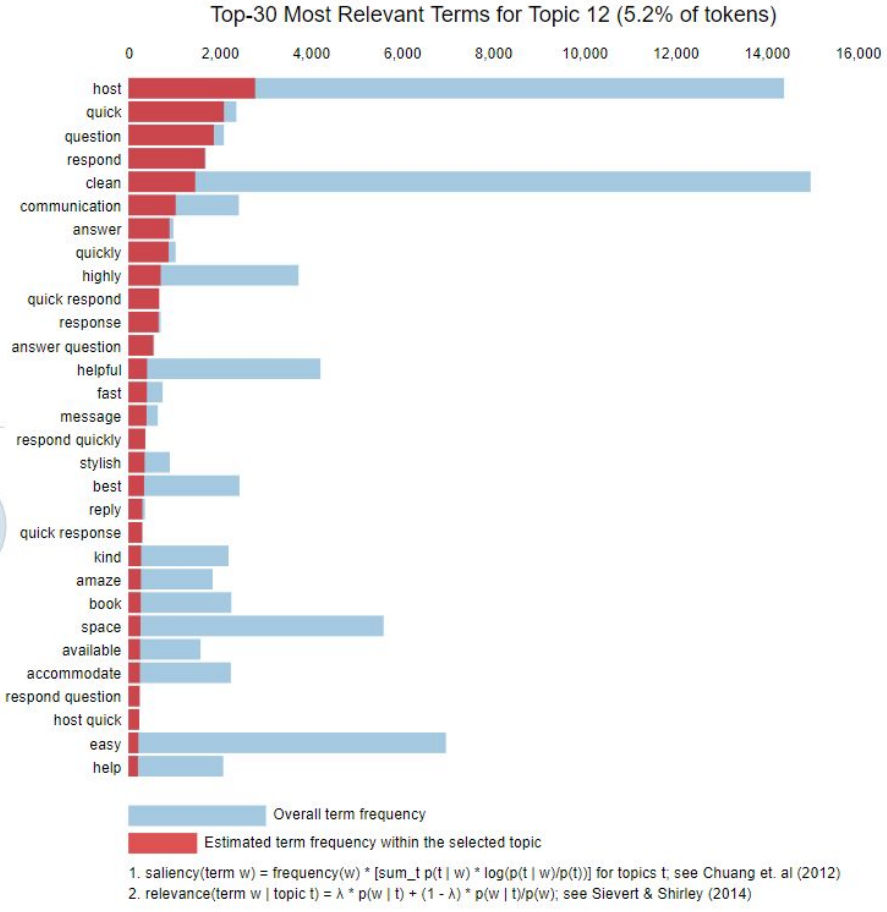
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Selected Topic: 12 | 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



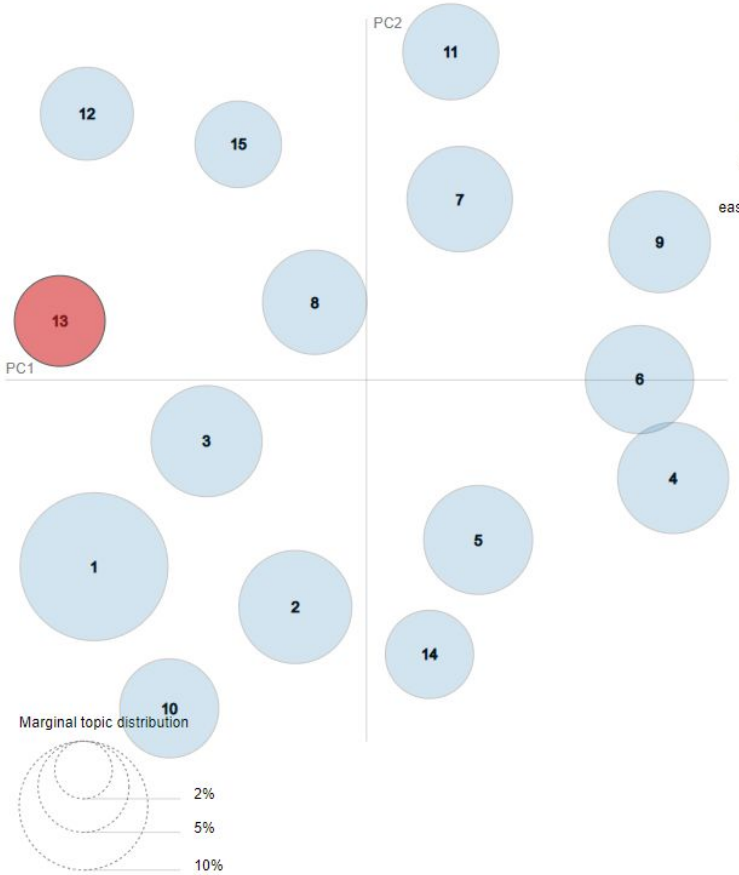
Selected Topic: 13 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

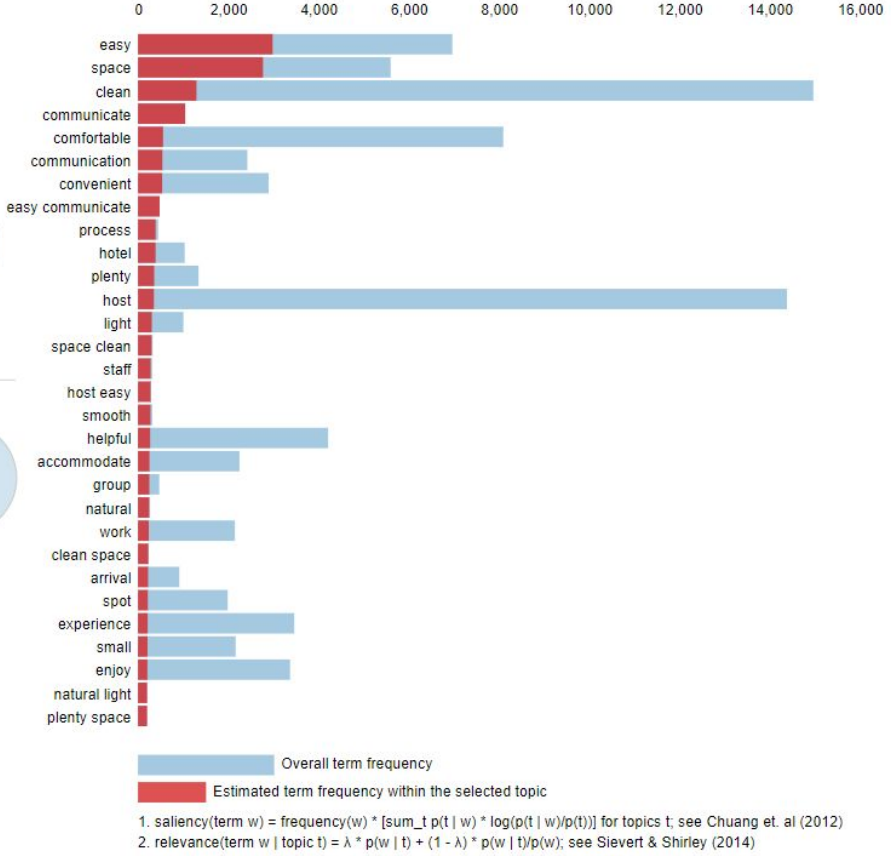
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 13 (5% of tokens)



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)

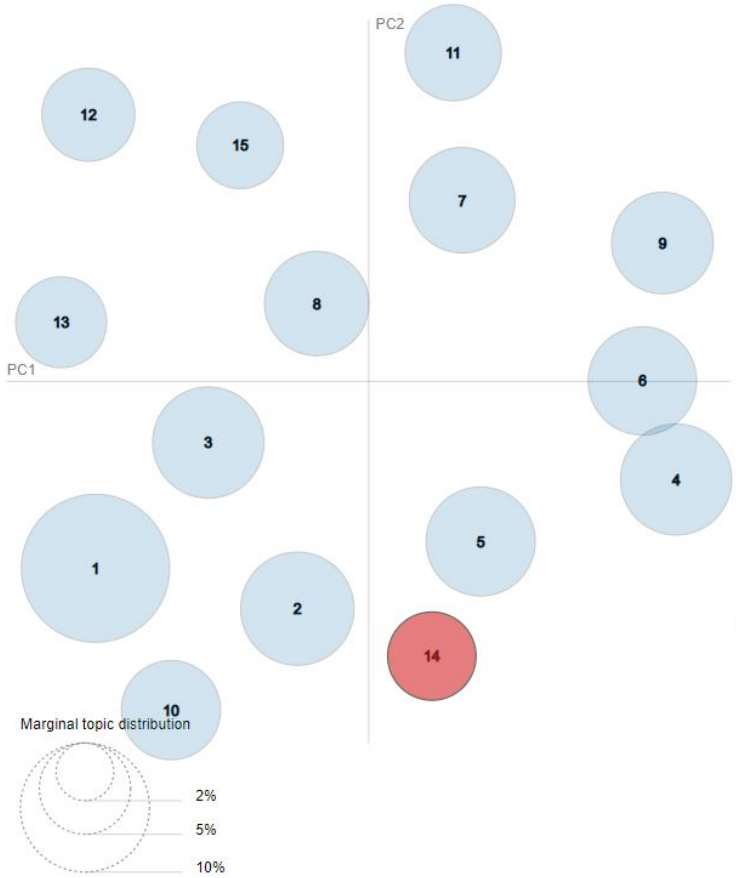
Selected Topic: 14 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

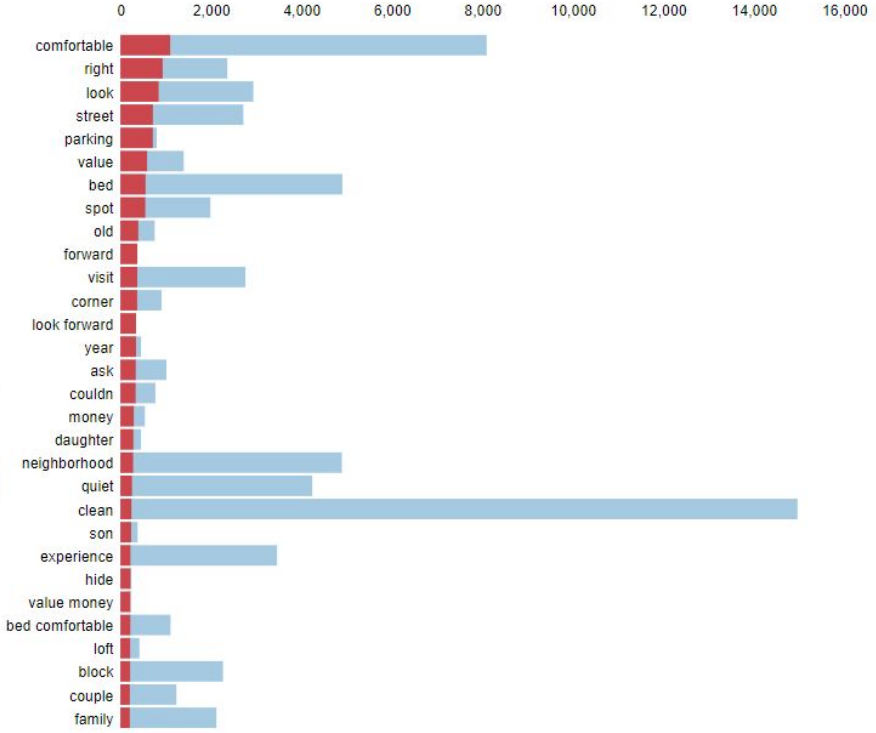
λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 14 (4.7% of tokens)



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)

Selected Topic: 15 Previous Topic Next Topic Clear Topic

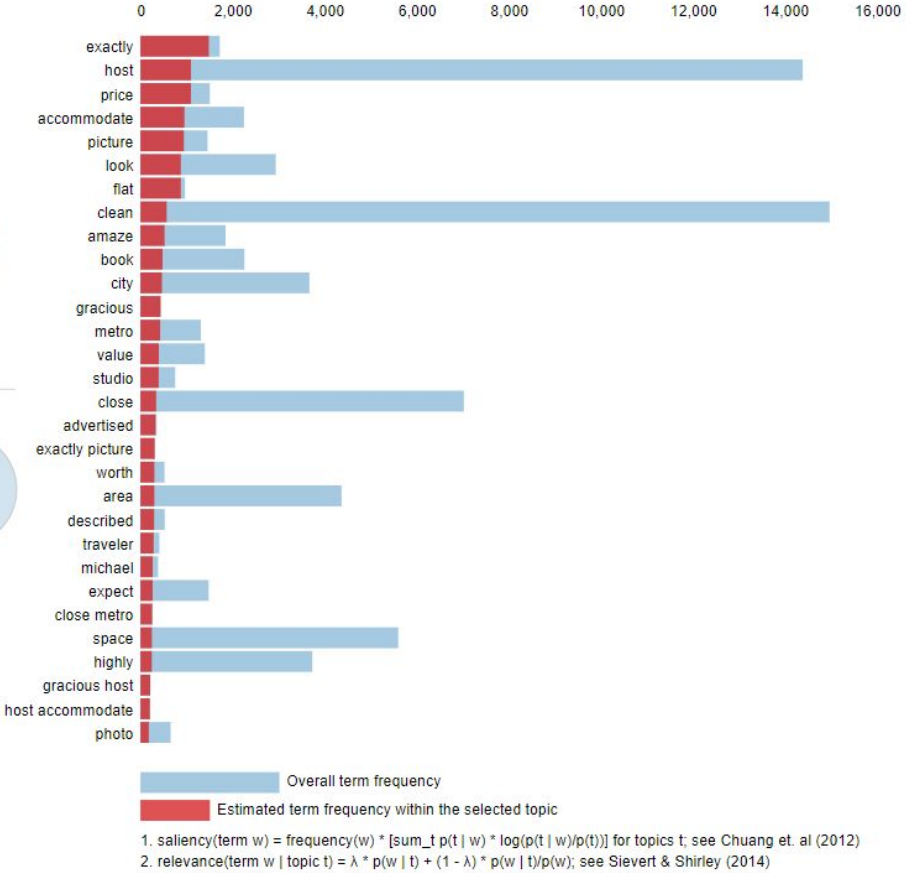
Slide to adjust relevance metric:(2)

λ = 1

0.0 0.2 0.4 0.6 0.8 1.0

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 15 (4.5% of tokens)



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)

Negative Topics:

Topic 1 - Inaccurate listings
Topic 2 - Check-in/out
Topic 3 - Bed/Bathroom
Topic 4 - Dirtiness and smell
Topic 5 - Uncomfortable sleep conditions
Topic 6 - Location
Topic 7 - Poor house maintenance
Topic 8 - Noise
Topic 9 - Hot Water/Heater
Topic 10 - Dirtiness
Topic 11 - Location
Topic 12 - Location general
Topic 13 - Value

Positive Topics Topics: