NLP Project
Hotel Review Analysis

METIS
Project V

What should hotels watch out for to improve their business?

Goal

Group relevant
keywords

Suggest key findings
for improvement

Workflow

Preprocess Data
NLTK Tokenizer

→

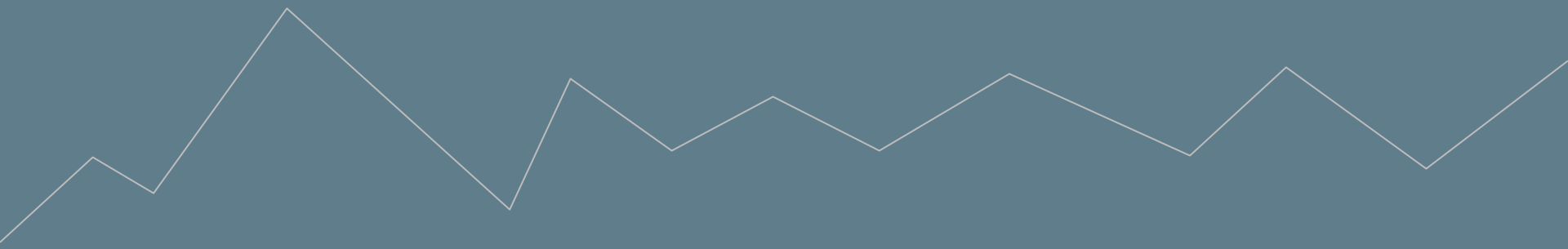
CountVectorize
ScikitLearn

→

LDA

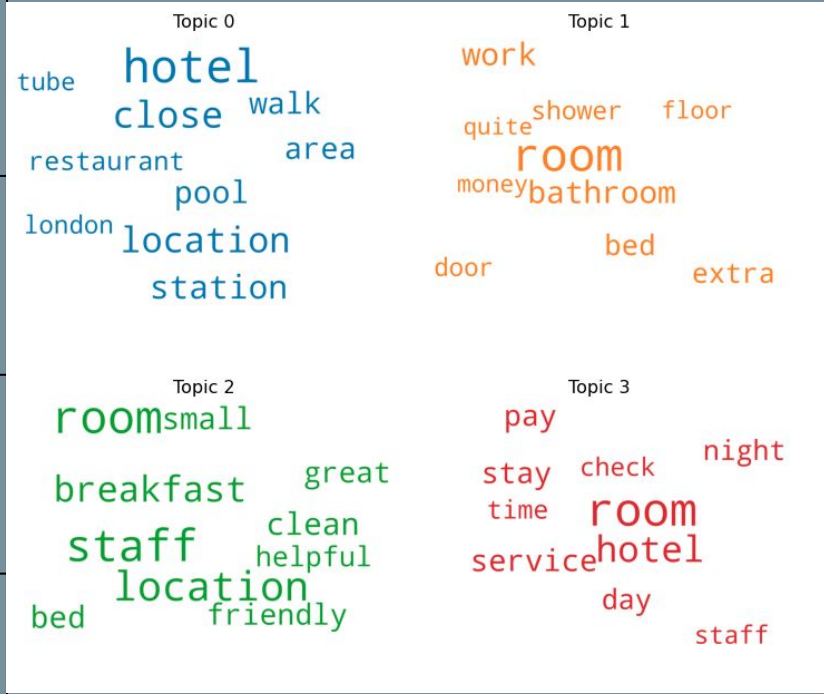
→

Interpret Findings
WordCloud
pyLDAvis

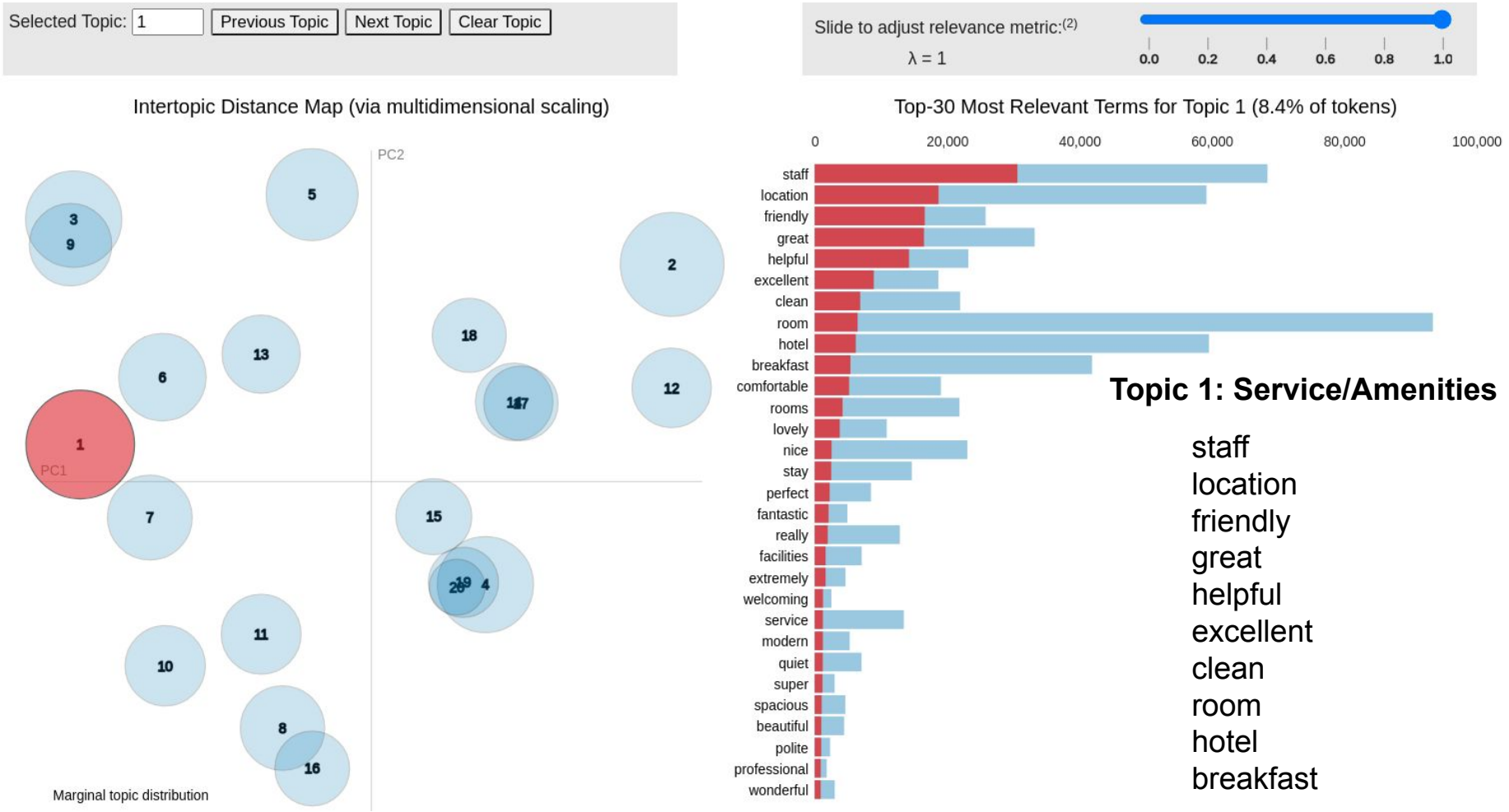


LDA Topic Modeling

	Topic 0: Hotel Surroundings	Topic 1: Hotel Amenities	Topic 2: Hotel Service	Topic 3: Hotel Check-In
1st	Hotel	Room	Room	Room
2nd	Close	Work	Staff	Hotel
3rd	Location	Bathroom	Location	Service

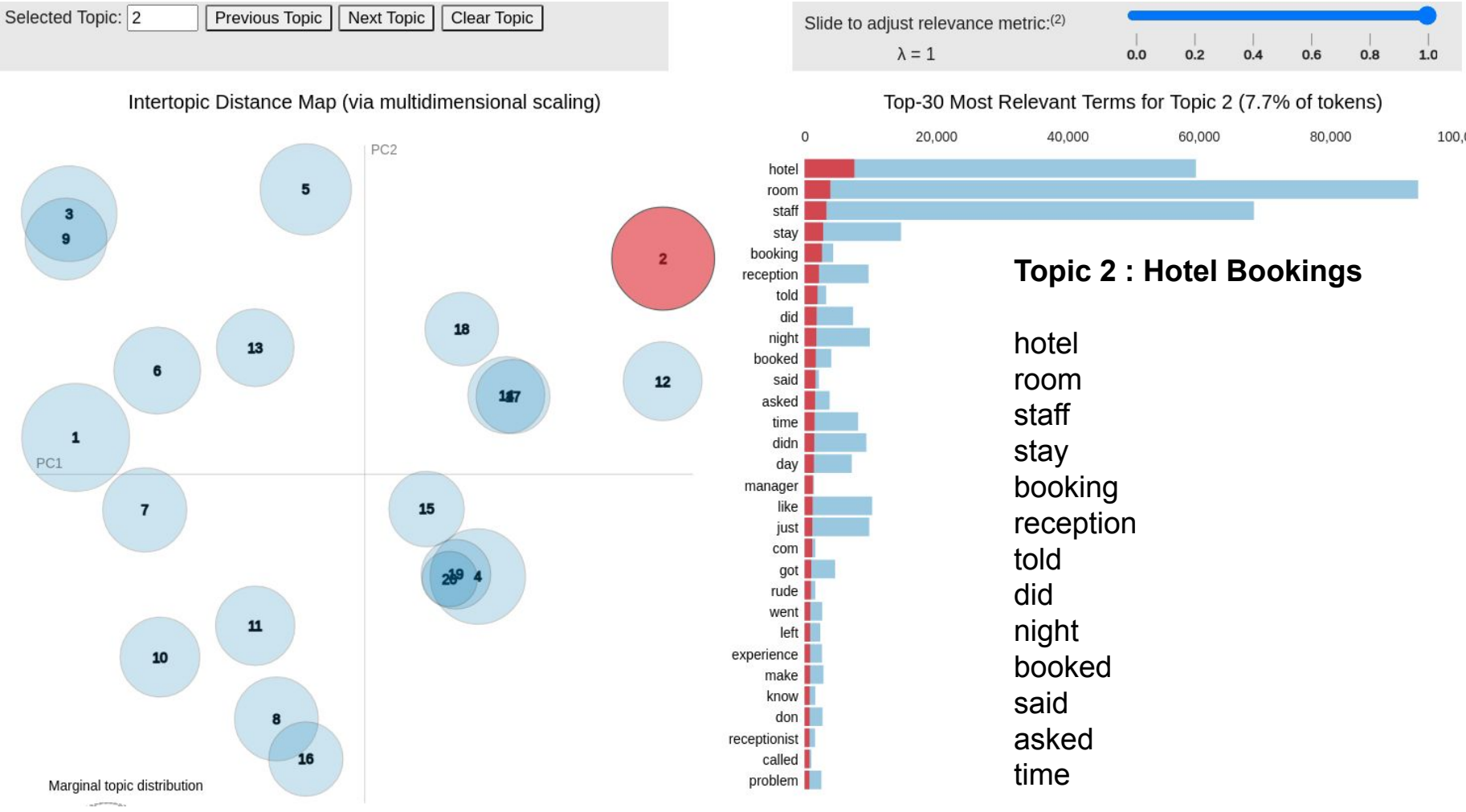


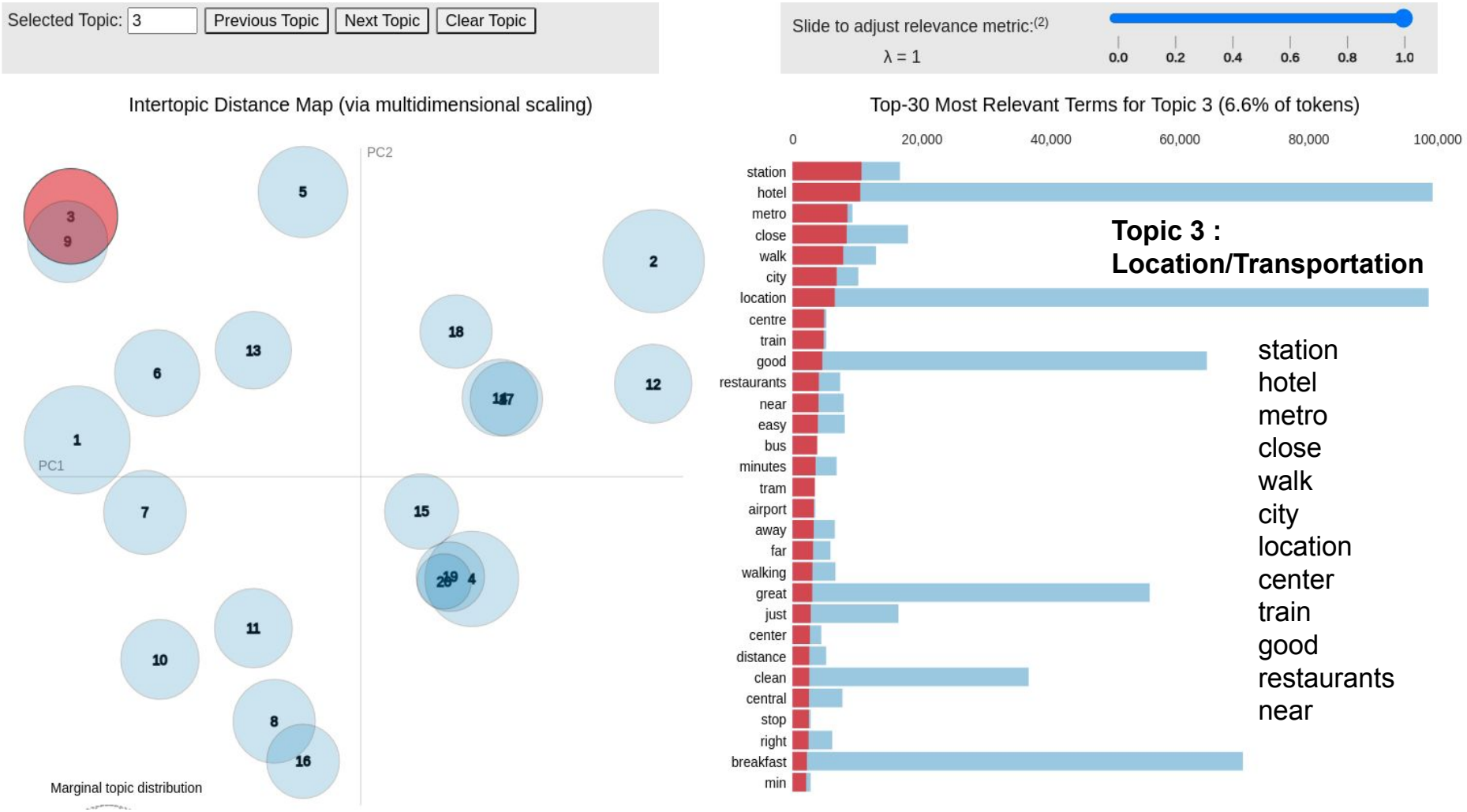
pyLDAvis Visualization

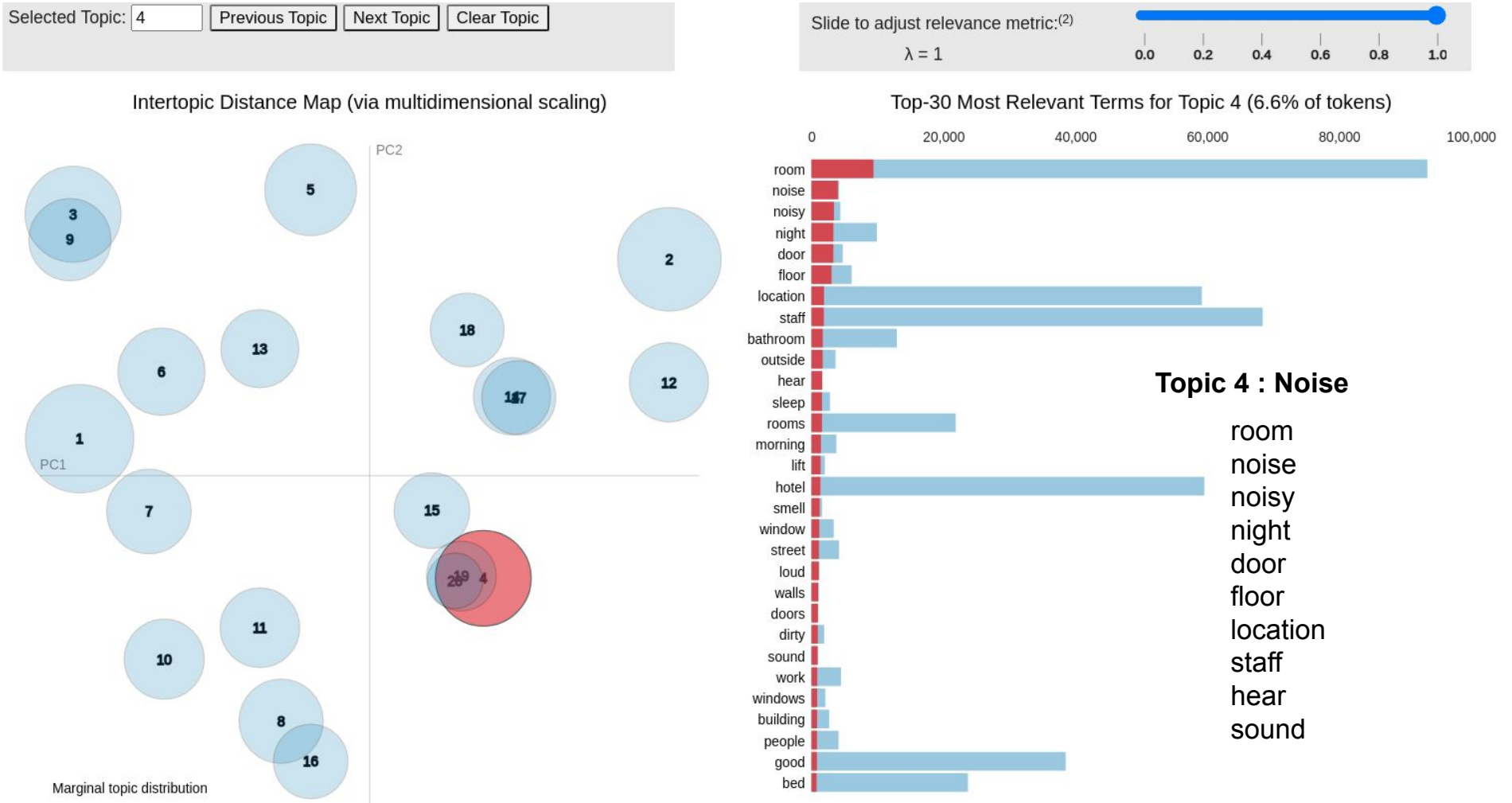


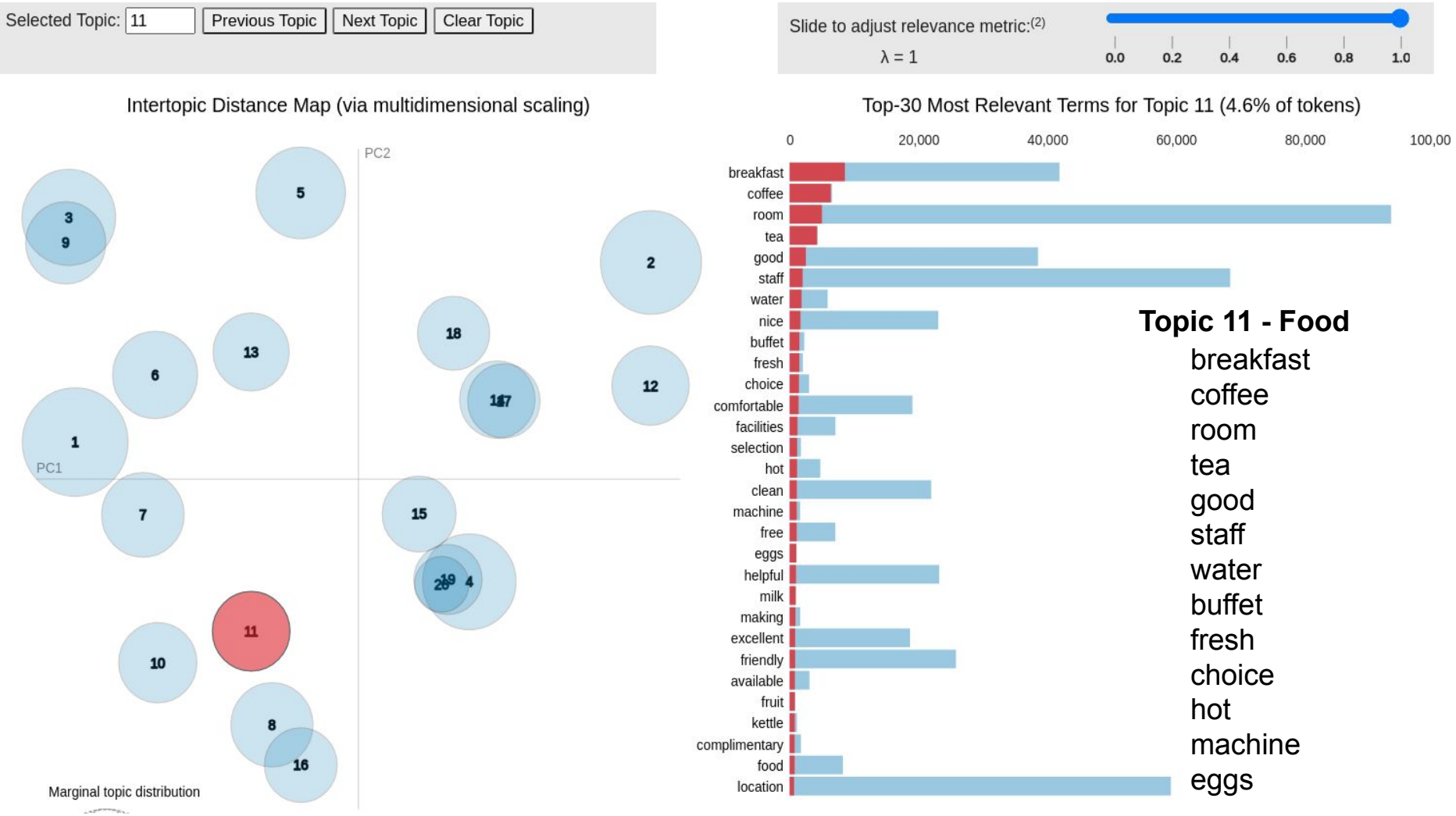
Topic 1: Service/Amenities

staff
location
friendly
great
helpful
excellent
clean
room
hotel
breakfast


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
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 2 (7.7% of tokens)
0
20,000
40,000
60,000
80,000
100,000
hotel
room
staff
stay
booking
reception
told
did
night
booked
said
asked
time
didn
day
manager
like
just
com
got
rude
went
left
experience
make
know
don
receptionist
called
problem
Topic 2 : Hotel Bookings
hotel
room
staff
stay
booking
reception
told
did
night
booked
said
asked
time


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)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 3 (6.6% of tokens)
0
20,000
40,000
60,000
80,000
100,000
station
hotel
metro
close
walk
city
location
centre
train
good
restaurants
near
easy
bus
minutes
tram
airport
away
far
walking
great
just
center
distance
clean
central
stop
right
breakfast
min
Topic 3 :
Location/Transportation
station
hotel
metro
close
walk
city
location
center
train
good
restaurants
near


Selected Topic: 4
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
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 4 (6.6% of tokens)
0
20,000
40,000
60,000
80,000
100,000
room
noise
noisy
night
door
floor
location
staff
bathroom
outside
hear
sleep
rooms
morning
lift
hotel
smell
window
street
loud
walls
doors
dirty
sound
work
windows
building
people
good
bed
Topic 4 : Noise
room
noise
noisy
night
door
floor
location
staff
hear
sound

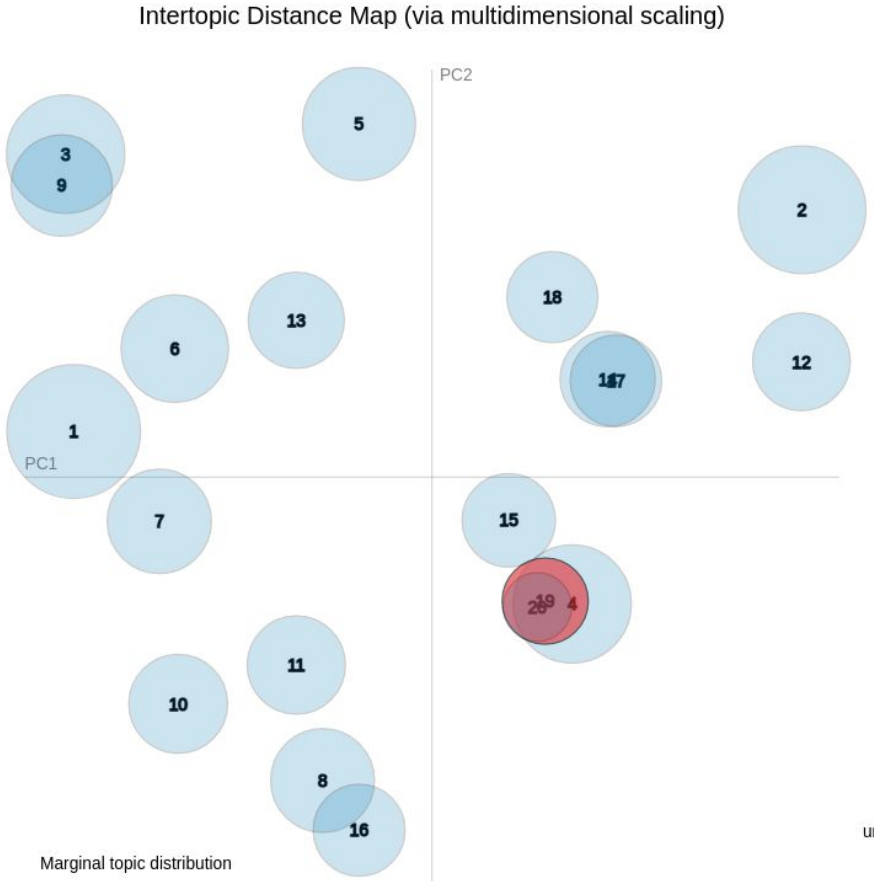

Selected Topic: 11
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Relevant Terms for Topic 11 (4.6% of tokens)
0
20,000
40,000
60,000
80,000
100,00
breakfast
coffee
room
tea
good
staff
water
nice
buffet
fresh
choice
comfortable
facilities
selection
hot
clean
machine
free
eggs
helpful
milk
making
excellent
friendly
available
fruit
kettle
complimentary
food
location
Topic 11 - Food
breakfast
coffee
room
tea
good
staff
water
buffet
fresh
choice
hot
machine
eggs

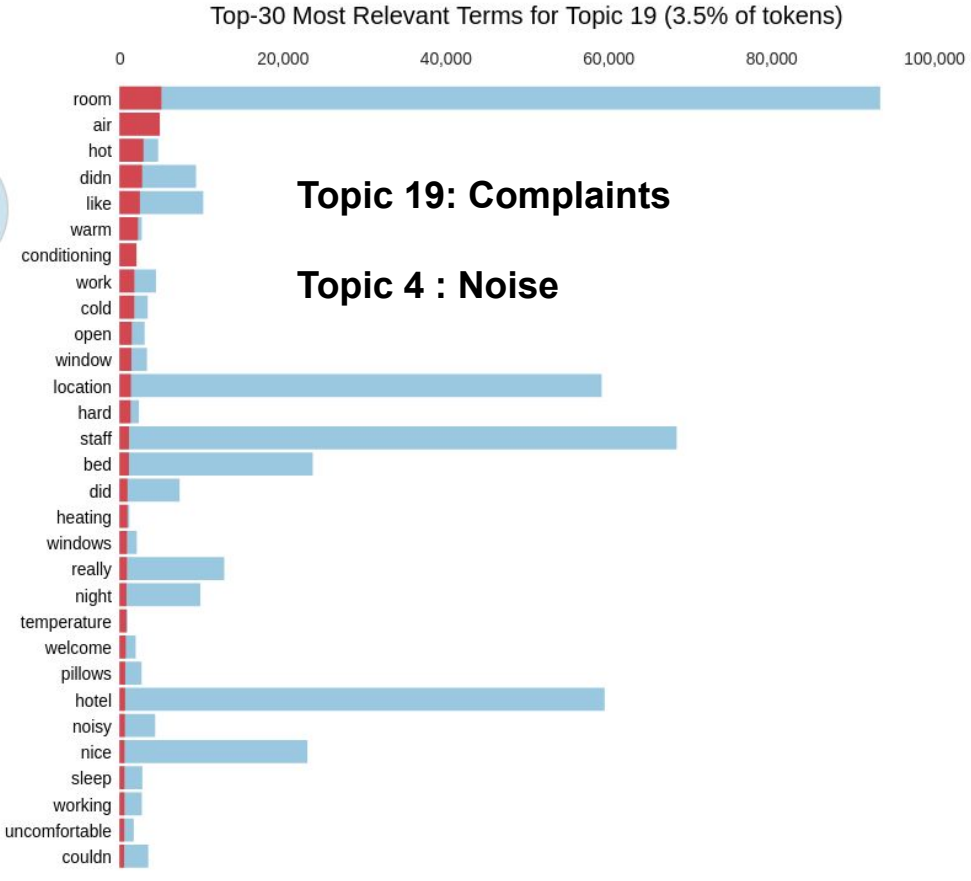
Overlapping Topics

Topic 9 - Proximity

Has high similarity with Topic 3 - (Location/Transportation)

Overlapping Topics





Topic 19: Complaints

Topic 4 : Noise

Insights/Improvements

- 5 main topics that hotel should focus on
- **Service, Bookings, Location, Noise, Food**
- Overall a good breakdown of topics
- Overlapping topics are somewhat distinguishable, but share common words

Improvements

- Coherence score metric to better recommend by topic
- Create Recommendation System for app users

Each bubble = a topic , larger the bubble higher percentage of the number of words in the corpus

Blue bar = overall frequency of each word in corpus

Red bars = estimated # of times a given term was generated by a given topic

Appendix

NMF model - Top 15 words

THE TOP 15 WORDS FOR TOPIC 0:

['quite', 'floor', 'big', 'air', 'work', 'window', 'shower', 'noisy', 'little', 'bit', 'view', 'bathroom', 'size', 'small', 'room']

THE TOP 15 WORDS FOR TOPIC 1:

['facility', 'superb', 'quiet', 'wifi', 'ideal', 'expensive', 'comfort', 'fantastic', 'convenient', 'cleanliness', 'price', 'central', 'staff', 'perfect', 'location']

THE TOP 15 WORDS FOR TOPIC 2:

['attentive', 'pleasant', 'super', 'efficient', 'fantastic', 'professional', 'polite', 'really', 'welcome', 'reception', 'lovely', 'extremely', 'helpful', 'friendly', 'staff']

THE TOP 15 WORDS FOR TOPIC 3:

['price', 'close', 'near', 'quality', 'station', 'shower', 'really', 'size', 'restaurant', 'wifi', 'food', 'facility', 'money', 'value', 'good']

LDA Topic Modeling

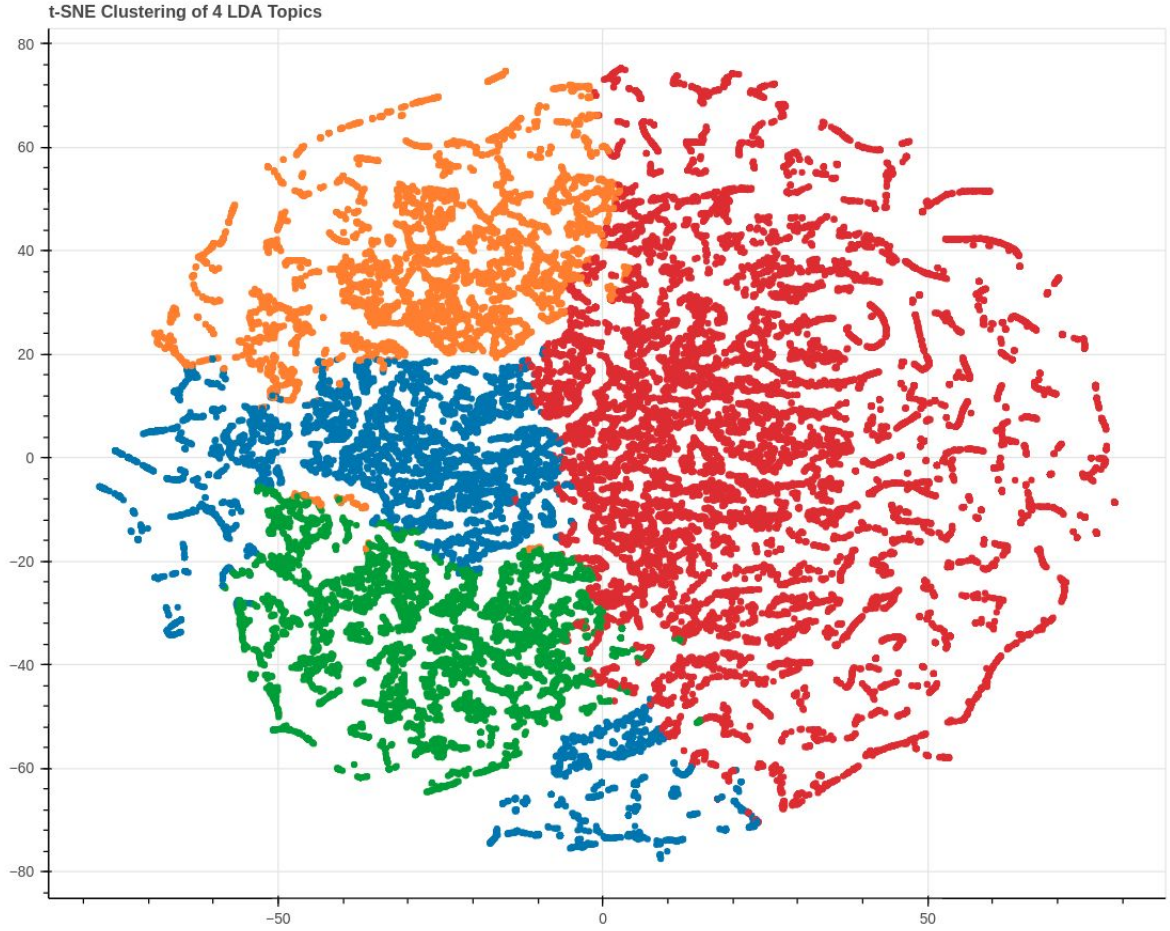
```
[(0,
  '0.067*"hotel" + 0.041*"close" + 0.038*"location" + 0.032*"station" + '
  '0.029*"pool" + 0.027*"area" + 0.027*"walk" + 0.019*"restaurant" + '
  '0.018*"london" + 0.017*"tube"'),
 (1,
  '0.080*"room" + 0.028*"work" + 0.028*"bathroom" + 0.023*"bed" + '
  '0.022*"extra" + 0.019*"shower" + 0.018*"door" + 0.016*"floor" + '
  '0.016*"money" + 0.015*"quite"'),
 (2,
  '0.088*"room" + 0.082*"staff" + 0.069*"location" + 0.052*"breakfast" + '
  '0.038*"clean" + 0.037*"bed" + 0.037*"small" + 0.034*"friendly" + '
  '0.034*"great" + 0.031*"helpful"'),
 (3,
  '0.045*"room" + 0.026*"hotel" + 0.018*"service" + 0.016*"stay" + 0.016*"pay" '
  '+ 0.015*"day" + 0.015*"night" + 0.013*"time" + 0.012*"check" + '
  '0.012*"staff"')]
```

Topic 0 - Hotel Surroundings

Topic 1 - Hotel Amenities

Topic 2 - Hotel Service

Topic 3 - Hotel Check-In



Red = **Topic 3**

Green = **Topic 2**

Orange = **Topic 1**

Blue = **Topic 0**