

Trip Advisor Hotel Reviews

NLP Project

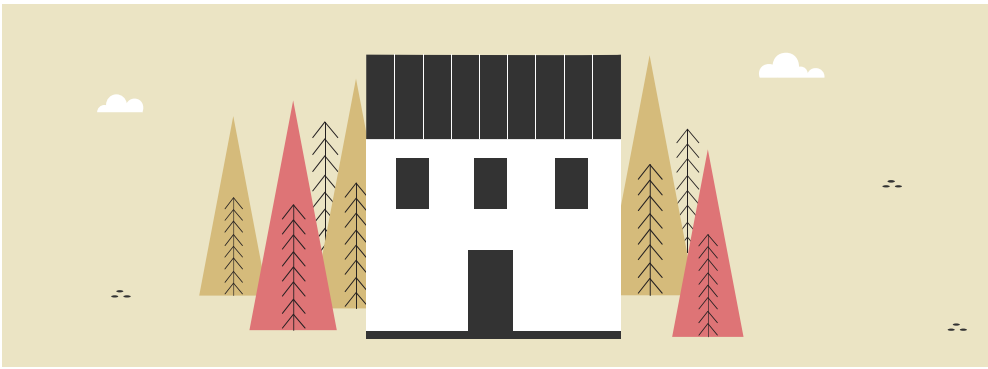
By: Amirah , Amal , Alanoud



What is Tripadvisor?

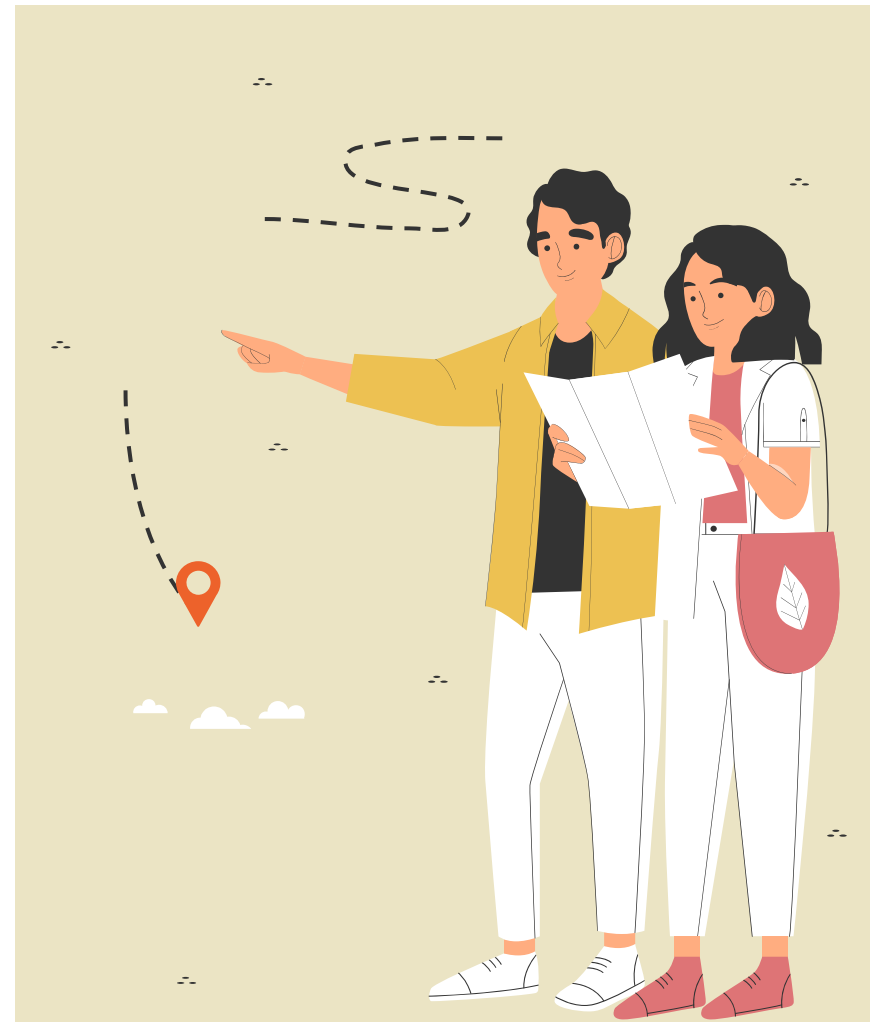
Tripadvisor is an American online travel company that operates a website with user-generated content and It also offers online hotel reservations, bookings for transportation and travel experiences.





Project Goals

selecting best model
performing data and
give us the best topic
modeling.

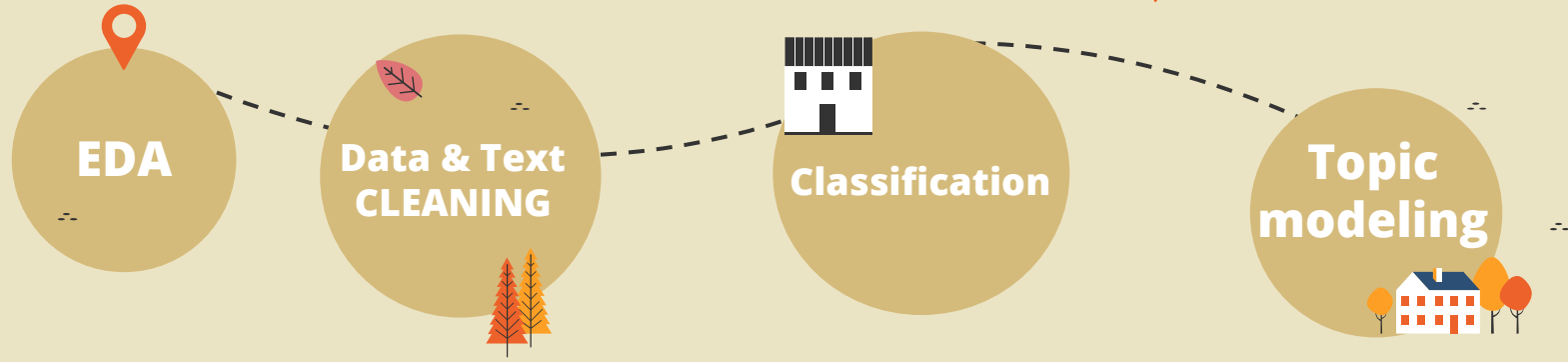




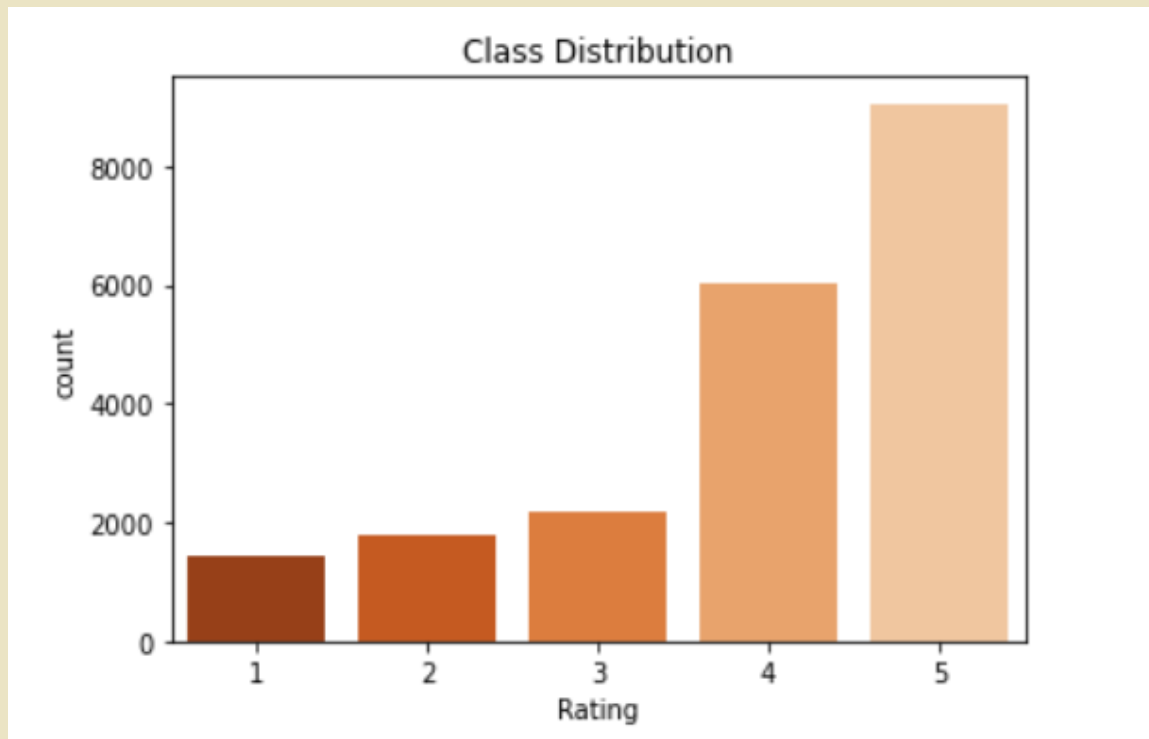
Datasets

- 20k rows of hotels reviews
- 2 columns: **review** and **Rating**

Workflow



EDA



Text Cleaning



Stop word



useless words

Like [also,tell,
meanwhile,however,arriv]



Symbols



Numbers

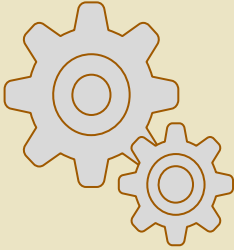


Extraspace

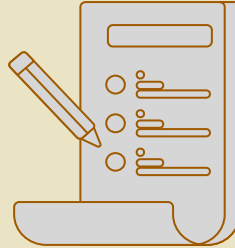


punctuations

Text Preprocessing



Contractions



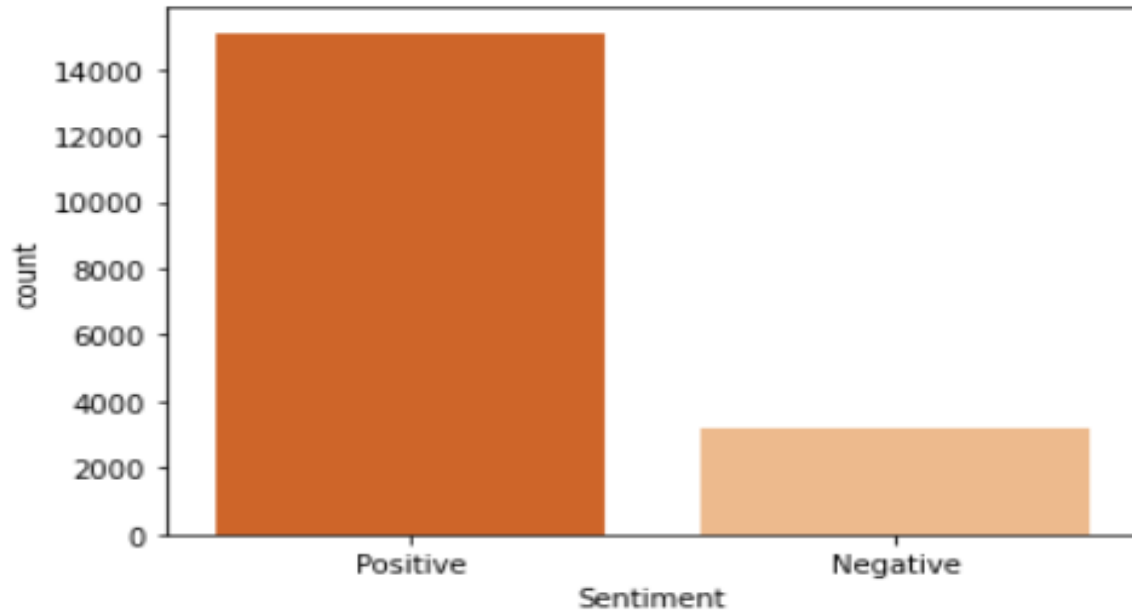
Lemmatization



Stemming

[illegible]

Sentiment Analysis





Classification




	LR1-TFIDF	LR2-TFIDF	Naive1-TFIDF	Naive2-TFIDF	AdaBoost1-TFIDF	AdaBoost2-TFIDF	SVM1-TFIDF	SVM2-TFIDF
Accuracy	0.944	0.905	0.832	0.829	0.914	0.910	0.949	0.922
Precision	0.943	0.897	0.831	0.830	0.930	0.925	0.952	0.916
Recall	0.991	0.999	1.000	0.997	0.970	0.970	0.988	0.997
F1 Score	0.966	0.945	0.908	0.906	0.950	0.947	0.970	0.955



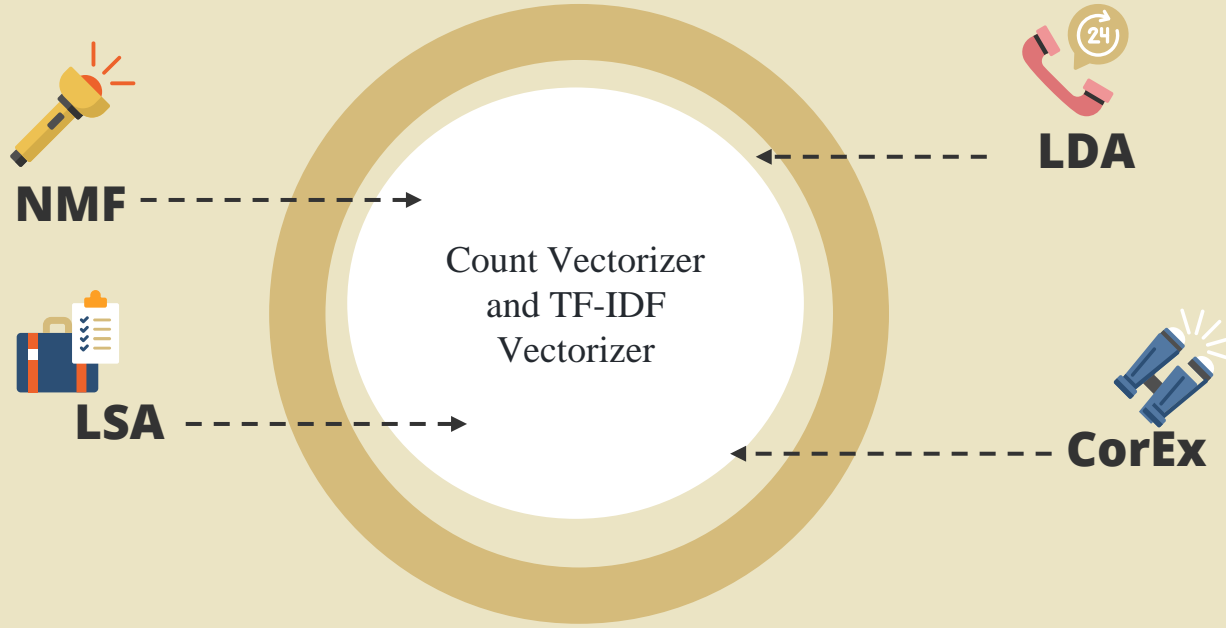


Classification

	LR1-TFIDF	LR2-TFIDF	Naive1-TFIDF	Naive2-TFIDF	AdaBoost1-TFIDF	AdaBoost2-TFIDF	 SVM1-TFIDF	SVM2-TFIDF
Accuracy	0.944	0.905	0.832	0.829	0.914	0.910	0.949	0.922
Precision	0.943	0.897	0.831	0.830	0.930	0.925	0.952	0.916
Recall	0.991	0.999	1.000	0.997	0.970	0.970	0.988	0.997
F1 Score	0.966	0.945	0.908	0.906	0.950	0.947	0.970	0.955



Topic Modeling



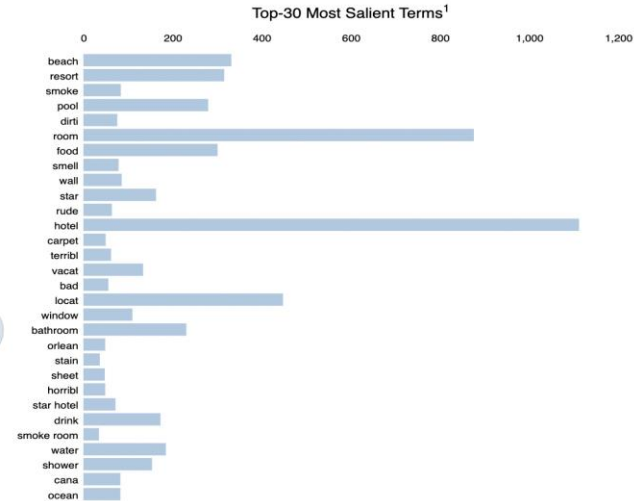
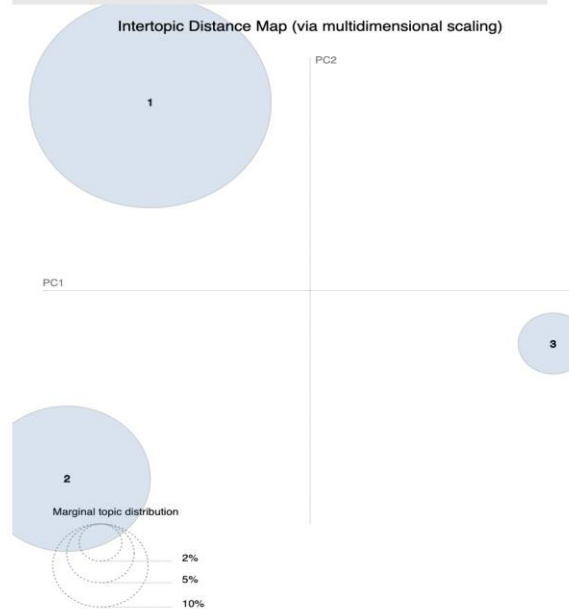
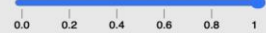
In [52]: `pyLDAvis.sklearn.prepare(lda_2, doc_word2, Tfidf_vectorizer)`

/Users/alanoudabdulaziz/opt/anaconda3/lib/python3.8/site-packages/ipykernel/ipkernel.py:287: DeprecationWarning: `should_run_async` will not call
ll' automatically in the future. Please pass the result to `transformed_cell` argument and any exception that happen during thetransform in `prepro
tuple` in IPython 7.17 and above.
and should_run_async(code)

Out[52]:

Selected Topic: Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)
 $\lambda = 1$



Overall term frequency
Estimated term frequency within the selected topic

1. $saliency(term, w) = frequency(w) \cdot [\sum_i p(t_i, w) \cdot \log(p(t_i, w)/p(t_i))]$ for topics t_i ; see Chuang et al (2012)
2. $relevance(term, w, t, topic, \lambda) = \lambda \cdot p(w, t, \lambda) + (1 - \lambda) \cdot p(w, t, \lambda)$; see Sievert & Shirley (2014)

And the model performing the best topic modeling is
LDA for TF-IDF Vectorizer with 3 topics.

Selected Topic: 0

Previous Topic

Next Topic

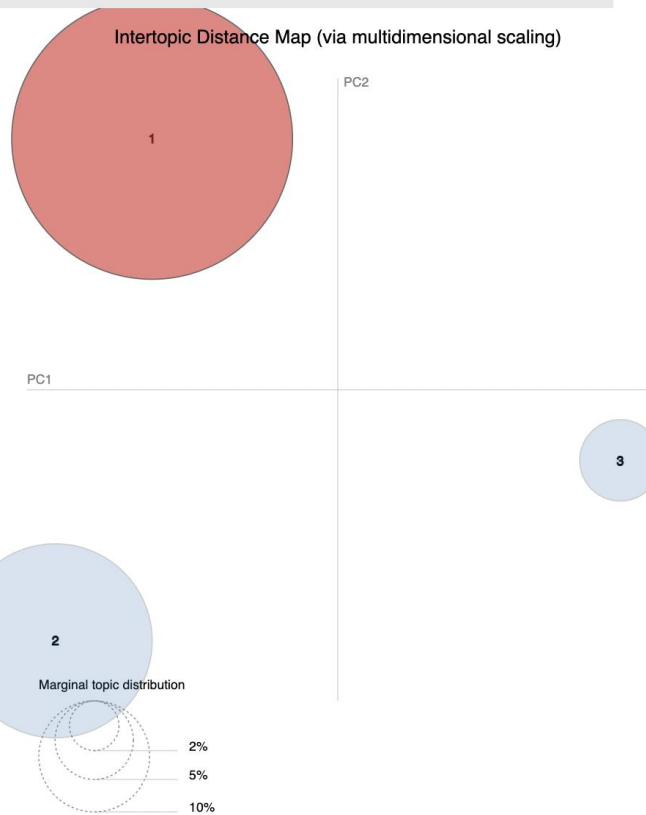
Clear Topic

Slide to adjust relevance metric:⁽²⁾

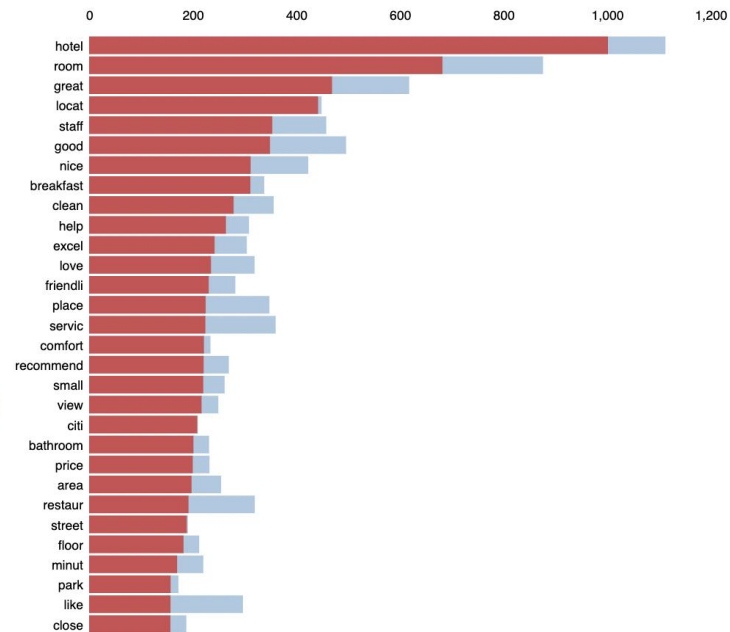
$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 1 (64.1% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t|w) * \log(p(t|w)/p(t))]$ for topics t ; see Chuang et. al. (2012)

2. $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley. (2014)

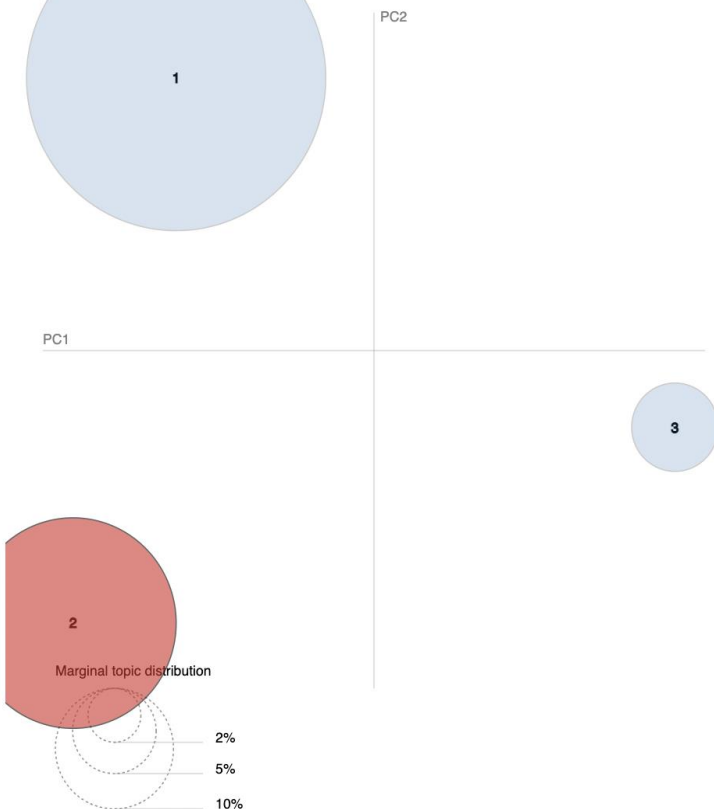
Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2)

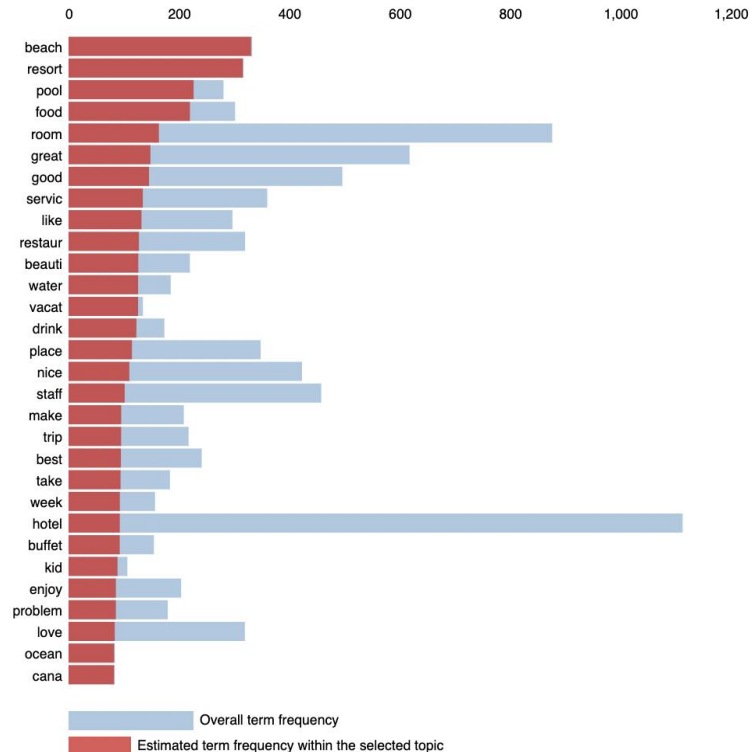
$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 2 (30.5% of tokens)



1. $\text{saliency}(\text{term } w) = \text{frequency}(w) \cdot \left[\sum_t p(t|w) \cdot \log\left(\frac{p(t|w)}{p(t)}\right) \right]$ for topics t ; see Chuang et. al (2012)

2. $\text{relevance}(\text{term } w | \text{topic } t) = \lambda \cdot p(w|t) + (1 - \lambda) \cdot p(w|t)/p(w)$; see Sievert & Shirley (2014)

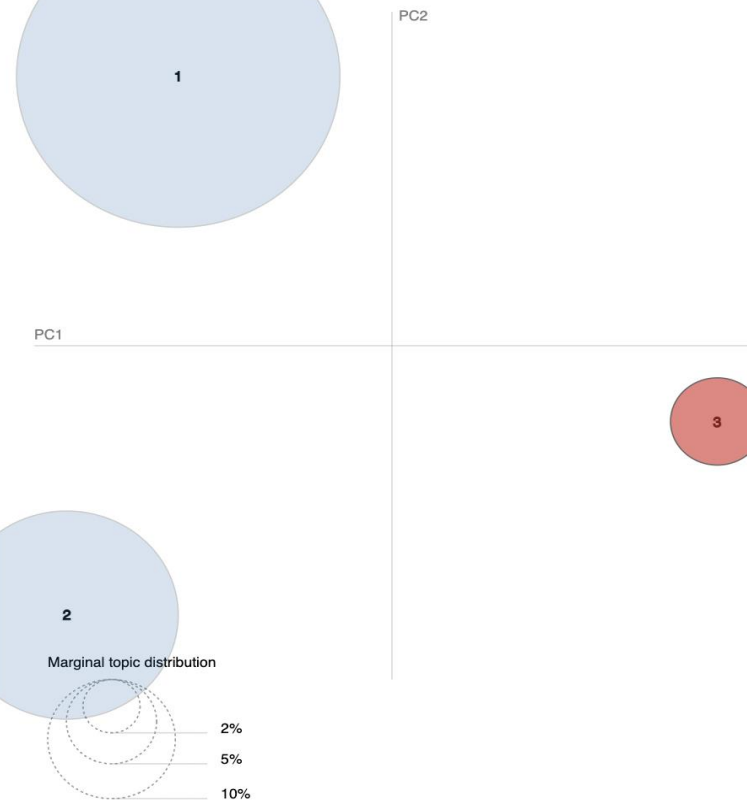
Selected Topic: 0

Slide to adjust relevance metric:(2)

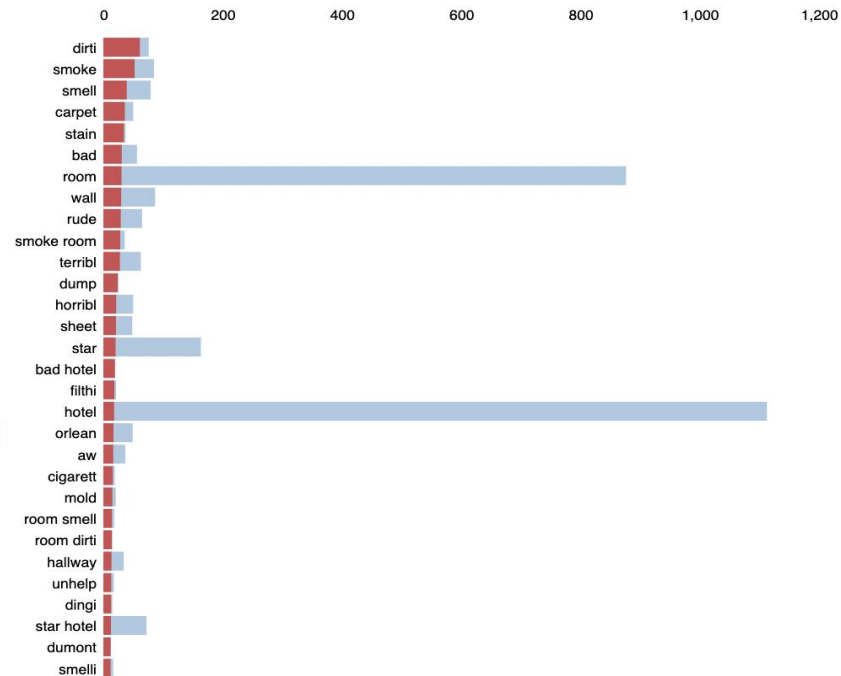
$\lambda = 1$

0.0 0.2 0.4 0.6 0.8 1

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 3 (5.4% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. $\text{saliency}(\text{term } w) = \text{frequency}(w) \cdot \left[\sum_t p(t|w) \cdot \log(p(t|w)/p(t)) \right]$ for topics t ; see Chuang et. al (2012)

2. $\text{relevance}(\text{term } w | \text{topic } t) = \lambda \cdot p(w|t) + (1 - \lambda) \cdot p(w|t)/p(w)$; see Sievert & Shirley (2014)



Future work

Add another dataset contain the hotels name to do recommendation system.

THANKS

