THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

TEXT MINING FOR HOTEL REVIEWS

Data Mining Academic

- Renzo Castagnino
- Cristina Giraldo
- Siqi Jiang

Topics to cover

1. Introduction
2. Smart Question
3. Summary of Data
4. Data Cleaning
5. Data Preprocessing
6. Text Mining
7. Text Mining Methods
    1. Naïve Bayes
    2. Support Vector Machine with RBF Kernel
    3. Linear Vector Machine with Lasso regression
8. Results
9. Conclusion

Introduction

1. The increasing of textual knowledge and unstructured data has caused a boost in text mining research.

1. In the case of the hotel industry, there is a vast amount of data that can be explore, specifically the reviews of the hotels.

1. Analyzing the comments will helps us understand the rating of the hotel

1. The insights could help customers in taking better decisions. For the hotels, it will help them understand the perception of their customers

Smart Question

Does the customers' reviews can be useful to predict the rating of the hotels?

Summary of Data

1. Dataset of US hotel reviews from 2016 to 2018.
2. 10,000 instances in a .csv file

id	dateadded	dateupdated	address	categories	primarycate	city	country	keys	latitude	longitude	name	postalcode	province	reviews_dat	reviews_dat	reviews_rat	reviews_sou	reviews_text	reviews_title	reviews_use	reviews_use	reviews_use	sourceurls	websites	location
AVwc252WI	2016-10-30T	2018-09-10T	5921 Valenc	Hotels,Hotel	Accommoda	Rancho Sant	US	us/ca/ranch	32.990959	-117.18614	Rancho Vale	92067	CA	2013-11-14T	2016-08-03T	5	https://www	Our experier	Best romantic vacation ever!!!!			Paula	http://www.	http://www.	POINT(-117.186136 32.990959)
AVwc252WI	2016-10-30T	2018-09-10T	5921 Valenc	Hotels,Hotel	Accommoda	Rancho Sant	US	us/ca/ranch	32.990959	-117.18614	Rancho Vale	92067	CA	2014-07-06T	2016-08-02T	5	https://www	Amazing plac	Sweet sweet serenity			D	http://www.	http://www.	POINT(-117.186136 32.990959)
AVwc252WI	2016-10-30T	2018-09-10T	5921 Valenc	Hotels,Hotel	Accommoda	Rancho Sant	US	us/ca/ranch	32.990959	-117.18614	Rancho Vale	92067	CA	2015-01-02T	2016-11-15T	5	https://www	We booked a	Amazing Property and Experience			Ron	http://www.	http://www.	POINT(-117.186136 32.990959)
AVwdOclqIN	2015-11-28T	2018-09-10T	7520 Teague	Hotels,Hotel	Accommoda	Hanover	US	us/md/hano	39.155929	-76.716341	Aloft Arunde	21076	MD	2016-05-15T	2016-05-21T	2	https://www	Currently in	Never again.	Richmond	VA	jaeem2016	http://www.	http://www.	POINT(-76.716341 39.155929)
AVwdOclqIN	2015-11-28T	2018-09-10T	7520 Teague	Hotels,Hotel	Accommoda	Hanover	US	us/md/hano	39.155929	-76.716341	Aloft Arunde	21076	MD	2016-07-09T	2016-07-31T	5	https://www	I live in Md a	ALWAYS GRE	Laurel	MD	MamaNiaOn	http://www.	http://www.	POINT(-76.716341 39.155929)
AVwdOclqIN	2015-11-28T	2018-09-10T	7520 Teague	Hotels,Hotel	Accommoda	Hanover	US	us/md/hano	39.155929	-76.716341	Aloft Arunde	21076	MD	2016-06-11T	2016-07-31T	5	https://www	I stayed here	Wonderful st	Laurel	MD	kevan777	http://www.	http://www.	POINT(-76.716341 39.155929)
AVwdOclqIN	2015-11-28T	2018-09-10T	7520 Teague	Hotels,Hotel	Accommoda	Hanover	US	us/md/hano	39.155929	-76.716341	Aloft Arunde	21076	MD	2016-04-30T	2016-05-05T	5	https://www	Beautiful roc	Worth the money			Princess F	http://www.	http://www.	POINT(-76.716341 39.155929)
AVwdOclqIN	2015-11-28T	2018-09-10T	7520 Teague	Hotels,Hotel	Accommoda	Hanover	US	us/md/hano	39.155929	-76.716341	Aloft Arunde	21076	MD	2016-06-24T	2016-07-17T	5	https://www	We stayed h	Great Hotel I	Clayton	NC	DebMurphy5	http://www.	http://www.	POINT(-76.716341 39.155929)
AVwdOclqIN	2015-11-28T	2018-09-10T	7520 Teague	Hotels,Hotel	Accommoda	Hanover	US	us/md/hano	39.155929	-76.716341	Aloft Arunde	21076	MD	2016-05-29T	2016-07-31T	5	https://www	I travel a lot	Short stay fo	Boston	MA	kayleighwills	http://www.	http://www.	POINT(-76.716341 39.155929)
AVwePiAX_7	2016-03-23T	2018-09-10T	315 SE Olym	Hotels,Hotel	Accommoda	Vancouver	US	us/wa/vancc	45.619212	-122.5252	Hampton Inr	98684	WA	2016-01-26T	2016-04-19T	5	https://www	In my line of	Amazing em	Portland		KristyWM	https://www	http://hamp	POINT(-122.525196 45.619212)
AVwePiAX_7	2016-03-23T	2018-09-10T	315 SE Olym	Hotels,Hotel	Accommoda	Vancouver	US	us/wa/vancc	45.619212	-122.5252	Hampton Inr	98684	WA	2016-05-03T	2016-05-13T	5	https://www	The staff is v	I loved our stay here			B M	https://www	http://hamp	POINT(-122.525196 45.619212)
AVwePiAX_7	2016-03-23T	2018-09-10T	315 SE Olym	Hotels,Hotel	Accommoda	Vancouver	US	us/wa/vancc	45.619212	-122.5252	Hampton Inr	98684	WA	2016-01-30T	2016-04-19T	5	https://www	Very friendly	Hampton Inr	Antioch		Cathleen S	https://www	http://hamp	POINT(-122.525196 45.619212)
AVwePiAX_7	2016-03-23T	2018-09-10T	315 SE Olym	Hotels,Hotel	Accommoda	Vancouver	US	us/wa/vancc	45.619212	-122.5252	Hampton Inr	98684	WA	2016-03-11T	2016-03-20T	5	http://www.	Upon arrivin	Perfection	Portland		1fiesty	https://www	http://hamp	POINT(-122.525196 45.619212)
AVwePiAX_7	2016-03-23T	2018-09-10T	315 SE Olym	Hotels,Hotel	Accommoda	Vancouver	US	us/wa/vancc	45.619212	-122.5252	Hampton Inr	98684	WA	2016-06-21T	2016-07-28T	5	https://www	This is a nice	Good hotel.	Port Orchard	WA	810michelle	https://www	http://hamp	POINT(-122.525196 45.619212)
AVwePiAX_7	2016-03-23T	2018-09-10T	315 SE Olym	Hotels,Hotel	Accommoda	Vancouver	US	us/wa/vancc	45.619212	-122.5252	Hampton Inr	98684	WA	2016-06-21T	2016-07-28T	5	https://www	Beautiful pro	Excellent!	San Diego	CA	travelchick1	https://www	http://hamp	POINT(-122.525196 45.619212)

Data Cleaning

1. Reduce features
2. Missing values
3. Convert to string
4. Noise Removal

```
def drop_columns(self):
    dropcols = ['id', 'dateadded', 'dateupdated', 'address', 'categories', 'primarycategories', 'keys', 'latitude',
                'longitude', 'postalcode', 'reviews_date', 'reviews_dateseen', 'reviews_sourceurls',
                'reviews_usercity', 'reviews_userprovince', 'reviews_username', 'sourceurls', 'websites', 'location'

    self.p_df = self.p_df.drop(dropcols, axis=1)
    return self.p_df
```

```
def missing_val(self):
    self.p_df.isnull().values.any()
    self.p_df["reviews_text"].isna().sum()
    self.p_df["reviews_title"].notnull().isna().sum()
    self.p_df['reviews_text'] = self.p_df['reviews_text'].dropna().reset_index(
        drop=True)
    self.p_df['reviews_text'] = self.p_df['reviews_text'].astype(str)
    return self.p_df
```

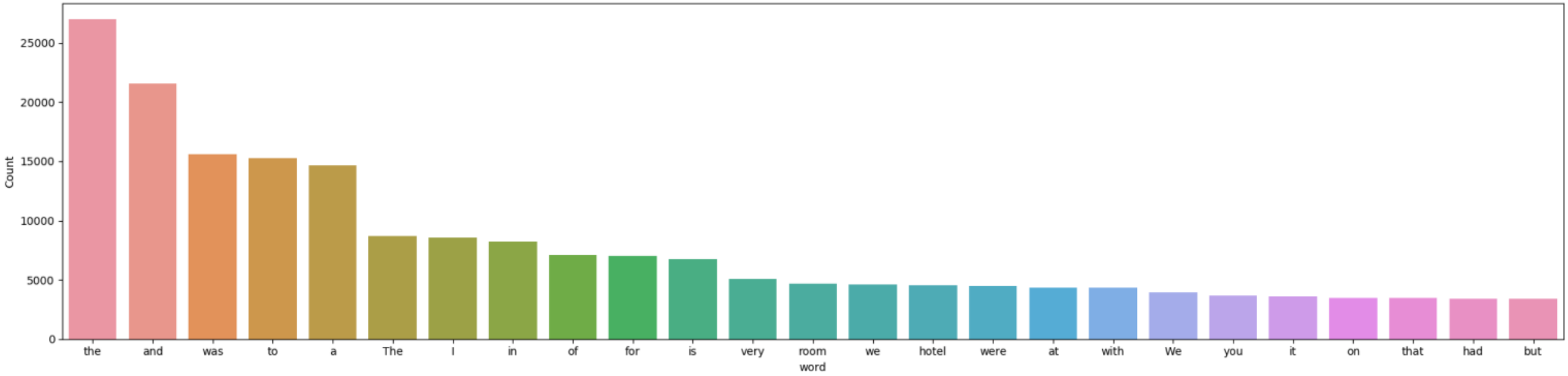
Data Preprocessing

1. Lowercase the reviews
2. Eliminate special characters
3. Delete common words
4. Remove Stop Words
5. Remove Rare Words

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

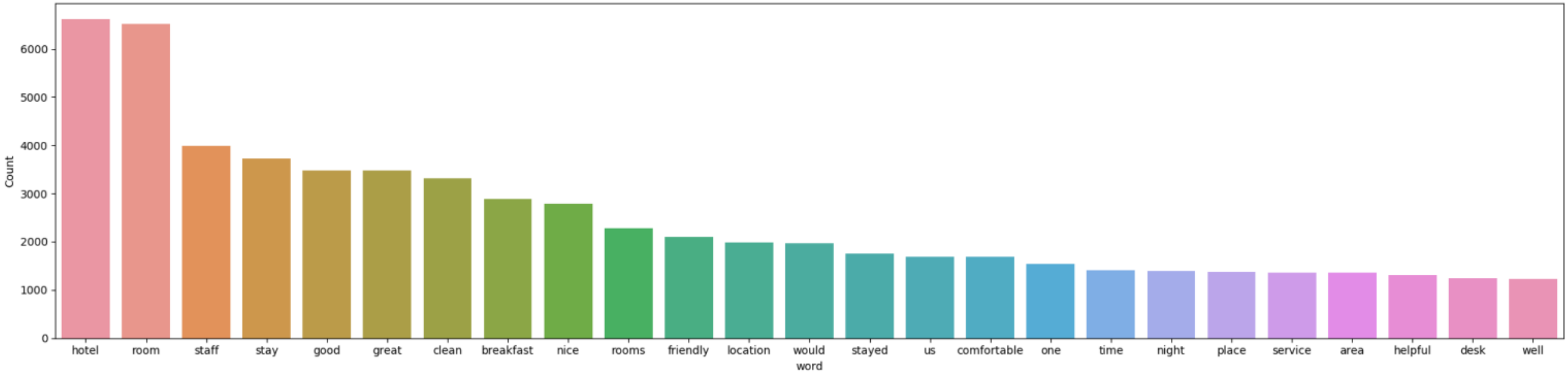
Data Preprocessing

- Frequency of words BEFORE the most frequent words removal



Data Preprocessing

- Frequency of words AFTER the most frequent words removal



Data Preprocessing

OTHER TEXT MINING TECHNIQUES

- Tokenization
- Lemmatization
- Stemming
- Spelling correction

Before the lemmatization and tokenization:

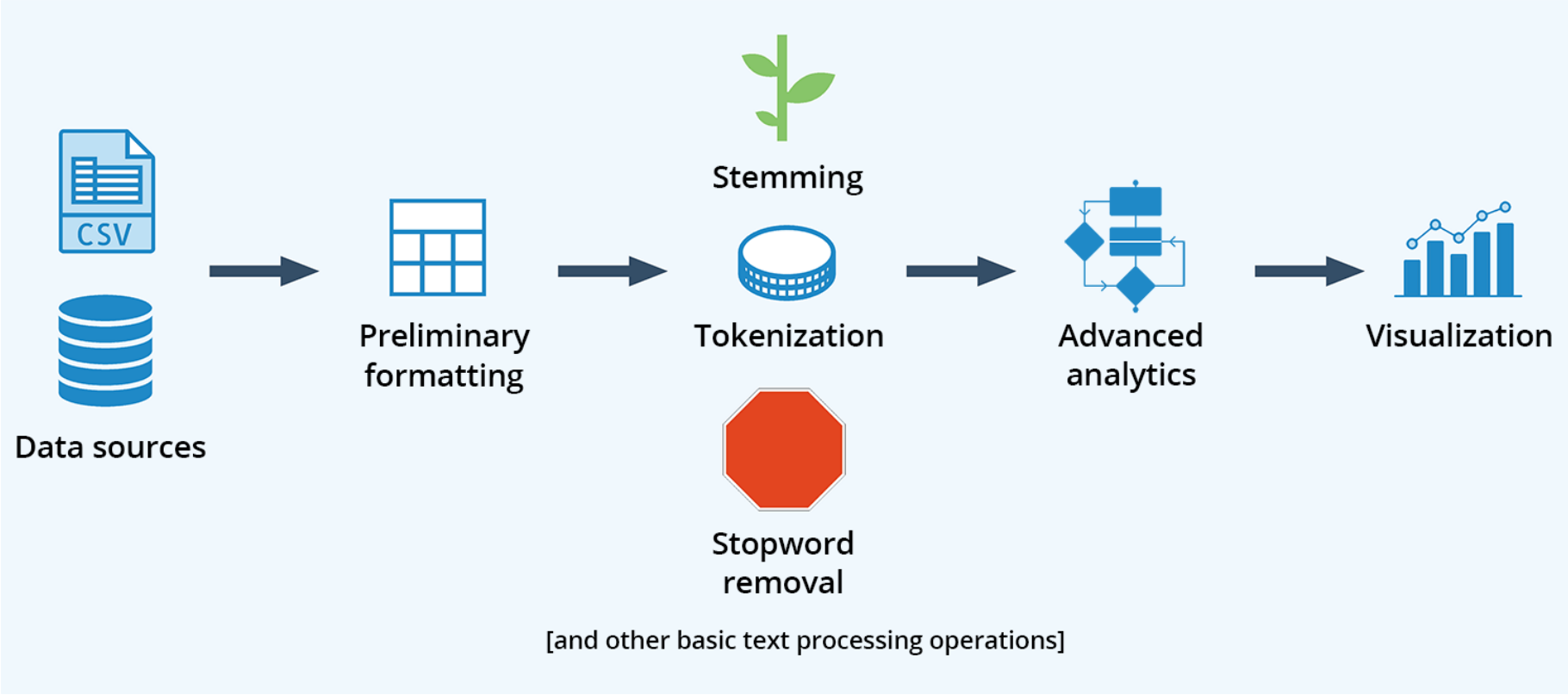
- experience rancho valencia absolutely perfect beginning end felt special happy stayed would come back heart beat

Tokenization:

- ['experience', 'rancho', 'valencia', 'absolutely', 'perfect', 'beginning', 'end', 'felt', 'special', 'happy', 'stayed', 'would', 'come', 'back', 'heart', 'beat']

Text Mining

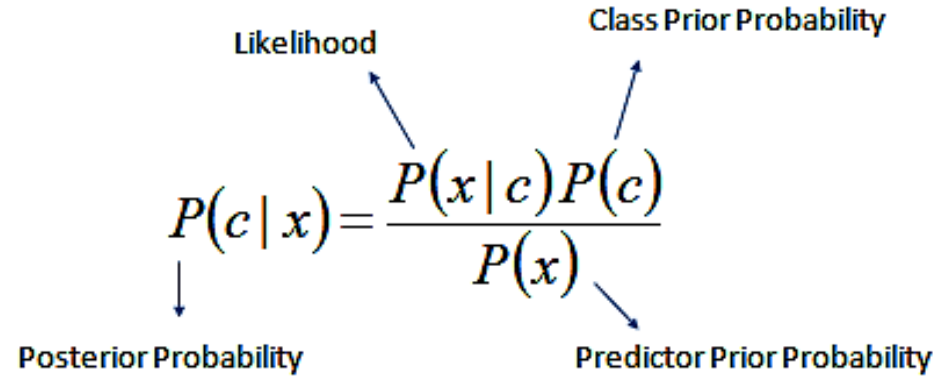
Also know as Text Analytics is the process of examining large collections of written resources to generate new information, and to transform the unstructured text into structured data for use in further analysis



Text Mining Methods - Naïve Bayes

1. This algorithm is based in probability. It is considered reliable and useful to work with big datasets

1. It has been used for a long time in classification problems such as text classification, spam filtering and sentiment analysis



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Results – Naïve Bayes

1. Sample of the results

4 => returned week long stay Americana want fancy don't stay complaints breakfast morning fine us bagel juice roll muffin coffee piece fruit good rooms clean yes older motel

4 => bad beds creaky thin walls hear everything room next door hallway housekeeping isn't consistent laundry delivery wrong room good location

4 => really nice smallish resort right lake geared towards families smaller children allow pets quite dogs upon checking gave us treats dog nice room clean resort nice little beach lots things kids

5 => bedroom suite right husband daughter enough space spread stay candlewood whenever visiting family nj like kitchen full size refrigerator microwave dishwasher two burner stove like free laundry place go

4 => bad bed small two people blanket thin staff ok friendly though good location new England style building room cute

Text Mining – Support Vector Machine (RBF)

1. Support vector machines is an algorithm that determines the best decision boundary between vectors that belong to the given group and vectors that do not belong to it.

1. The RBF kernel is defined as below. Th$||\mathbf{x} - \mathbf{x}'||^2$ as the squared euclidean distance between the two feature vectors.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right)$$

Results – Support Vector Machine (RBF)

1. The model has an accuracy of 46.20%.

1. It doesn't help with the performance by increasing the dimensional space using the RBF Kernel.

Text Mining – Linear Vector Machine With Lasso Regression

1. With this model, we are trying to improve the results obtain by the support vector machine, and have a better accuracy in the predictive model

1. We are also using LASSO regression to solve the optimization problem where t is a tuning parameter:

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$

$$s.t.\ y_i(\mathbf{w}^\intercal \mathbf{x}_i - b) \geq 1 - \xi_i \qquad i = 1, \ldots, m$$

$$\xi_i \geq 0$$

$$\min \|X\beta - b\|_2^2 \qquad (1)$$

$$s.t.\ \sum_{j=1}^{d}|\beta_j| \leq t \qquad (2)$$

Text Mining – Linear Vector Machine With Lasso Regression

1. Example of result of the algorithm:

1: dump rude staff disgusting room people refused need repair breakfast front room immediately worst pest

2: service staffs micro reset bad cleanliness everything old needs major made difficult go dinner falling apart complained

3: sparse unacceptable disorganized incompetent pillow case end night show age lot desired today better places

4: ac okay door bad room great hotel bad pool however good hotel bad valet nice price

5: great experience lovely thank cant wait love hotel amazing excellent exceptional loved hotel one best

The accuracy score of this model is 50%.

Conclusion

1. It is possible to analysis and predict the ratings of a hotel based on the reviews.

1. Comparing the 3 results, we can say the Linear vector machine performs better.

1. In the 3 models, the rating was between 45% to 50%. Which we can conclude that the data was not big enough to train the models.

4. Model can be improved by rescaling the rating

QUESTIONS