A Project Report

On

“Hotel Cluster Prediction using Machine Learning”

Submitted By:

Pulak Sinha (1606014)

Rohit Kumar (1606039)

Kumar Nishant Raj (1606110)

Aryan Roy (1506053)

Under the Supervision of:

Prof. A.K.Dudyala

Department of Computer Science and Engineering

National Institute of Technology Patna

January-April, 2019

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

(An Institute under Ministry Of HRD, Govt. of India)

ASHOK RAJPATH, PATNA-800005 (BIHAR)

CERTIFICATE

This is to certify that Pulak Sinha (1606014), Rohit Kumar (1606039), Kumar Nishant Raj (1606110), Aryan Roy (1506053) have carried out the project entitled “Hotel Cluster Prediction using Machine Learning” as their 6th semester Minor Project-I (6CS191) under my supervision.

(Signature)

Prof. A.K.Dudyala

Project Supervisor

CSE Department

(Signature)

Dr. Prabhat Kumar

Head of Department

CSE Department

राष्ट्रीय प्रौद्योगिकी संस्थान पटना
NATIONAL INSTITUTE OF TECHNOLOGY PATNA

DECLARATION

We hereby declare that this project work for Minor Project-I (6CS191) entitled **“Hotel Cluster Prediction using Machine Learning”** has been carried out by us under the supervision of **Prof. A.K.Dudyala , Department of Computer Science and Engineering, NIT Patna** . No part of this project has been submitted for the award degree or diploma to any other Institute.

S.No.	Name	Roll No.	Signature
1.	Pulak Sinha	1606014	
2.	Rohit Kumar	1606039	
3.	Kumar Nishant Raj	1606110	
4.	Aryan Roy	1506053	

Place: Patna

Date: 2nd May 2019

ACKNOWLEDGEMENT

We hereby take the privilege to express our gratitude to all the people who were directly or indirectly involved in the execution of this work, without whom this project would not have been a success.

We extend our deep gratitude, respect and obligation to our project supervisor, **Prof. A.K.Dudyala**, for his timely suggestions and encouragement.

We further express our gratitude to the Head of the Computer Science and Engineering Department, **Dr. Prabhat Kumar** for being a constant source of inspiration.

Our heartiest thank to our classmates who have supported us in all possible ways. Words are inadequate to express our gratitude to our parents and friends who have been supportive all the time. We would also like to thank our institution and the faculty members without whom this project would have been a distant reality.

Pulak Sinha (1606014)

Rohit Kumar (1606039)

Kumar Nishant Raj (1606110)

Aryan Roy (1506053)

Table of Content

SYNOPSIS

The objective of a Recommender System is to recommend relevant items for users, based on their preference. Preference and relevance are subjective and they are generally inferred by items users have consumed previously. The cold start problem is a well known and well researched problem for recommender systems, where system is not able to recommend items to users. due to three different situation i.e. for new users, for new products and for new websites. Content-based filtering is the method that solve this problem. Our system first uses the metadata of new products when creating recommendations, while visitor action is secondary for a certain period of time. And our systems recommend a product to a user based upon the category and description of the product.

CHAPTER 1

INTRODUCTION

1.1 Recommendation Engine

A **recommender engine** or a **recommendation engine** is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in commercial applications. Recommender systems are utilised in a variety of areas, and are most commonly recognised as playlist generators for video and music services like Netflix, YouTube and Spotify, product recommenders for services such as Amazon, or content recommenders for social media platforms such as Facebook and Twitter. These systems can operate using a single input, like music, or multiple inputs within and across platforms like news, books, and search queries. Recommender systems have been developed to explore research articles and experts, collaborators, financial services, and life insurance.

1.2 Machine Learning

Machine learning (ML) is the study of algorithms and mathematical models that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in the applications of email filtering, detection of network intruders, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers.

1.3 Cluster Prediction

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

CHAPTER 2

LITERATURE REVIEW

2.1 PROBLEM STATEMENT

To build a Recommendation System to predict most relevant booking outcome (hotel cluster) for an user event based on their search and other attributes associated with that user event.

We are interested in predicting which hotel group a user is most likely to view/ book next. The different parameters based on which similar hotels for a search are grouped together includes historical price, customer star ratings, geographical locations relative to city centre Customer feedback Extra amenities (like Wi-Fi connection, dining, etc.). Preferences and relevances are subjective and they are generally inferred by items users have consumed previously.

2.2 Collaborative Filtering

A system with collaborative filtering recommends items for a user which is liked by other users whose preference is alike. The collaborative filtering uses not the content of items but the similarity of users or items. In the newer, narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

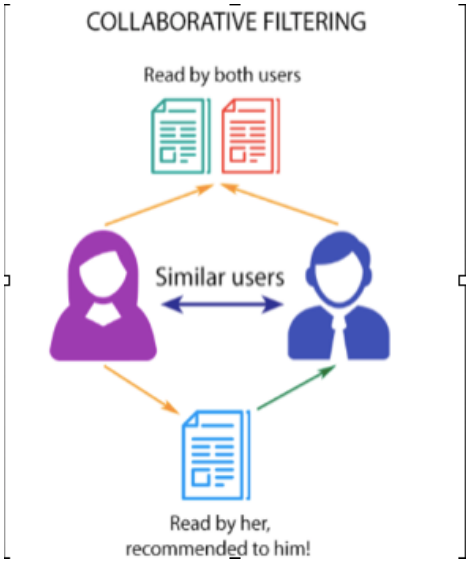


Fig 2.2(a) Collaborative Filtering

2.3 Content Based Filtering

A system with content-based filtering makes recommendation based on comparing user profile with the characteristic of content model. Content-based filtering uses the only characteristics of items in recommendation. Content-based recommendation systems may be used in a variety of domains ranging from recommending web pages, news articles, restaurants, television programs, and hotels. The advantage of content-based filtering is that it doesn't have a cold-start problem.

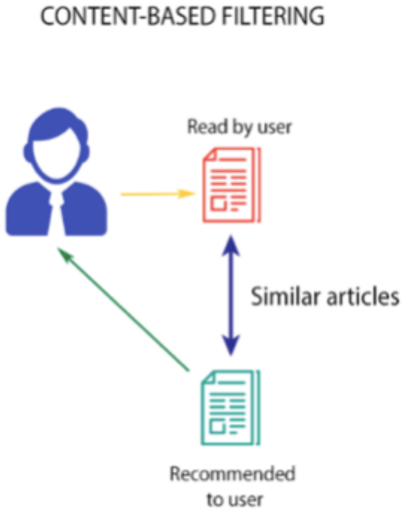


Fig 2.3(a) Content Based Filtering

The cold start problem is a well known and well researched problem for recommender systems, where system is not able to recommend items to users. due to three different situation i.e. for new users, for new products and for new websites.
Content-based filtering is the method that solve this problem. Our system first uses the metadata of new products when creating recommendations, while visitor action is secondary for a certain period of time. And our systems recommend a product to a user based upon the category and description of the product.

2.4 Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of obects on the basis of similarity and dissimilarity between them.

2.4.1 Clustering Methods :

1. **Density-Based Methods :** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters.

2. **Partitioning Methods :** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimise an objective criterion similarity function such as when the distance is a major parameter. Ex- K-means clustering algorithm.

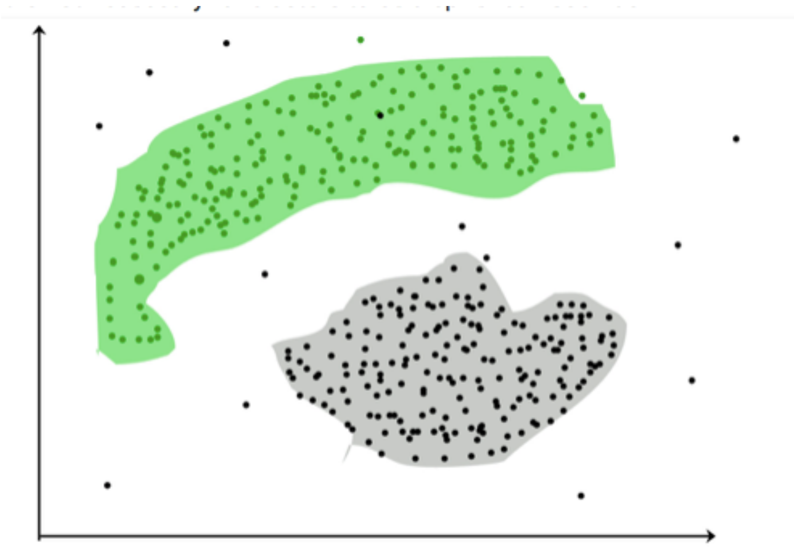


Fig 2.3(b) Cluster formation

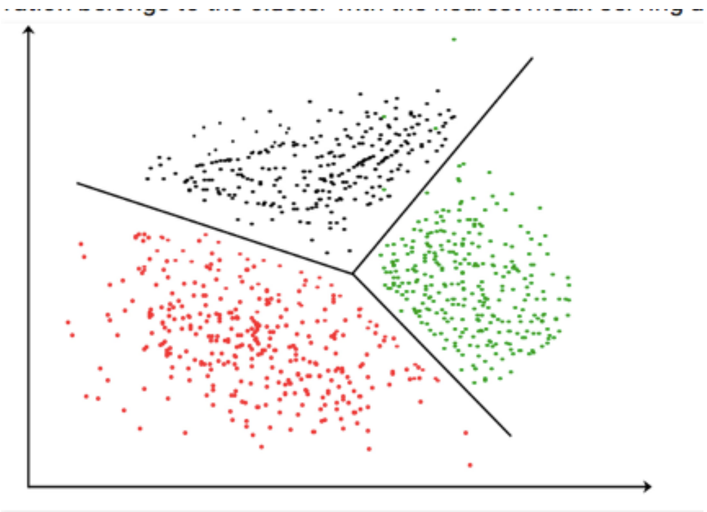


Fig 2.3(c) K-means cluster formation

CHAPTER 3

DESIGN AND METHODOLOGY

3.1 Basic Working

At first the dataset is collected from the required website. Then the dataset is processed (normalisation). After this step the normalised data will be processed in a selected model and then the prediction and error calculation is done.

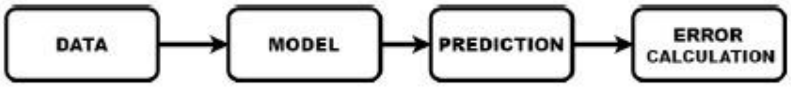


Fig 3.1(a) Basic Model

3.2 System Architecture

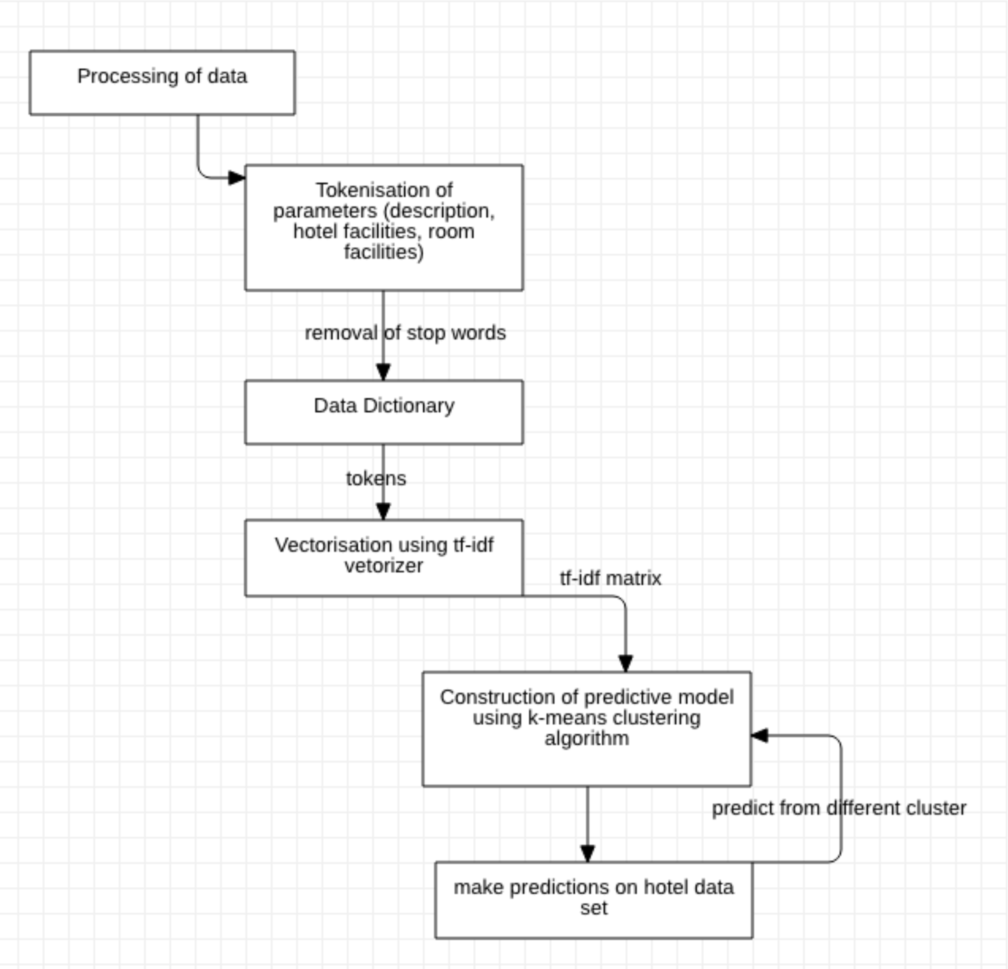

Processing of data
Tokenisation of parameters (description, hotel facilities, room facilities)
removal of stop words
Data Dictionary
tokens
Vectorisation using tf-idf vetorizer
tf-idf matrix
Construction of predictive model using k-means clustering algorithm
predict from different cluster
make predictions on hotel data set

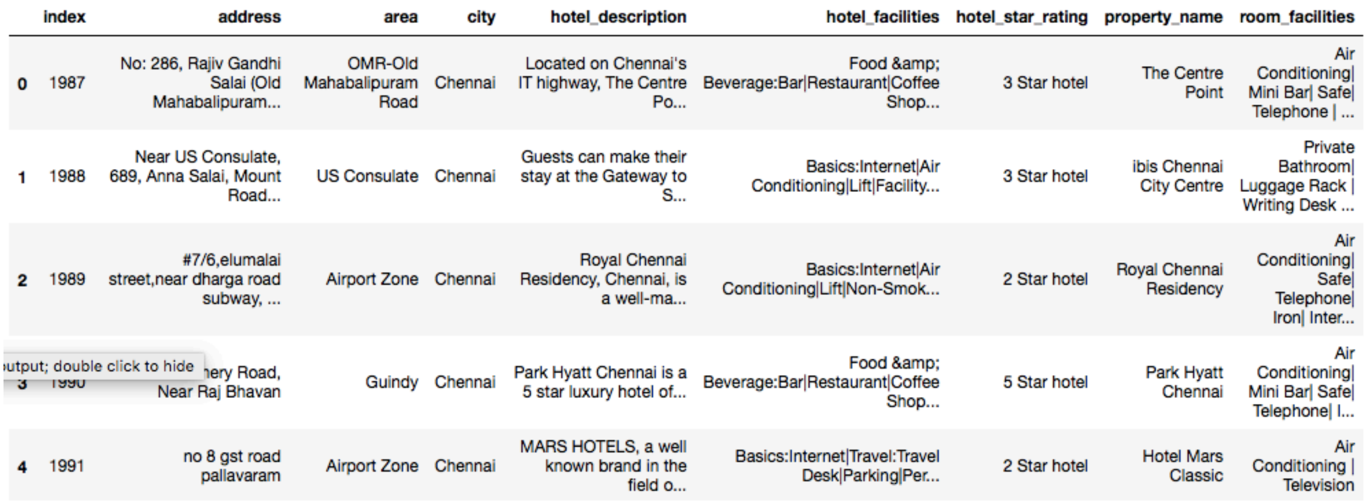
Fig. 3.2(a) System Architecture

The hotel prediction method shown above has mainly four methods:

1. Data Extraction and tokenisation
2. Creating data dictionary
3. Construction of Predictive Model by vectorisation and clustering
4. Prediction

3.2.1 Data Extraction and tokenisation

Hotel cluster data was obtained from cleartrip.com . This hotel booking website offers its dataset from which historical data for various hotels can be obtained having different parameters.



	index	address	area	city	hotel_description	hotel_facilities	hotel_star_rating	property_name	room_facilities
0	1987	No: 286, Rajiv Gandhi Salai (Old Mahabalipuram...	OMR-Old Mahabalipuram Road	Chennai	Located on Chennai's IT highway, The Centre Po...	Food &amp; Beverage:Bar|Restaurant|Coffee Shop...	3 Star hotel	The Centre Point	Air Conditioning| Mini Bar| Safe| Telephone | ...
1	1988	Near US Consulate, 689, Anna Salai, Mount Road...	US Consulate	Chennai	Guests can make their stay at the Gateway to S...	Basics:Internet|Air Conditioning|Lift|Facility...	3 Star hotel	ibis Chennai City Centre	Private Bathroom| Luggage Rack | Writing Desk ...
2	1989	#7/6,elumalai street,near dharga road subway, ...	Airport Zone	Chennai	Royal Chennai Residency, Chennai, is a well-ma...	Basics:Internet|Air Conditioning|Lift|Non-Smok...	2 Star hotel	Royal Chennai Residency	Air Conditioning| Safe| Telephone| Iron| Inter...
3	1990	nery Road, Near Raj Bhavan	Guindy	Chennai	Park Hyatt Chennai is a 5 star luxury hotel of...	Food &amp; Beverage:Bar|Restaurant|Coffee Shop...	5 Star hotel	Park Hyatt Chennai	Air Conditioning| Mini Bar| Safe| Telephone| I...
4	1991	no 8 gst road pallavaram	Airport Zone	Chennai	MARS HOTELS, a well known brand in the field o...	Basics:Internet|Travel:Travel Desk|Parking|Per...	2 Star hotel	Hotel Mars Classic	Air Conditioning | Television

Fig. 3.2(b) cleartrip.com dataset

The above data was then tokenised into a format suitable for use with our prediction model by performing the following steps:

1. Merging the columns hotel_description, hotel_facilities and room facilities
2. Removing stop words from the obtained data of the above three columns

3.2.2 Creating data-dictionary

With the help of tokenised data we get a dictionary of useful words (no stop words). The removal of stop words are analysed by visualising frequency distribution of tokens generated in the above step.

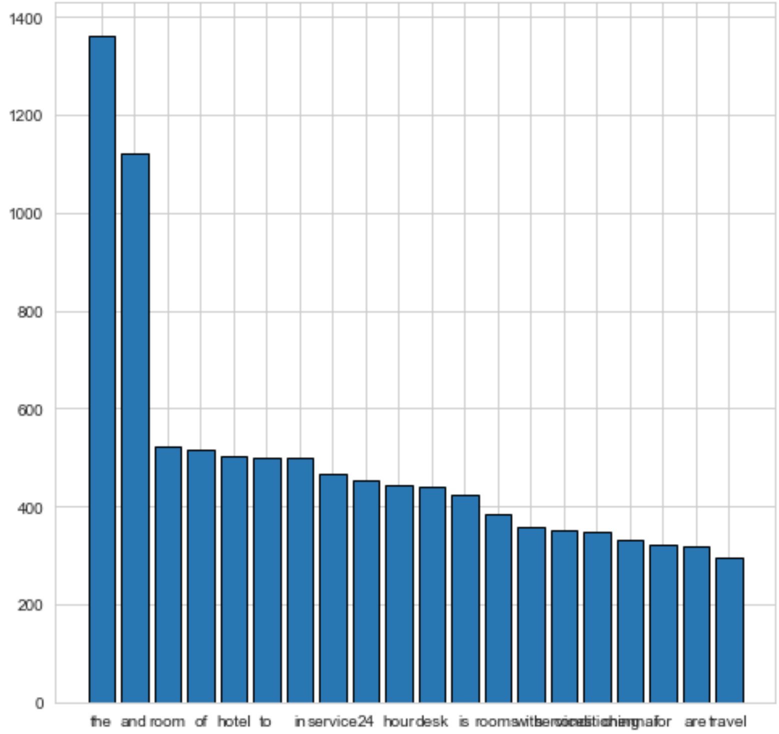
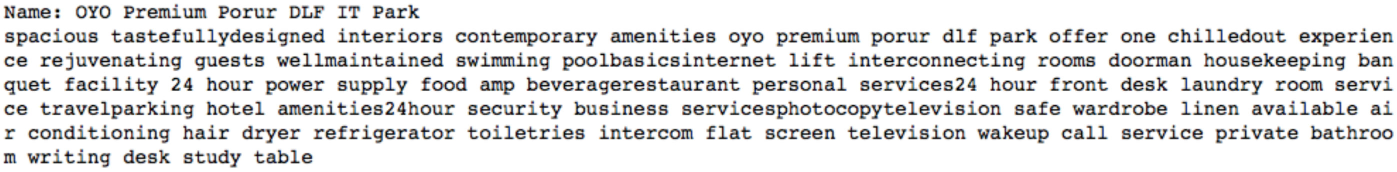


Fig. 3.2(b) Frequency distribution of stop words



```
Name: OYO Premium Porur DLF IT Park
spacious tastefullydesigned interiors contemporary amenities oyo premium porur dlf park offer one chilledout experien
ce rejuvenating guests wellmaintained swimming poolbasicsinternet lift interconnecting rooms doorman housekeeping ban
quet facility 24 hour power supply food amp beveragerestaurant personal services24 hour front desk laundry room servi
ce travelparking hotel amenities24hour security business servicesphotocopytelevision safe wardrobe linen available ai
r conditioning hair dryer refrigerator toiletries intercom flat screen television wakeup call service private bathroo
m writing desk study table
```

Fig. 3.2(c) Data obtained after removal of stop words

3.2.3 Construction of predictive model by vectorisation and clustering

The data dictionary is passed into a TF-IDF (Term Frequency-Inverse Document Frequency) vectoriser to obtain a tf-idf matrix. The obtained matrix is then passed to a k-means clustering algorithm to obtain clusters. Based on distances calculated by the algorithm labels are made which serves as a predicting parameter for the model.

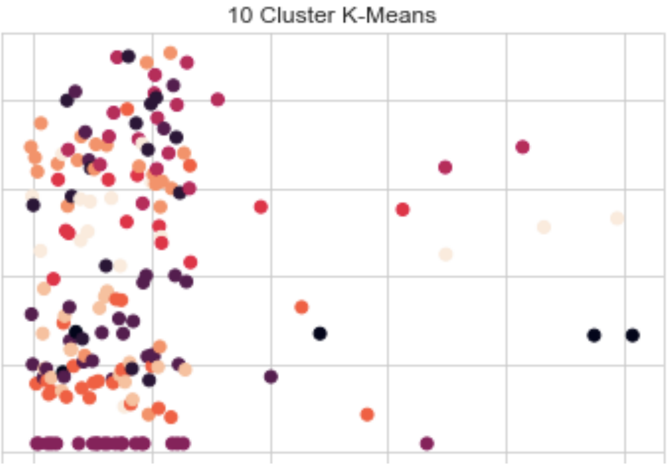
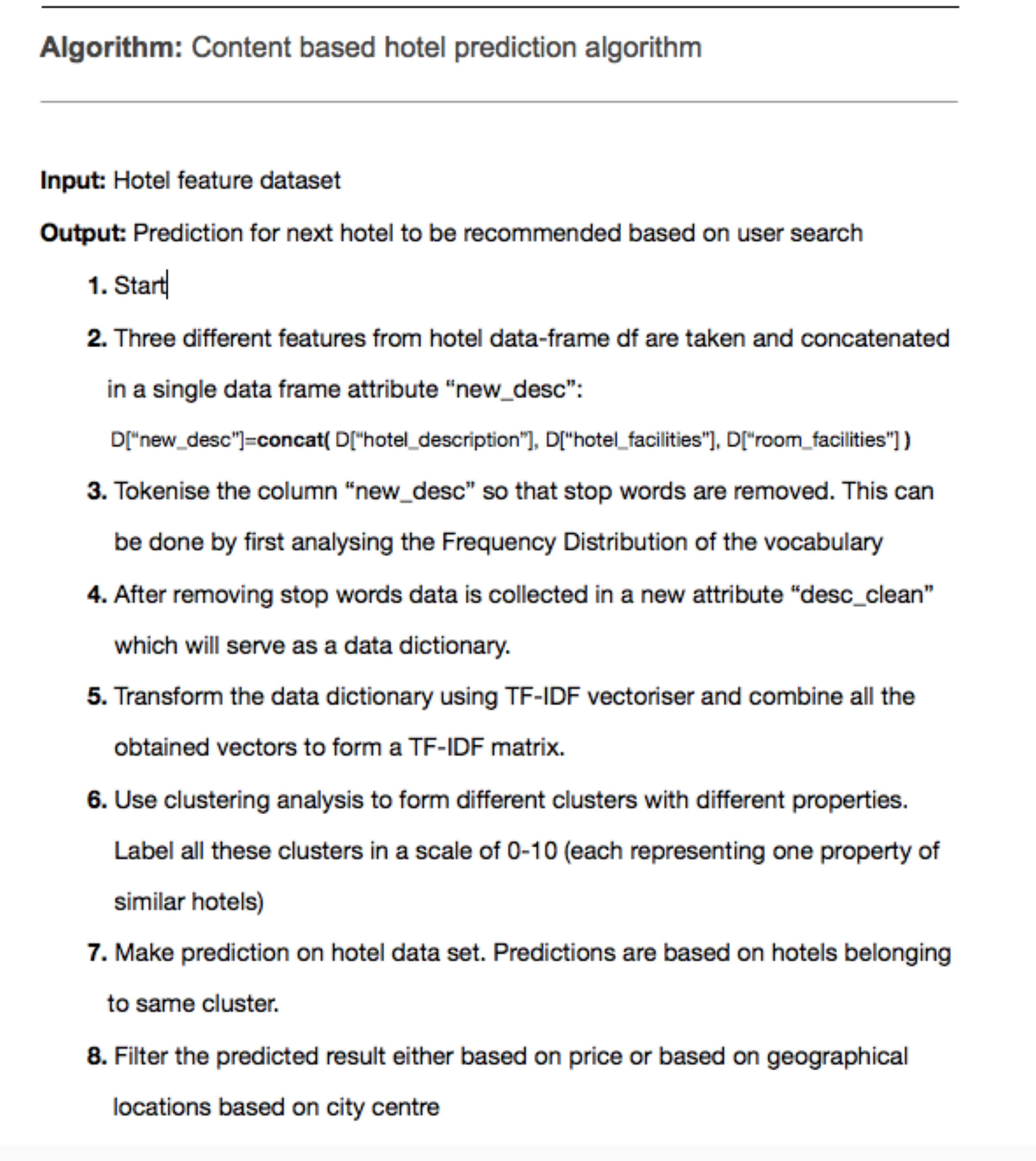


Fig. 3.2(d) Data obtained after removal of stop words

3.2.4 Prediction

Based on distances calculated by the algorithm labels are made which serves as a predicting parameter for the model. These labels define each cluster and predictions are made which belongs to same cluster. These predictions are further filtered based on either price or geographical area of locality as prompted by the user.

3.3 ALGORITHM



Algorithm: Content based hotel prediction algorithm

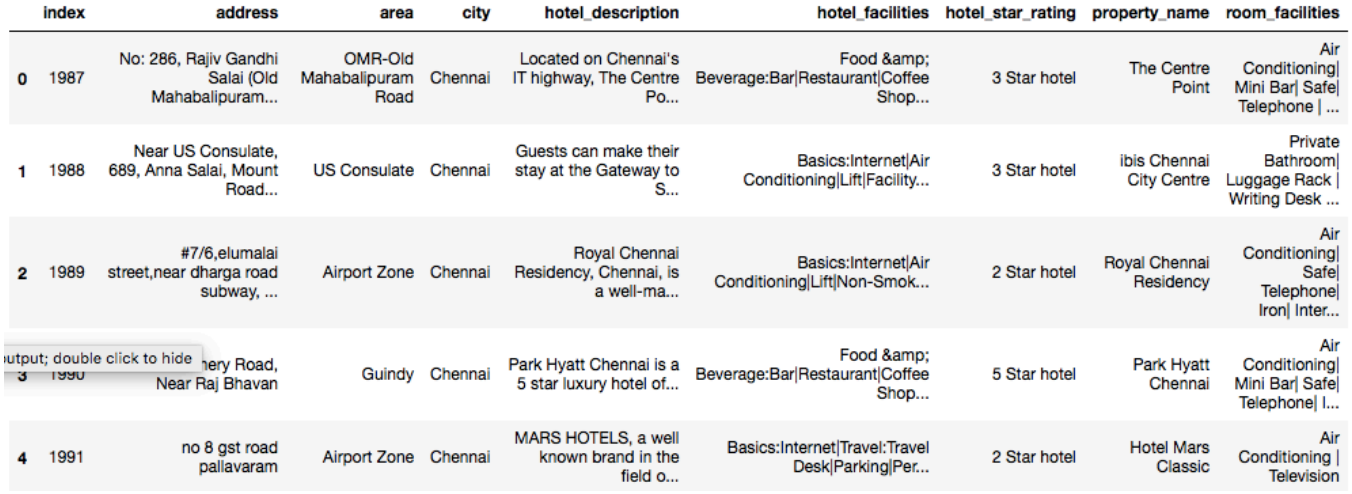
Input: Hotel feature dataset

Output: Prediction for next hotel to be recommended based on user search

1. Start
2. Three different features from hotel data-frame df are taken and concatenated in a single data frame attribute "new_desc":

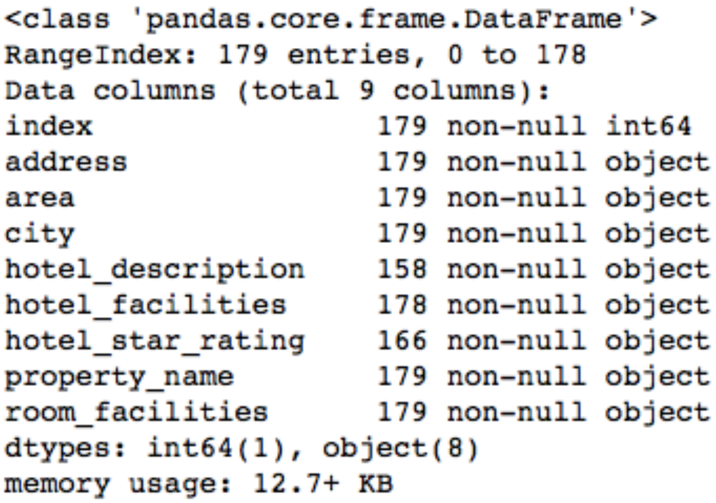
   D["new_desc"]=**concat(** D["hotel_description"], D["hotel_facilities"], D["room_facilities"] **)**
3. Tokenise the column "new_desc" so that stop words are removed. This can be done by first analysing the Frequency Distribution of the vocabulary
4. After removing stop words data is collected in a new attribute "desc_clean" which will serve as a data dictionary.
5. Transform the data dictionary using TF-IDF vectoriser and combine all the obtained vectors to form a TF-IDF matrix.
6. Use clustering analysis to form different clusters with different properties. Label all these clusters in a scale of 0-10 (each representing one property of similar hotels)
7. Make prediction on hotel data set. Predictions are based on hotels belonging to same cluster.
8. Filter the predicted result either based on price or based on geographical locations based on city centre

3.4 Dataset



	index	address	area	city	hotel_description	hotel_facilities	hotel_star_rating	property_name	room_facilities
0	1987	No: 286, Rajiv Gandhi Salai (Old Mahabalipuram...	OMR-Old Mahabalipuram Road	Chennai	Located on Chennai's IT highway, The Centre Po...	Food &amp; Beverage:Bar|Restaurant|Coffee Shop...	3 Star hotel	The Centre Point	Air Conditioning| Mini Bar| Safe| Telephone | ...
1	1988	Near US Consulate, 689, Anna Salai, Mount Road...	US Consulate	Chennai	Guests can make their stay at the Gateway to S...	Basics:Internet|Air Conditioning|Lift|Facility...	3 Star hotel	ibis Chennai City Centre	Private Bathroom| Luggage Rack | Writing Desk ...
2	1989	#7/6,elumalai street,near dharga road subway, ...	Airport Zone	Chennai	Royal Chennai Residency, Chennai, is a well-ma...	Basics:Internet|Air Conditioning|Lift|Non-Smok...	2 Star hotel	Royal Chennai Residency	Air Conditioning| Safe| Telephone| Iron| Inter...
3	1990	hery Road, Near Raj Bhavan	Guindy	Chennai	Park Hyatt Chennai is a 5 star luxury hotel of...	Food &amp; Beverage:Bar|Restaurant|Coffee Shop...	5 Star hotel	Park Hyatt Chennai	Air Conditioning| Mini Bar| Safe| Telephone| I...
4	1991	no 8 gst road pallavaram	Airport Zone	Chennai	MARS HOTELS, a well known brand in the field o...	Basics:Internet|Travel:Travel Desk|Parking|Per...	2 Star hotel	Hotel Mars Classic	Air Conditioning | Television

Fig. 3.4(a) Data-frame head

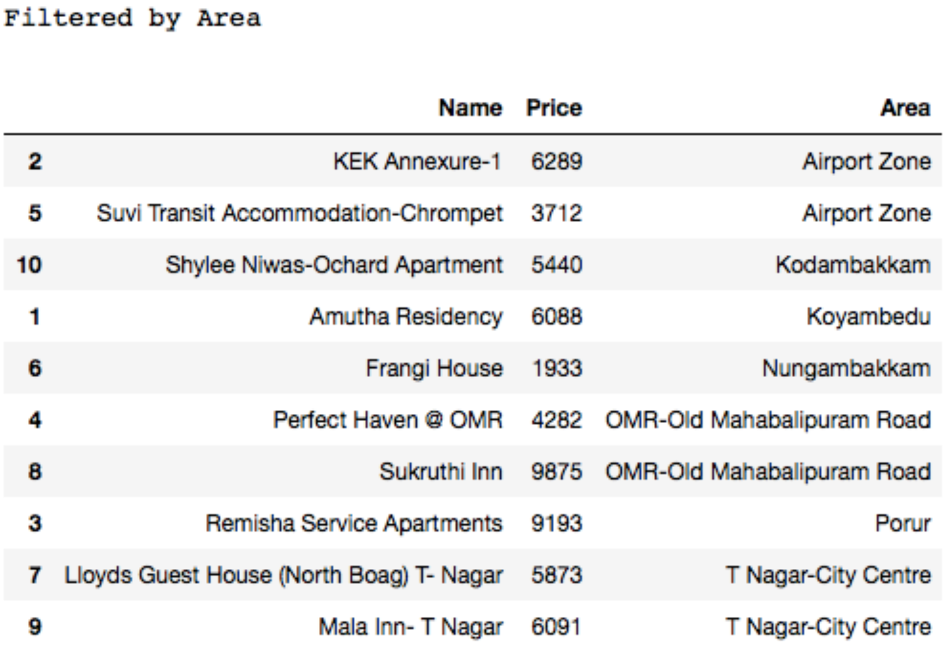


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179 entries, 0 to 178
Data columns (total 9 columns):
index                179 non-null int64
address              179 non-null object
area                 179 non-null object
city                 179 non-null object
hotel_description    158 non-null object
hotel_facilities     178 non-null object
hotel_star_rating    166 non-null object
property_name        179 non-null object
room_facilities      179 non-null object
dtypes: int64(1), object(8)
memory usage: 12.7+ KB
```

Fig. 3.4(b) **Data-frame info**

CHAPTER 4

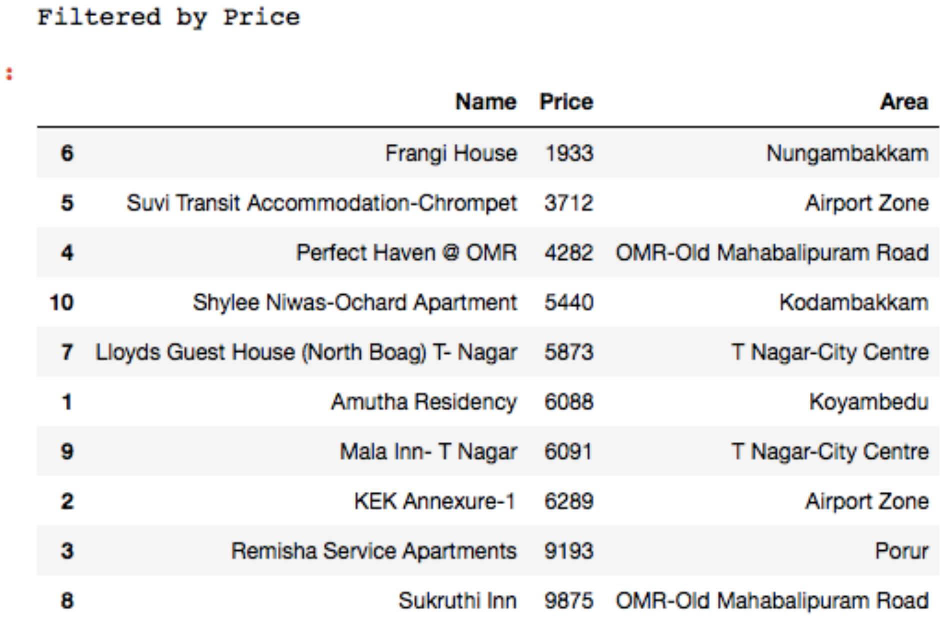
Results



```
Filtered by Area
```

	Name	Price	Area
2	KEK Annexure-1	6289	Airport Zone
5	Suvi Transit Accommodation-Chrompet	3712	Airport Zone
10	Shylee Niwas-Ochard Apartment	5440	Kodambakkam
1	Amutha Residency	6088	Koyambedu
6	Frangi House	1933	Nungambakkam
4	Perfect Haven @ OMR	4282	OMR-Old Mahabalipuram Road
8	Sukruthi Inn	9875	OMR-Old Mahabalipuram Road
3	Remisha Service Apartments	9193	Porur
7	Lloyds Guest House (North Boag) T- Nagar	5873	T Nagar-City Centre
9	Mala Inn- T Nagar	6091	T Nagar-City Centre

Fig 4(a) Recommendations filtered by locality (area)



```
Filtered by Price
```

	Name	Price	Area
6	Frangi House	1933	Nungambakkam
5	Suvi Transit Accommodation-Chrompet	3712	Airport Zone
4	Perfect Haven @ OMR	4282	OMR-Old Mahabalipuram Road
10	Shylee Niwas-Ochard Apartment	5440	Kodambakkam
7	Lloyds Guest House (North Boag) T- Nagar	5873	T Nagar-City Centre
1	Amutha Residency	6088	Koyambedu
9	Mala Inn- T Nagar	6091	T Nagar-City Centre
2	KEK Annexure-1	6289	Airport Zone
3	Remisha Service Apartments	9193	Porur
8	Sukruthi Inn	9875	OMR-Old Mahabalipuram Road

Fig 4(b) Recommendations filtered by price

CHAPTER 5

Conclusion

Content-based recommendation systems may be used in a variety of domains ranging from recommending web pages, news articles, restaurants, television programs, and hotels. The advantage of content-based filtering is that it doesn’t have a cold-start problem. Content-based filtering is the method that solve this problem. Our system first uses the metadata of new products when creating recommendations, while visitor action is secondary for a certain period of time. And our systems recommend a product to a user based upon the category and de scription of the product.

REFERENCES:

[1] Alexandra Roschina, John Cardiff, and Paolo Rosso, TWIN: Personality-based Intelligent Recommender System, Journal of intelligent and Fuzzy Systems (2015)

[2] Michael Arruza, John Pericich, Michael Straka, The Auto mated Travel Agent: Hotel Recommendations Using Machine Learning, June 6 2016

[3] Ryosuke Saga, Yoshihiro Hayashi, and Hiroshi Tsuji, Hotel Recommender System Based on User's Preference Transition, 2008 IEEE International Conference on Systems, Man and Cy bernetics (SMC 2008)

[4] Personalized Hotel Recommendation based on Social Net works Shaowu Liu and Gang Li