San Francisco Airbnb

Sneha Somaya, Yin Yin Teo, Erin Liu, and Chandrima Sabharwal

Introduction

Problem Statement

Simplifying the SF Airbnb Search Process

Our goal is to develop a model that allows customers to find suitable Airbnb's in the city of San Francisco based on pre-selected features such as bedrooms, neighbourhood, price and more. While the Airbnb website offers a few filters to search with, our model goes beyond filtering, exploring similarities across neighbourhoods. Customers also have the option to narrow their search based on how our model clusters Airbnb's!

Project Context & Motivation

User Persona

- By developing a recommendation model that suggests Airbnb’s as per personal choices, we can help various types of customers select Airbnb’s from myriads of options available in San Francisco that be-fit their preferences in a more efficient manner.
- Such a recommendation tool and a user-interactive interface enables user-friendly service and attracts more customers to utilize it.

The Data

➢ The Raw Dataset

The datasets:

- *listings.csv.gz*
- *reviews.csv.gz*

They were downloaded and obtained through the ***Inside Airbnb*** website. It contains publicly sourced, cleansed and aggregated data directly from the actual Airbnb site.

➢ The Raw Dataset: *listings.csv.gz*

accommodates	bathrooms	bathrooms_text	bedrooms	beds	amenities	price	minimum_nights
3	NaN	1 bath	1.0	2.0	["Stove", "Refrigerator", "Coffee maker", "Iro...	$131.00	2
5	NaN	1 bath	2.0	3.0	["Heating", "Smoke alarm", "Dryer", "Wifi", "F...	$235.00	30
2	NaN	4 shared baths	1.0	1.0	["Heating", "Smoke alarm", "Dryer", "Wifi", "C...	$56.00	32
1	NaN	2 shared baths	1.0	1.0	["Stove", "Refrigerator", "Long term stays all...	$45.00	5
2	NaN	4 shared baths	1.0	1.0	["Heating", "Smoke alarm", "Dryer", "Wifi", "C...	$56.00	32

listings.csv.gz contains 6998 rows x 74 columns

- Contains detailed data on SF Airbnb's: listing id, host name, total host listings, neighbourhood location, min/max nights stayed, availability, number of beds/bathrooms, customer review scores, listing price, etc.

➢ The Raw Dataset: *reviews.csv.gz*

date	reviewer_id	reviewer_name	comments
2009-07-23	15695	Edmund C	Our experience was, without a doubt, a five st...
2009-08-03	26145	Simon	Returning to San Francisco is a rejuvenating t...
2009-09-27	25839	Denis	We were very pleased with the accommodations a...
2009-11-05	33750	Anna	We highly recommend this accomodation and agre...
2010-02-13	15416	Venetia	Holly's place was great. It was exactly what I...
...	...	...	...
2020-10-01	275852101	Daniel	It's a very nice place and Bennett is very res...
2020-10-04	352755405	Lauren	Very nice room for a great price!
2020-09-28	72180528	Omer Faruk	Charming house with spacious 3 three bedrooms.
2020-10-03	326482683	Yevgeniy	I was the first one to stay and for a new memb...
2020-10-04	7955718	Alice	Great space which was clean and easy to get to...

reviews.csv.gz contains 316517 rows x 6 columns

- Contains information about customer reviews: date reviewed, name of the reviewer, listing id, reviewer id, and user's comments

Approaches & Algorithms

The Approach

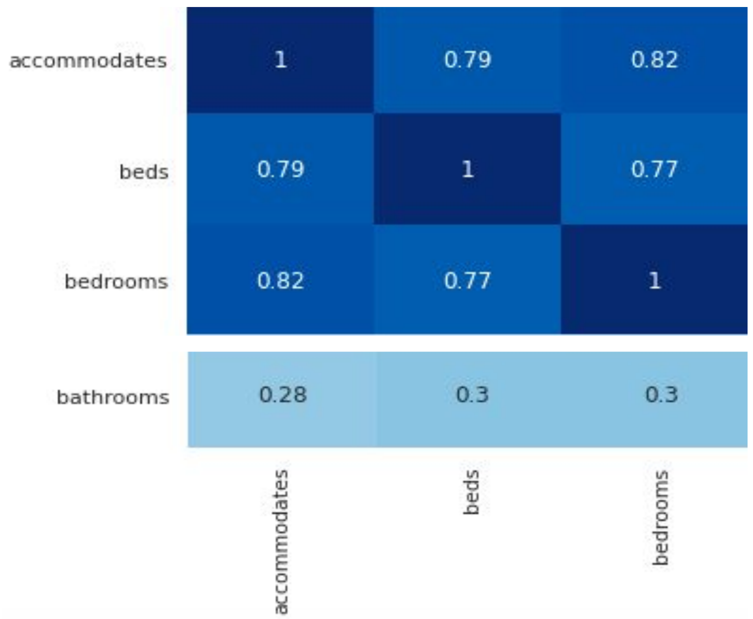
1. Data Cleaning / EDA
2. TF-IDF
3. Sentiment Analysis
4. K-Means Clustering
5. User Interface (Results)

Data Cleaning / EDA

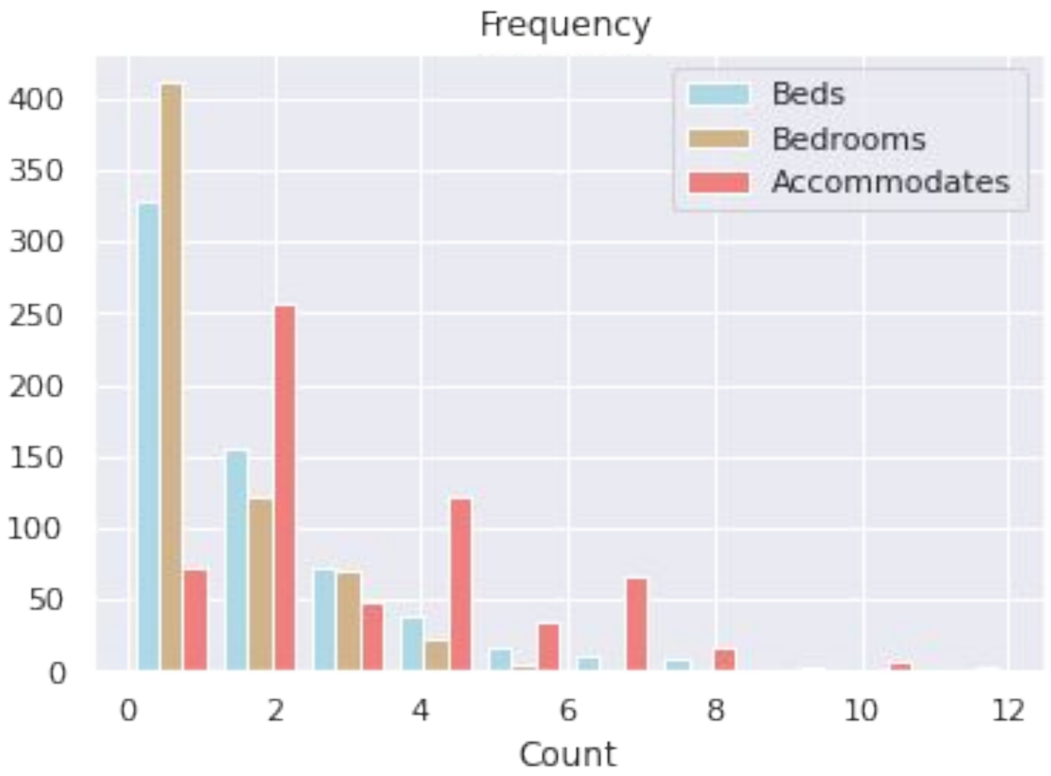
1. First, we performed an inner join on *listings.csv.gz* and *reviews.csv.gz*
2. Next, we dropped unnecessary columns:
    a. 'listing_url', 'scrape_id', 'last_scraped', 'id_x','host_url', 'host_thumbnail_url', 'neighbourhood_group_cleansed', 'minimum_minimum_nights', 'maximum_minimum_nights', 'minimum_maximum_nights', 'maximum_maximum_nights', 'minimum_nights_avg_ntm', 'maximum_nights_avg_ntm', 'calendar_updated', 'license', 'id_y'
3. From the simplified joined table, we looked only at *listing_id* and *comments*, grouping comments by each listing to gather customer opinions on each listing
4. Finally, we separated the joined table based on categorical and numerical values (used for EDA in next few slides)

➢ Correlation Analysis

SF Airbnb's that have more **beds**, **bedrooms**, and **bathrooms** tend to be booked by users with large numbers of occupants, with correlations *.79*, *.82*, and *.28*, respectively.



➢ Data Visualisations



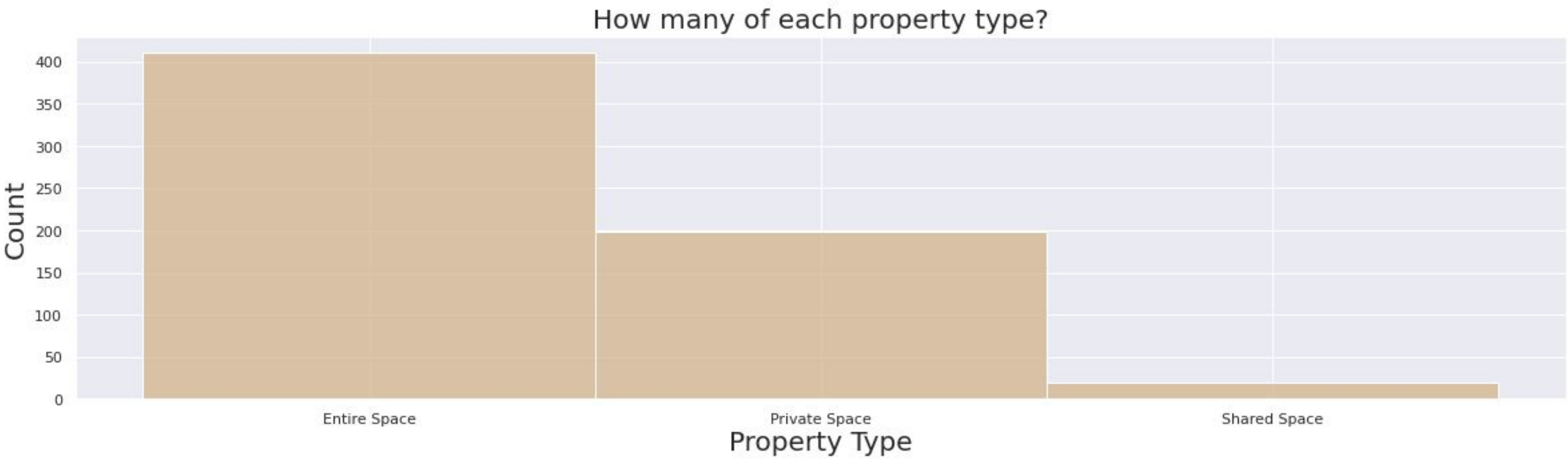
We wanted to understand how the Airbnb's differ based on beds, bedrooms and accommodation. It's quite clear that most Airbnb's are meant for up to 2-4 people with a few that go above 6.

➢ Data Visualisations

It seems that higher ratings given by reviewers address pricier Airbnb listings as compared to lower ratings.

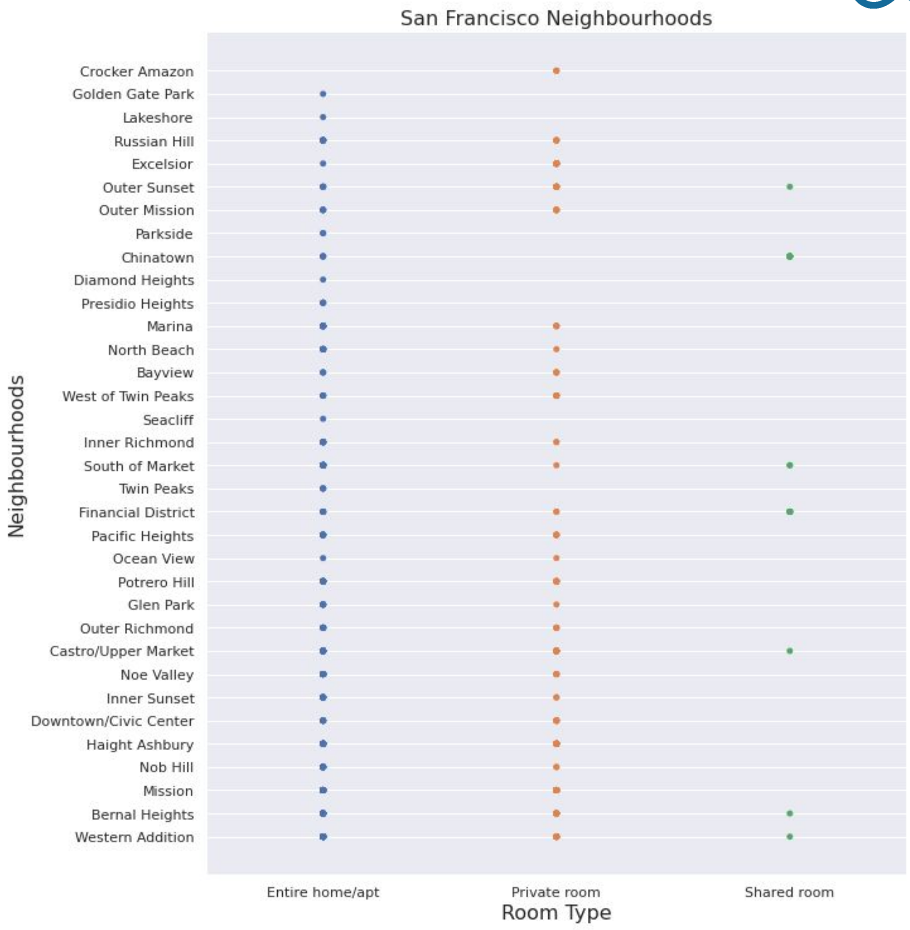
➢ Data Visualisations

Most Airbnb's provide the entire space for stay, and very few offer shared spaces, however it does exist.



➢ Data Visualisations

Almost all neighbourhoods offer entire homes/apartments, and many offer private rooms as well. With few shared rooms available, it is also not as spread out in the city.



Approach & Algorithm - Part 1

1. TFIDF

- We used TF-IDF to quantify the Airbnb descriptions provided by hosts and use it for clustering the listings.

2. Sentiment Analysis

- Goal: To understand how positive or negative users' experiences were with their stay at the different Airbnb's
- Preprocess the text:
  - Lowercase, special characters, stopwords, stemming
- Calculate sentiment score (polarity range: -1 to 1)

Approach & Algorithm - Part 2

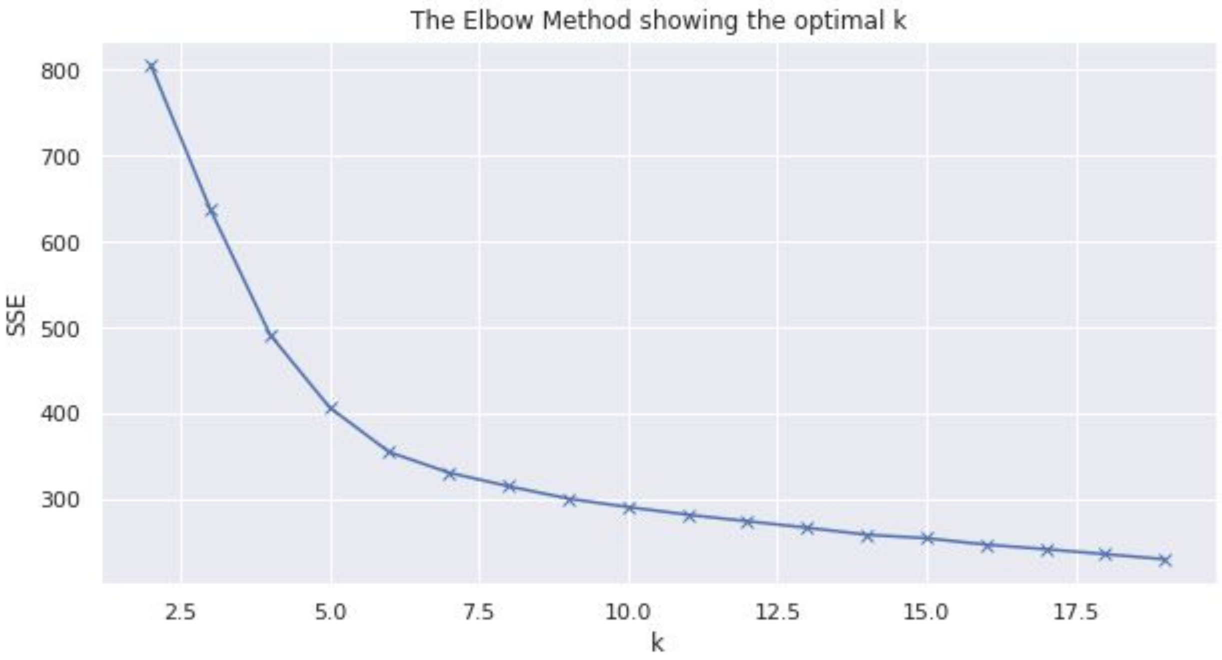
3. K-Means

- Objective: Use K-means algorithm to group similar Airbnb's into clusters based on a combination of key numerical features and text analysis
- Calculated TF-IDF scores on Airbnb descriptions provided by the hosts
- Performed sentiment analysis on Airbnb reviews by users to account for user experience
- Perform recursive clustering on a user's preferred group of Airbnb's to develop more distinct groups of similar Airbnb's
- This allows user to refine their search at each level

Results & Conclusion

Results - KMeans

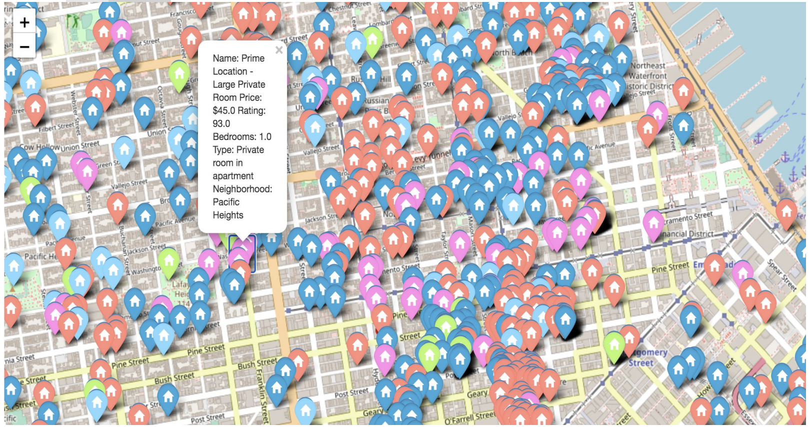
Optimal no. of clusters using elbow method: k = 5



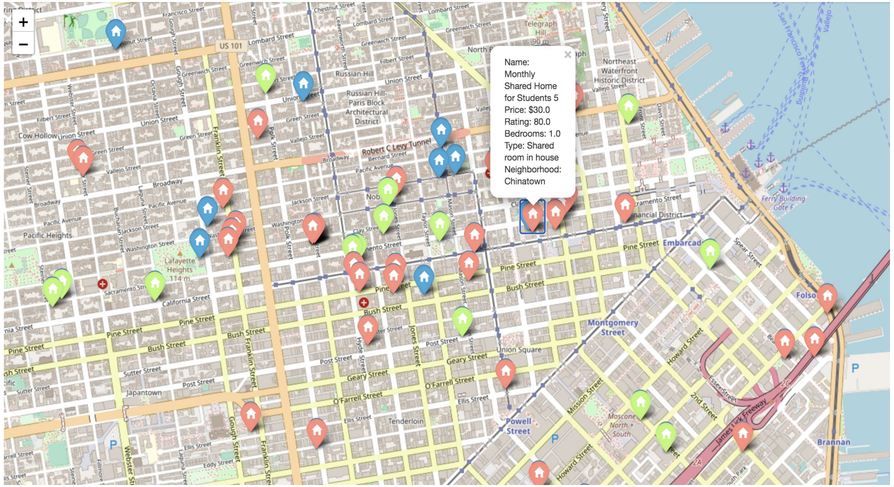
Results - User Interface

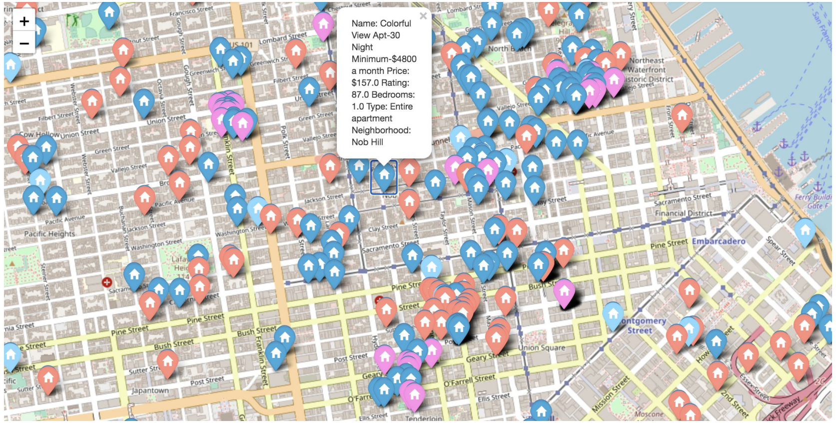
- Used Geopandas and Folium libraries to develop a map of Airbnb listings in San Francisco.
- The map displays the different clusters of Airbnb's in five different colors based on the results of the elbow method for finding the optimal number of clusters (k=5)
- The user can move around, zoom-in/out, and click on a listing to see details of the listing.
- The map interface makes it easier for a user to retrieve information for similar Airbnb's and decide which listing to book.

Map 1: Shows the clusters of all the Airbnb's of San Francisco



Map 3: After choosing group "pink"





Map 2: After choosing group "blue"

Key Findings

- Cluster 4 (light blue) had highest average price
  - Corresponds to highest average no. accommodations/beds/baths
  - But had lowest average TF-IDF & sentiment score
- Clusters 1 (blue) and 3 (pink) had similar average price (also lowest)
  - Both had highest average TF-IDF score among clusters, though not significant enough to conclude anything
  - Cluster 3 can accommodate 1 more person on average but had lower review score
- Clusters 0 (light red) and 3 (pink) had highest average maximum stay
  - One with a lower price point, the other with a higher
- Cluster 2 (light green) had highest average review score but also lowest no. reviews
- Average review sentiment scores relatively stable across all clusters

➢ Conclusion

- Not all Airbnb's within a neighborhood are similar, as one may assume
- Sentiment analysis on users reviews is not a good measure of UX -> bad predictor on similar Airbnb's
- Similar descriptions are not a good predictor of similar Airbnb's
- A slight difference in one feature can significantly affect assigned clusters
- Optimal k changes as you recursively cluster or focus on a specific cluster
- Next steps for improvement:
    - An option for users to select features they deem important and perform clustering based on user selected features
    - Improved UI for users to select & filter clusters

➢ Some challenges we faced...

- The initial steps of this projects that include data collection, processing and exploring were definitely the most time consuming bits - once you have so much data, figuring out next steps and making a plan to uncover interesting insights is not as easy as it seems!
- We wanted to incorporate as much knowledge from our data science courses as we could - learning concepts is simple, but applying them to different scenarios is where it gets interesting.
- Modelling is not as simple as writing a couple lines of code - for a model to work elegantly, fine-tuning and understanding its effect on each feature is important. While our K-Means model does quite well, we know there's place for improvement, through more tuning for more depth in learning.

Thank you!

Appendix

- Data source: http://insideairbnb.com/get-the-data.html
- Link to code (Colab Notebook): https://colab.research.google.com/drive/1EG1FoOHKYmu0Q5WYMwNvkaAdg_Wp6dJI#scrollTo=3IlhbXXZq2Gp