In [1]:
```
import pandas as pd
import nltk
from nltk.corpus import stopwords
```

In [2]:
```
LA_reviews = pd.read_csv("~/Desktop/142_Project/LA_reviews.csv")
Oak_reviews = pd.read_csv("~/Desktop/142_Project/Oak_reviews.csv")
PG_reviews = pd.read_csv("~/Desktop/142_Project/PG_reviews.csv")
Santa_Cruz_reviews = pd.read_csv("~/Desktop/142_Project/Santa_Cruz_reviews.csv")
SC_reviews = pd.read_csv("~/Desktop/142_Project/SC_reviews.csv")
SD_reviews = pd.read_csv("~/Desktop/142_Project/SD_reviews.csv")
SF_reviews = pd.read_csv("~/Desktop/142_Project/SF_reviews.csv")
SM_reviews = pd.read_csv("~/Desktop/142_Project/SM_reviews.csv")
```

In [3]:
```
LA_listings = pd.read_csv("~/Desktop/142_Project/LA_listings_with_zip.csv")
Oak_listings = pd.read_csv("~/Desktop/142_Project/Oak_listings_zip.csv")
PG_listings = pd.read_csv("~/Desktop/142_Project/PG_listings_zip.csv")
Santa_Cruz_listings = pd.read_csv("~/Desktop/142_Project/Santa_Cruz_listings_zip.csv")
SC_listings = pd.read_csv("~/Desktop/142_Project/SC_listings_zip.csv")
SD_listings = pd.read_csv("~/Desktop/142_Project/SD_listings_zip.csv")
SF_listings = pd.read_csv("~/Desktop/142_Project/SF_listings_zip.csv")
SM_listings = pd.read_csv("~/Desktop/142_Project/SM_listings_zip.csv")
```

In [39]:

```
CA_reviews = pd.concat([LA_reviews, Oak_reviews, PG_reviews, Santa_Cruz_
reviews, SC_reviews, SD_reviews, SF_reviews, SM_reviews])
CA_listings = pd.concat([LA_listings, Oak_listings, PG_listings, Santa_C
ruz_listings, SC_listings, SD_listings, SF_listings, SM_listings])
CA_reviews.head()
```

```
/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:2: FutureW
arning: Sorting because non-concatenation axis is not aligned. A future
version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=True'.

To retain the current behavior and silence the warning, pass sort=False
```

Out[39]:

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	109	449036	2011-08-15	927861	Edwin	The host canceled my reservation the day befor...
1	109	74506539	2016-05-15	22509885	Jenn	Me and two friends stayed for four and a half ...
2	344	79805581	2016-06-14	2089550	Drew & Katie	We really enjoyed our stay here in Burbank! Th...
3	344	120725697	2016-12-11	32602867	Christopher	I had a ton of fun learning to play Go with Fu...
4	344	123800867	2016-12-30	35822259	May	The host canceled this reservation the day bef...

In [5]:

```
# Filtering out data in the past three years
CA_reviews = CA_reviews[CA_reviews['date'] >= '2017-01-01']
```

In [6]:

```
# drop rows in LA_reviews with missing columns
CA_reviews.dropna(inplace=True)
```

```
/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:2: Setting
WithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-d
ocs/stable/indexing.html#indexing-view-versus-copy
```

In [7]:

```
# Randomly select 3 reviews from all reviews with the same listing
ids = CA_reviews['listing_id'].unique()
CA_Selected = pd.DataFrame([])
for unique_id in ids:
    curr = CA_reviews[CA_reviews['listing_id'] == unique_id]
    if curr.shape[0] < 3:
        CA_Selected = pd.concat([CA_Selected, curr])
        continue
    temp = curr.sample(3)
    CA_Selected = pd.concat([CA_Selected, temp])
```

In [8]:

```
CA_Selected.shape
```

Out[8]:

```
(170014, 6)
```

In [9]:

```
# The VADER lexicon gives the sentiment of individual words
print("".join(open("vader_lexicon.txt").readlines()[:10]))
```

```
$:      -1.5    0.80623 [-1, -1, -1, -1, -3, -1, -3, -1, -2, -1]
%)      -0.4    1.0198  [-1, 0, -1, 0, 0, -2, -1, 2, -1, 0]
%-)     -1.5    1.43178 [-2, 0, -2, -2, -1, 2, -2, -3, -2, -3]
&-:     -0.4    1.42829 [-3, -1, 0, 0, -1, -1, -1, 2, -1, 2]
&:      -0.7    0.64031 [0, -1, -1, -1, 1, -1, -1, -1, -1, -1]
( '}{' )        1.6     0.66332 [1, 2, 2, 1, 1, 2, 2, 1, 3, 1]
(%      -0.9    0.9434  [0, 0, 1, -1, -1, -1, -2, -2, -1, -2]
('-:    2.2     1.16619 [4, 1, 4, 3, 1, 2, 3, 1, 2, 1]
(':     2.3     0.9     [1, 3, 3, 2, 2, 4, 2, 3, 1, 2]
((-:    2.1     0.53852 [2, 2, 2, 1, 2, 3, 2, 2, 3, 2]
```

In [40]:

```
# Extract sentiment polarity for each word
sentiment = pd.read_csv("vader_lexicon.txt", header = None, delimiter =
'\t', index_col = 0)
sentiment['polarity'] = sentiment[1]
sentiment = sentiment.drop(columns = [1,2,3])
sentiment = sentiment.rename(index={0:'word'})
sentiment.iloc[300:311,:]
```

Out[40]:

	polarity
0	
lmfao	2.5
lmso	2.7
lol	2.9
lolz	2.7
lts	1.6
ly	2.6
ly4e	2.7
lya	3.3
lyb	3.0
lyl	3.1
lylab	2.7

In [11]:

```
# lowercase the test in comments
CA_Selected['comments'] = CA_Selected['comments'].str.lower()
```

In [12]:

```
# get rid of punctuations in comments
# save regex in punc_re

punc_re = r'[^\w\s]'

CA_Selected['no_punc'] = CA_Selected['comments'].str.replace(punc_re, '
 ')
```

```
In [13]: # Remove stopwords
         stop = stopwords.words('english')
         CA_Selected['cleaned'] = CA_Selected['no_punc'].apply(lambda x: ' '.join
         ([word for word in x.split() if word not in (stop)]))
         CA_Selected.head()
```

Out[13]:

	listing_id	id	date	reviewer_id	reviewer_name	comments	no_punc	cle
7	344	339541321	2018-10-21	68239654	Gerardo	the house was beautiful, all the rooms were ni...	the house was beautiful all the rooms were ni...	house beaut rooms home backy
6	344	315417394	2018-08-28	208896709	Lauren	my family had a wonderful stay at melissa’s ho...	my family had a wonderful stay at melissa s ho...	family wond stay meliss home rooms large.
5	344	126784029	2017-03-03	587914	Nicolee	learning to play go with fuko was great fun! w...	learning to play go with fuko was great fun w...	learni play g fuko g fun in fuko k
23	2708	425630420	2019-03-18	13903098	Tami	chas. is an excellent host, and this is a grea...	chas is an excellent host and this is a grea...	chas excell host g locati great space
13	2708	222420570	2017-12-29	155985882	Manami	ﾎｽﾄはとても親切で英語の発音の仕方など丁寧に教えてくれた。\n\n部屋はﾘﾋﾞﾝｸﾞを板で...	ﾎｽﾄはとても親切で英語の発音の仕方など丁寧に教えてくれた \n\n部屋はﾘﾋﾞﾝｸﾞを板で...	ﾎｽﾄは も親切 語の発 仕方な 寧に教 くれた 屋はﾘ を板て って..

In [14]:
```
# Use bag of words to remove infrequent terms.
# Step 1: Tokenize the Sentences. Step 2: Create a dictionary of word fr
equencies
# Step 3: Only keep terms that appear 10 or more times.

CA_wordfreq = {}
for sentence in CA_Selected['cleaned']:
    tokens = nltk.word_tokenize(sentence)
    for token in tokens:
        if token not in CA_wordfreq.keys():
            CA_wordfreq[token] = 1
        else:
            CA_wordfreq[token] += 1
```

In [38]:
```
CA_infreq = [k for (k,v) in CA_wordfreq.items() if v < 10]
CA_infreq[:10]
```

Out[38]:
```
['fuko',
 '4000',
 'players',
 'fundamentals',
 'strategies',
 'ﾎｽﾄはとても親切で英語の発音の仕方など丁寧に教えてくれた',
 '部屋はﾘﾋﾞﾝｸﾞを板で仕切っているのであまりﾌﾟﾗｲﾊﾞｼｰが守られていない',
 '朝になると電話の音やｷｯﾁﾝの音や話し声がそのまま聞こえる',
 'ﾊﾘｳｯﾄﾞの中心地から歩いて15分程なのでﾛｹｰｼｮﾝは最高はとても便利',
 'ｽｰﾊﾟｰやその他ｼｮｯﾌﾟ']
```

In [17]:
```
CA_Selected['cleaned'] = CA_Selected['cleaned'].apply(lambda x: ' '.join
([word for word in x.split() if word not in (CA_infreq)]))
CA_Selected
```

Out[17]:

	listing_id	id	date	reviewer_id	reviewer_name	comments	
7	344	339541321	2018-10-21	68239654	Gerardo	the house was beautiful, all the rooms were ni...	the ho beaut rooms ni...
6	344	315417394	2018-08-28	208896709	Lauren	my family had a wonderful stay at melissa’s ho...	my fa a wor stay a s ho..
5	344	126784029	2017-03-03	587914	Nicolee	learning to play go with fuko was great fun! w...	learni go wi was g w...
23	2708	425630420	2019-03-18	13903098	Tami	chas. is an excellent host, and this is a grea...	chas i excell and th grea..
13	2708	222420570	2017-12-29	155985882	Manami	ﾎｽﾄはとても親切で英語の発音の仕方など丁寧に教えてくれた。\n\n部屋はﾘﾋﾞﾝｸﾞを板で...	ﾎｽﾄは 切で英 音の仕 丁寧に くれた 屋はﾘ 板で..
19	2708	306500549	2018-08-12	48408491	Andrew	charles place is lovely and in a great locatio...	charle lovely great
45	2732	160591707	2017-06-14	13797405	Edward D	all was as described. i was totally at home a...	all wa descr was t home
47	2732	501542842	2019-08-03	144393977	Yizhou	i stayed here for 3 months. it’s definitely a ...	i staye for 3 r s defi
46	2732	348918624	2018-11-15	178015720	Жибек	it was unforgettable vacation. everything was ...	it was unforg vacati every ...

	listing_id	id	date	reviewer_id	reviewer_name	comments	
180	5728	213120452	2017-11-19	1681073	Nick	a quiet oasis of calm and good vibes. the tiny...	a quie calm vibes
285	5728	435213686	2019-04-09	106737024	Katherine	a hidden gem in a lovely neighborhood. small i...	a hidc in a lc neigh small
276	5728	419753324	2019-03-04	245914489	Monica	sweet little place, full of hidden quirks. i w...	sweet place hidde w...
421	5729	252941277	2018-04-12	4950905	Federico	great place and location, beautiful design, ma...	great and lc beaut desig
508	5729	466881110	2019-06-09	2264742	Su Gyeong	this place is awesome! you will never regret c...	this p awes will ne regret
491	5729	439320517	2019-04-18	180153470	Zac	the place was amazing!\nwould recommend if com...	the pl amazi \nwou recom com..
598	5843	350875555	2018-11-20	11484819	Capucine	great place, very nice garden, easy checkin an...	great very r garde check
633	5843	519031403	2019-08-27	9831031	Angelica	sanni & helene's home is a gourgeus cottage in...	sanni home gourg cottag
557	5843	199743329	2017-10-02	4860311	Sun	we used to live in marina del rey for 7 years ...	we us in ma rey fo ...
656	6033	228064229	2018-01-17	164733571	Chelsea	located just outside of the city and near all ...	locate outsic city ai all ...

	listing_id	id	date	reviewer_id	reviewer_name	comments	
654	6033	220463387	2017-12-22	47853584	Elba	the host canceled this reservation the day bef...	the hc cance reserv day b
652	6033	140777627	2017-03-31	70785841	Lauren	lovely host, bright and peaceful room, thank y...	lovely bright peace thank
670	6931	450227800	2019-05-08	68031652	Marcus	charles is a superhost for a reason. i stayed ...	charle super reaso ...
668	6931	271248096	2018-05-31	175283825	Joey	i stayed with chas for two months and he was t...	i staye chas month was t.
672	6931	474102167	2019-06-22	224210468	Dan	easy, simple and nice stay. location is wonder...	easy s and n locati wond
679	7874	437492058	2019-04-14	255288923	Thomas	great host. everything described is accurate.	great every descr accur
681	7874	476203935	2019-06-25	253858949	Esther	stayed for 5 weeks. great host, clean and comf...	staye weeks host c comf.
675	7874	365824943	2019-01-01	77621246	Eunjae Liv	cozy house with friendly and nice family. high...	cozy l with fi and n high..
839	7992	409773799	2019-02-08	79413694	Karima	the place is amazing, tucked away in a cozy ne...	the pl amazi tucke a coz

	listing_id	id	date	reviewer_id	reviewer_name	comments	
731	7992	139277496	2017-03-24	7880962	Gillian	great last minute stay at tom's place. had a ...	great minut tom s had a
771	7992	227430826	2018-01-14	85347283	Jesse S	this was my first experience with air bnb, and...	this w first e with a and...
...	...	...	...	...	...	...	...
366633	38979868	541130081	2019-10-04	269782288	Victor	awesome place! great location and excellent va...	awes place locati excell
366635	38984642	544691900	2019-10-10	299968663	Kevin	.	
366636	38984642	545637912	2019-10-12	218094678	Chan	joyce had been a great host for my weekend tri...	joyce a grea my w tri...
366637	38985258	546586277	2019-10-13	281436537	Matthew	this place was a amazing. sonya was great and...	this p a ama sonya great
366638	39059847	546662295	2019-10-13	163121515	Cynthia	what a steal for the price!\nmy partner and i ...	what the pr partn
366639	39137417	543048610	2019-10-07	225801399	Deonicia	the host canceled this reservation 5 days befo...	the ho cance reserv days
366640	39157824	544269028	2019-10-09	300834900	Leah	communication was great and the place was real...	comm was g the pl real...
366641	39225180	546511107	2019-10-13	211726938	John	nice location and a great view of the ocean. ...	nice l and a view ocea

	listing_id	id	date	reviewer_id	reviewer_name	comments	
366642	39258642	544849536	2019-10-11	2627694	Joon	the host canceled this reservation 7 days befo...	the ho cance reserv days
96	208345	328283145	2018-09-25	131989446	Tom	very relaxed environment, angela and her famil...	very r envirc angel famil..
90	208345	260705300	2018-05-04	4477059	Dom	wonderful place, i highly recommend!	wond place recom
84	208345	129424413	2017-01-30	35840150	Kazuo	my son and myself stayed at their loft and enj...	my sc mysel at the enj...
270	796558	187881815	2017-08-27	73695579	Krzysztof	so. \nthis was my first abnb experience, hones...	so \nt my fir exper hones
269	796558	184477587	2017-08-19	93427832	Yuanjian	lisa is very nice and the room is very comfort...	lisa is and th is ver comfc
282	796558	212018918	2017-11-15	158708574	Stephen	close to the airport and walkable to the many ...	close airpor walka many
381	1126135	402512262	2019-01-18	114749813	Yiwei	good place to stay!	good stay
382	1126135	426693134	2019-03-21	93664141	Akira	this is a great place to stay if you need to b...	this is place you n b...
377	1126135	138669956	2017-03-20	121178686	Garrick	great experience. everyone open and accepting...	great exper every and accep
405	1157718	134877123	2017-03-01	108577218	Raquel	me gustó muchísimo, la atención , las personas...	me gu much atenc perso

	listing_id	id	date	reviewer_id	reviewer_name	comments	
416	1157718	313664317	2018-08-25	190298870	William	vic is the best. perfect to live at during sum...	vic is perfec at dur sum..
407	1157718	153283797	2017-05-20	49818055	Dr. Michael J.	best place to stay in the south bay. location ...	best p stay in south locati
451	1334378	189740840	2017-09-02	145422726	Vicente	very nice and quiet place.	very n quiet
454	1334378	223803475	2018-01-01	33672517	Kéf	i stayed at the menlo park hackerhome over chr...	i stay menlo hacke over c
465	1334378	343776599	2018-11-01	95618652	Hikari	hackerhome is similar to a hostel in taiwan or...	hacke simila hoste taiwan
495	1602634	246107838	2018-03-24	178556632	Wilson	excellent! great hospitality and even provides...	excell hospit even provid
492	1602634	134734562	2017-02-28	30724210	Luis	people i shared the space was great, it is pre...	peopl the sp great
498	1602634	428285876	2019-03-24	148754820	Linda	i went in late march and the place was pretty ...	i went march place pretty
517	1602655	442567503	2019-04-23	177275886	Duane	convenient location. friendly roommates. ...	conve locati friend roomm
512	1602655	140761568	2017-03-31	30724210	Luis	great place!	great
515	1602655	346188380	2018-11-07	221012382	Troy	it was great staying at a place that valued th...	it was stayin place valued

```
170014 rows × 8 columns
```

In [18]:

```
# convert the comments into a tidy format to make sentiments easier to calculate
tidy_format = CA_Selected['cleaned'].str.split(expand = True).stack().reset_index(level = 1).rename(columns={'level_1': "num", 0:"word"})
tidy_format
```

Out[18]:

	num	word
7	0	house
7	1	beautiful
7	2	rooms
7	3	nice
7	4	homey
7	5	feel
7	6	backyard
7	7	nice
7	8	place
7	9	relax
7	10	long
7	11	day
7	12	pool
7	13	great
7	14	house
7	15	super
7	16	close
7	17	shopping
7	18	areas
7	19	surrounded
7	20	stores
7	21	places
7	22	eat
7	23	family
7	24	really
7	25	enjoyed
7	26	stay
6	0	family
6	1	wonderful
6	2	stay
...	...	...
498	76	space

	num	word
498	77	bunk
498	78	beds
498	79	lots
498	80	closet
498	81	space
498	82	extra
498	83	hangers
498	84	separate
498	85	larger
498	86	bathroom
498	87	room
498	88	bit
498	89	different
517	0	convenient
517	1	location
517	2	friendly
517	3	roommates
517	4	kind
517	5	college
517	6	dorm
517	7	room
517	8	feel
512	0	great
512	1	place
515	0	great
515	1	staying
515	2	place
515	3	valued
515	4	guests

3752036 rows × 2 columns

In [19]:

```
# for each comment, find the sentiment of each word.
# Calculate the sentiment of each comment by taking the sum of the senti
ments of its words.

df = pd.merge(sentiment, tidy_format, how = 'inner', left_index = True,
right_on = 'word')

CA_Selected['review_polarity'] = df.groupby(df.index)[['polarity']].agg(
sum)['polarity']
CA_Selected[['review_polarity']] = CA_Selected[['review_polarity']].fill
na(value=0)
```

In [20]:

```
CA_Selected = CA_Selected[['listing_id', 'review_polarity']]
CA_Reviews = CA_Selected.groupby('listing_id').mean().round(2)
```

In [21]:

```
CA_Reviews.head()
```

Out[21]:

	review_polarity
listing_id	
6	19.37
344	18.30
958	11.60
2708	13.37
2732	19.20

In [22]:

```
CA_listings = CA_listings[['id', 'zip']]
CA_listings.head()
```

Out[22]:

	id	**zip**
0	109.0	90056
1	344.0	91506
2	2708.0	90046
3	2732.0	90405
4	2864.0	90706

In [31]:

```
df = pd.merge(CA_listings, CA_Reviews, how='inner', left_on='id', right_
index=True).set_index('zip')
Reviews = df[['review_polarity']]
Reviews.head()
```

Out[31]:

	review_polarity
zip	
91506	18.30
90046	13.37
90405	19.20
90066	14.13
90066	7.47

In [30]:

```
temp = pd.read_csv("CA_Review.csv")
temp.head()
```

Out[30]:

	zip	review_polarity
0	90001	6.72
1	90002	6.14
2	90003	6.27
3	90004	7.40
4	90005	7.35

In [29]:

```
result = temp.groupby("zip").mean().round(2)
result.head()
```

Out[29]:

	review_polarity
zip	
90001	6.72
90002	6.14
90003	6.27
90004	7.40
90005	7.35

In [27]:

```
result.to_csv(r'CA_Review.csv')
```