Which option is better?
Analysis on On-Campus Housing Vs Off-Campus Housing

Using data scraped from top subreddits of universities in US

Jeet Ketan Thaker

CSC 440 - Final Project

Fall 2019

INTRODUCTION

Background and Motivation

Every student who wants to pursue a higher education asks him/herself this question at some point. “Which is better among On-Campus vs. Off-Campus housing?” This is an important question after all. It decides the place the student will be living and spending most of his/her time during their studies. The choice of taking one type of housing against another can have a huge impact on the student’s study patterns and life in school in general. There seems to be no particular method to once and for all decide which housing option is better. The type of housing definitely depends from student to student and lots of other factors. But this is not to say we cannot infer anything about this question.

In this study what we will attempt to do is to get “*the wisdom of the crowd*”. Even though the exact answer to the question of housing needs will change from student to student, we can still ascribe to the wisdom of the crowd and see what is the collective opinion of thousands of students across the nation. In this study we will run a sentiment analysis and figure out the sentiment of students who chose the option of on-campus housing and compare it against the students who chose off-campus housing. For our study we will be limiting ourselves to universities in the US and taking the collective sentiment of students attending those universities. The United States has universities spanning its lands and further in the study we would also like to cluster universities according to the four cardinal directions North , South , West and East. This will help us in understanding if the sentiment changes with different clusters.

Below is a map of the US showing how universities are located in different areas :

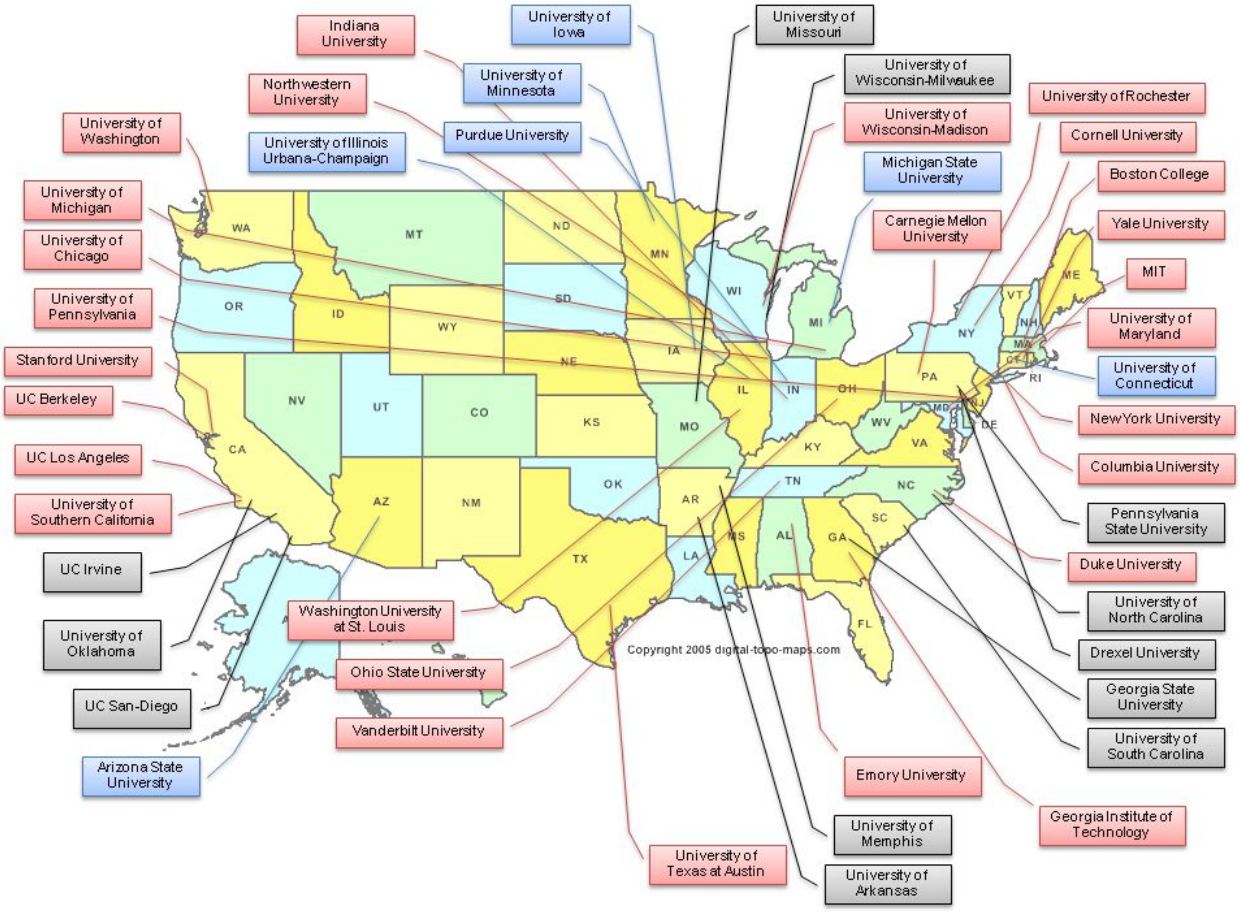


Figure 1 : A map of the US showing universities in different locations (Source : https://aacademy.files.wordpress.com/2009/09/university_map.png)

Problem Statement

Does the happiness level of a student vary with the housing option he/she chooses?

For example : For universities located in the north which face a lot of snow, a student would be more happy to be on campus as to avoid travelling. Along with this there are multiple factors contributing to the happiness level of a student in an environment.

Problem Importance

As iterated before , this question of housing is important. Every student is confused and thinks about which housing option should he go for. Happiness levels are not only determined by the housing option you choose after all, it very much so is also dependant on the environment. By doing this study we are trying to find how a students happiness level would change given the on-campus or off-campus *environment.*

For example : Doing this study we got to know about how in one university, dog-walking was normal on campus and by choosing to live on-campus students had opportunities to interact with the pet and their sentiment which we calculated through their posts were joyous.

METHODOLOGY

Data Collection

The data is collected scraping the social media platform “Reddit”. We tried collecting data from “Twitter” too but it had some limitations which we will elaborate on here.

The main problem collecting Twitter data for our use case was the disbalance of samples between on-campus and off-campus candidates. The way to find on-campus students was to use querying using hashtags and similarly for off-campus students. The samples we got from querying for off-campus housing were mostly parents who were studying part-time or people with jobs who were studying part-time. The samples for on-campus housing consisted of mostly teens and students just starting their undergraduate degree. To compare between the sentiments of these two groups would be ill-advised. As parents or people with jobs they have a whole set of responsibilities other than which the samples from the on-campus search have. If there were a difference in happiness levels between these groups we could not be sure whether it is because of the location of their housing needs or because of some other factors. Besides we couldn't find many people tweeting about off-campus activities and such.

Reddit, in this case provided a good solution. Reddit has these specialized *subreddits* where people from a particular community can post/talk about topics concerning a particular topic. Also, it is a platform that many undergraduate and graduate students use which makes it advantageous for our purposes.

We searched for the most subscribed *university subreddits* on this platform. These are subreddits dedicated to a particular university like for our university of Rochester we have the r/UofR. In where people talk about events and things concerning only about UofR.

The universities we considered for this project are (in descending order of population in terms of subscriber counts) :

- University of Illinois at Urbana-Champaign
- University of California - Berkeley
- University of Texas - Austin
- University of Waterloo
- Texas A&M
- Georgia Institute of Technology
- Ohio State University
- Rochester Institute of Technology (RIT)

- University of Toronto
- Virginia Tech
- University of Central Florida
- University of Michigan
- University of California - Los Angeles
- Rutgers University
- Purdue University
- Penn State University
- University of Maryland
- University of British Columbia
- University of California - Santa Barbara
- Michigan State University
- University of California - San Diego
- University of Georgia
- Arizona State University
- University of Washington
- Florida State University
- University of Pittsburgh
- University of Wisconsin Madison
- Rensselaer Polytechnic Institute
- McGill University
- University of California - Davis
- North Carolina State University
- University of Minnesota
- California Polytechnic State University
- University of California - Santa Cruz
- University of Connecticut
- Louisiana State University
- Indiana University
- University of Southern California
- Massachusetts Institute of Technology
- University of Houston
- Northeastern University
- University of Virginia
- University of Florida
- New York University - NYU

(*Source : A reddit comment by the user, first_user_cs)*

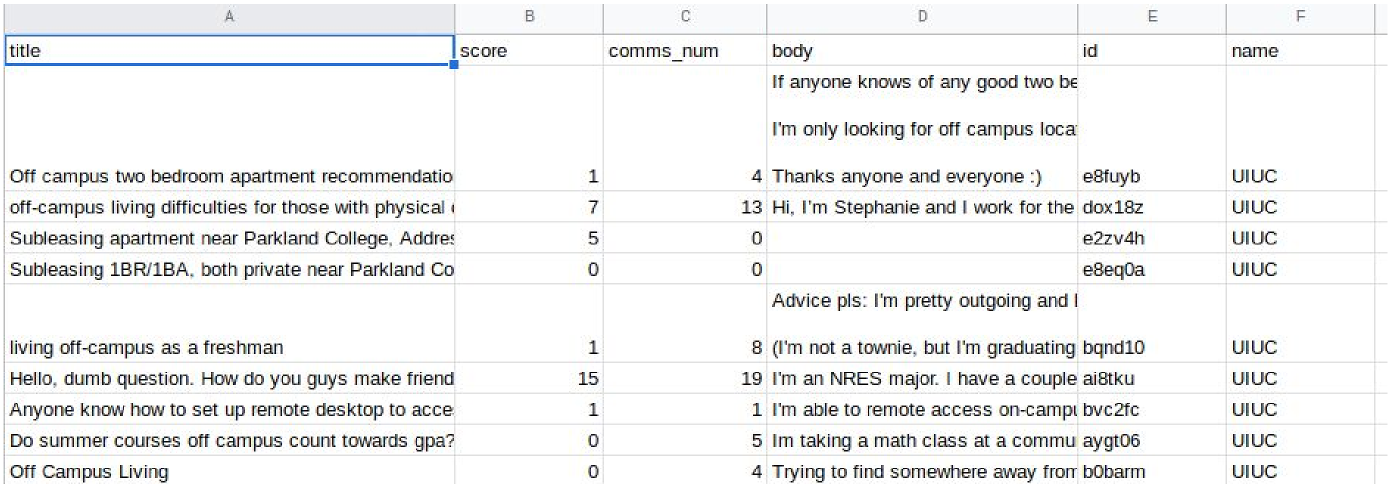
For data collection, we scraped subreddits of all the above listed universities with search keywords , “on campus” , “off campus” , “housing” and “dorm”.

The data collected from each post was :

- The title of the post
- The body of the post
- The karma the post gathered (analogous to facebook likes)
- The number of comments on the post
- The id of the post (each post on reddit has a unique id)

Once the data was collected it was sorted and converted into 4 different csv files. One for each of the search query. This data has all the information about the post elucidated above plus an additional field pointing to the name of the university from which the post was gathered.

Below is a screenshot of the data from “off_campus.csv” that we made



A	B	C	D	E	F
title	score	comms_num	body	id	name
Off campus two bedroom apartment recommendatio	1	4	If anyone knows of any good two be I'm only looking for off campus loca Thanks anyone and everyone :)	e8fuyb	UIUC
off-campus living difficulties for those with physical	7	13	Hi, I'm Stephanie and I work for the	dox18z	UIUC
Subleasing apartment near Parkland College, Addres	5	0		e2zv4h	UIUC
Subleasing 1BR/1BA, both private near Parkland Co	0	0		e8eq0a	UIUC
living off-campus as a freshman	1	8	Advice pls: I'm pretty outgoing and I (I'm not a townie, but I'm graduating	bqnd10	UIUC
Hello, dumb question. How do you guys make friend	15	19	I'm an NRES major. I have a couple	ai8tku	UIUC
Anyone know how to set up remote desktop to acce	1	1	I'm able to remote access on-campu	bvc2fc	UIUC
Do summer courses off campus count towards gpa?	0	5	Im taking a math class at a commu	aygt06	UIUC
Off Campus Living	0	4	Trying to find somewhere away from	b0barm	UIUC

Figure 2 : Screenshot of data from ‘off-campus.csv’

Sentiment Analysis using this data

We will be doing sentiment analysis on this dataset , which is 4 csv files named, “off-campus.csv” , “on-campus.csv” , “housing.csv” and “dorm.csv”.

The way we will do this is by using *VADER* which is a tool used for sentiment analysis. We will be making use of this tool for our needs. It takes in a text and gives out scores for the positive sentiment in the text , the negative sentiment in the text , the neutral sentiment in the text and the combination of all the above the compound sentiment in the text. Since VADER works only for one text at a time we will have to code it up so it can take in datasets and give out the positive , negative and neutral sentiment over the whole dataset.

To do sentiment analysis on our data we will need to clean it up a little. We will go through the dataset (the title , the body , the id , the karma of each post) and replace any NaN values by 0. We do this as VADER cannot handle NaN values. When given an input of 0 to VADER it gives out a zero sentiment, so replacing NaN’s by 0 won't affect our sentiment scores.

Next step in our preprocessing will be to add 1 to karma value of every post. Like said before, VADER cannot give you a sentiment for a dataset. It only gives sentiments corresponding to a single text. The way we make VADER compute sentiment of our data is by getting a weighted average of sentiments of each post. The weights are the karma values associated with the particular post. This is done so we have a fair representation of sentiment values. If a post’s sentiment resonates within people reading the post they increase the karma value by one otherwise they decrease it by one. Hence if there is a post which has a positive sentiment and not many people have upvoted the post we can assume that the event was not happy in general and was very specific in the way it induced happiness in the poster. In contrast, if there is a post with a positive sentiment which also has a large number of upvotes we can assume it was a very happy event and lots of people agree with the sentiment and therefore we must give more representation to events like those. It is for this reason we do a weighted average of sentiments. This is why we need to add 1 to karma values of every post , because if a post has 0 karma then it wouldn't be represented at all.

Sentiment analysis on the entire dataset

We load the csv file named "off_campus.csv" and "on_campus.csv" into python. Each of these contain the title , body and karma values of the posts contained within itself. We will use both the title and body which are both just plain texts written in natural language to calculate the sentiment value of the post. We will calculate the sentiment obtained from the title of the post and then calculate the sentiment obtained from the body of the post and average the result in order to obtain the sentiment of a single post. This is done as sometimes sentiment is missing from titles or sometimes it is missing from the body or sometimes there is no body or sometimes there is no title. We will average the sentiment of the two in order to get the true sentiment of the post.

This is how we find sentiment of a single post :

- Sentiment_of_title = VADER_SENTIMENT(post['title'])
- Weighted_Sentiment_title = post['score'] * Sentiment_of_title
- Sentiment_of_body = VADER_SENTIMENT(post['body'])

- Weighted_Sentiment_body = post[‘score’] * Sentiment_of_body
- Weighted_Sentiment_post = (Weighted_Sentiment_title + Weighted_Sentiment_body) /2

Where score is the karma value attributed to the post. The above “Weighted_Sentiment_post” will give us the sentiment value of a single post.
Now that we have the sentiment from a single post we can find the sentiment of the entire dataset. We will go through all the posts in “off_campus.csv” and for each post record the sentiment values of each post using the above method. Finally once we have all the sentiment values (pos , neg , neu) we can take the average of these values to obtain the sentiment of the dataset. We follow a similar procedure to find the sentiment of “on_campus.csv”

Sentiment analysis via region

One thing that could be helpful would be to categorize each university with a region. So each university would then belong to either *North , South , East or West.* After doing this we collect the posts corresponding to these regions from our dataset. Now we can follow the same procedure as the above section to calculate off_campus sentiments and on_campus sentiments based on the region.
The way each university is assigned a region is by manually clustering them together. I have drawn regions corresponding to North , South , West and East regions. If a university falls in a region with the North label then its classified as a North university.

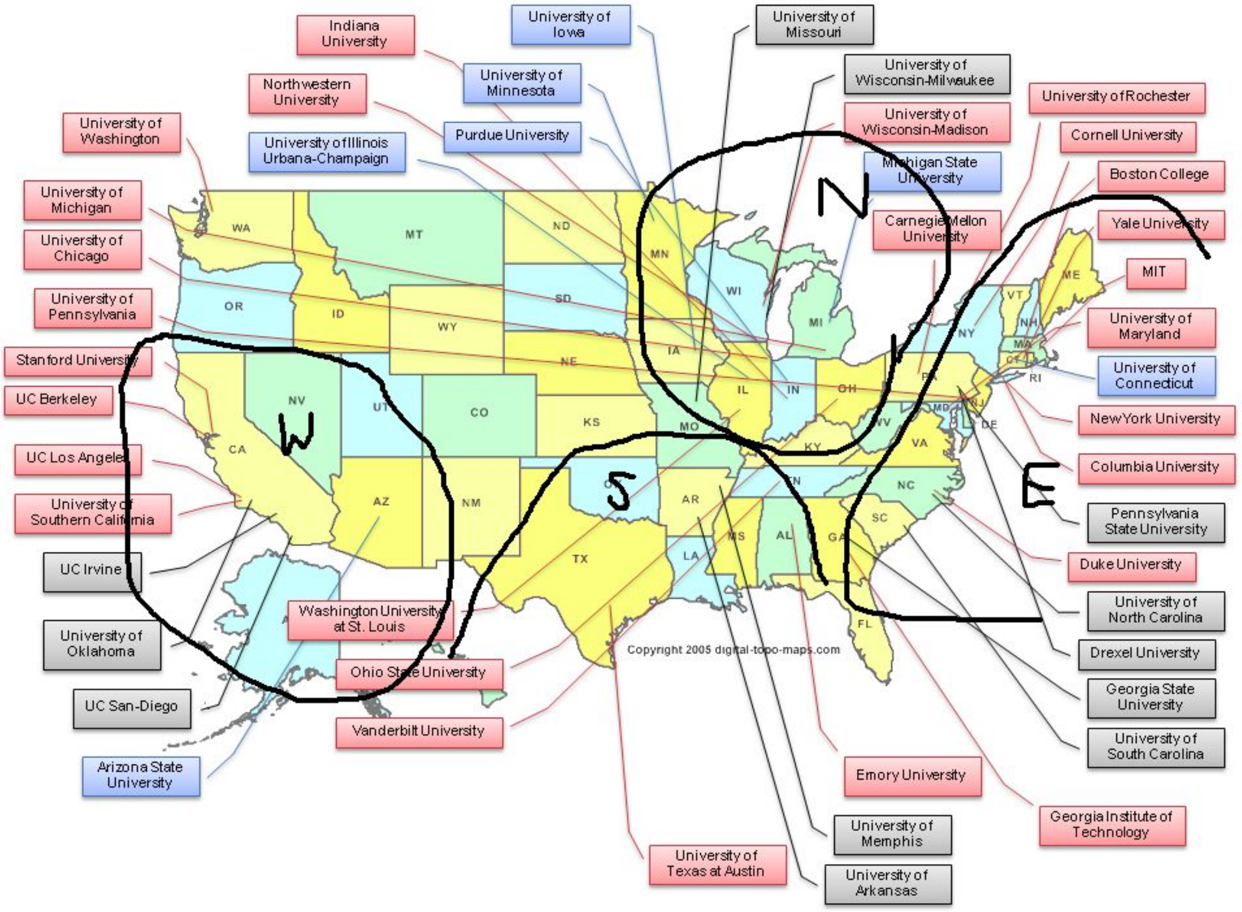


Figure 3 : The map with manually drawn regions under which universities fall *(Source : https://aacademy.files.wordpress.com/2009/09/university_map.png)*

This is how I have divided the universities up for analysis by region:

- University of Illinois at Urbana-Champaign :- N
- University of California - Berkeley :- W
- University of Texas - Austin :- S
- University of Waterloo :- E
- Texas A&M :- S
- Georgia Institute of Technology :- E
- Ohio State University :- N
- Rochester Institute of Technology (RIT) :- E
- University of Toronto :- E
- Virginia Tech :- E
- University of Central Florida :- E
- University of Michigan :- N

- University of California - Los Angeles :- W
- Rutgers University :- E
- Purdue University :- N
- Penn State University :- E
- University of Maryland :- E
- University of British Columbia :- OUTLIER
- University of California - Santa Barbara :- W
- Michigan State University :- N
- University of California - San Diego :- W
- University of Georgia :- E
- Arizona State University :- W
- University of Washington :- OUTLIER
- Florida State University :- E
- University of Pittsburgh :- E
- University of Wisconsin Madison :- N
- Rensselaer Polytechnic Institute :- E
- McGill University :- OUTLIER
- University of California - Davis :- W
- North Carolina State University :- E
- University of Minnesota :- N
- California Polytechnic State University :- W
- University of California - Santa Cruz :- W
- University of Connecticut :- E
- Louisiana State University :- S
- Indiana University :- N
- University of Southern California :- W
- Massachusetts Institute of Technology :- E
- University of Houston :- S
- Northeastern University :- E
- University of Virginia :- E
- University of Florida :- E
- New York University - NYU :- E

Data Pre-Processing for frequent itemset mining

We also would like to see if there are some interesting words/group of words that occur together across a lot of the posts. Doing frequent itemset mining will help to gain insights into the issues that students talk about the most. In order to do this, there will be a variety of preprocessing steps we will need to take. They are all explained below :

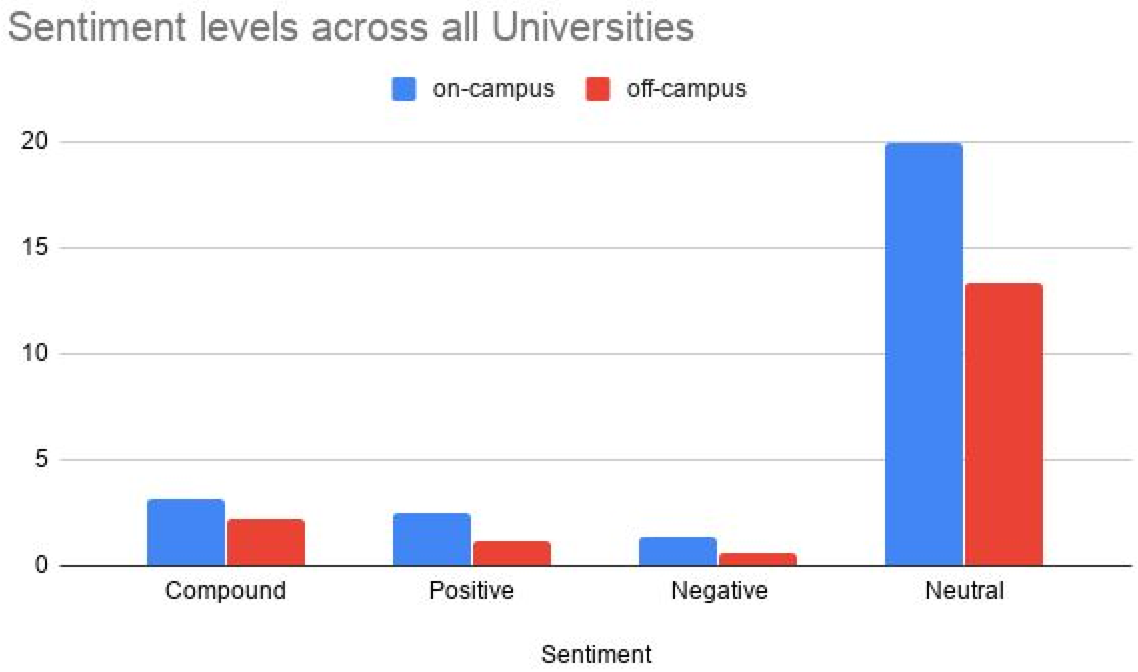
- Load your data into python from the respective csv files.
- Replace NaN by 0.
- Convert every word into lower case as we dont want to make a distinction between Rabbit and rabbit.
- Remove all punctuation , brackets etc. Along with those also remove ‘&#x200B’ and ‘&nbsp’ which appear in the dataset. This was found by manually skimming through the data.
- Next is to import english stop words from the nltk library. Stop words are words like , ‘a’ , ‘an’ , ‘by’ , ‘he’ , ‘she’ etc. Words which don’t add any meaning to our sentence. We do not want to do frequent itemset mining and get these words as the most popularly used words.
- Remove the words ‘off’ and ‘on’ from the list of stop words as these are important words to us.
- Go through the database and keep only words which do not appear in the list of stop words. Now we have a list of words for each post where each word has some sort of meaning.
- Still there could be different variants of the same word which may not give good results if we do a frequent itemset mining. For example , dog and dogs represent the same thing but are different words. To overcome this we will use a process called lemmatization.
- Lemmatization converts words to their root words. So dogs will turn into dog , and the same will happen to all other words. We lemmatize each word using the nltk library. It has a Lemmatization function inbuilt into it.

Now our data is ready for a frequent itemset mining algorithm. We will use the algorithm FP-Growth , to implement this algorithm we will be using the external python library called *pyfpgrowth.*

RESULTS AND EVALUATION

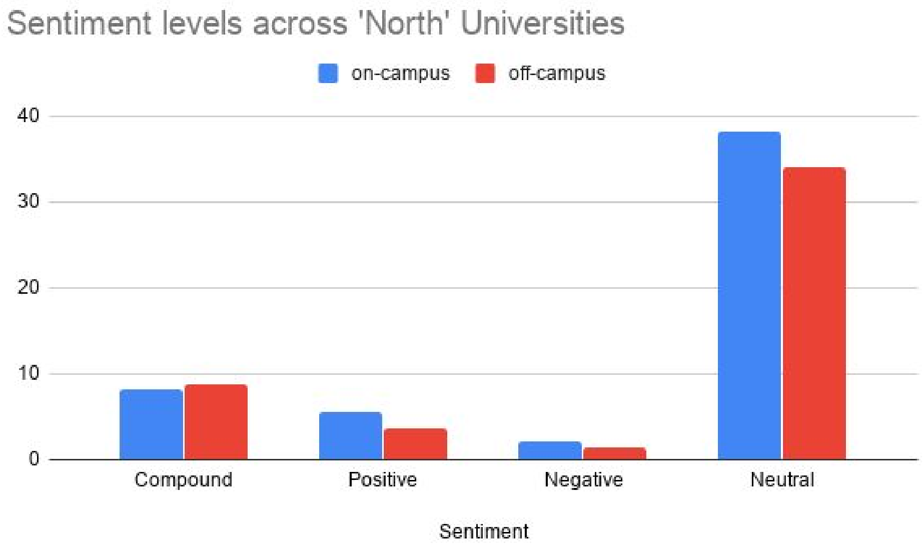
Results from Sentiment Analysis

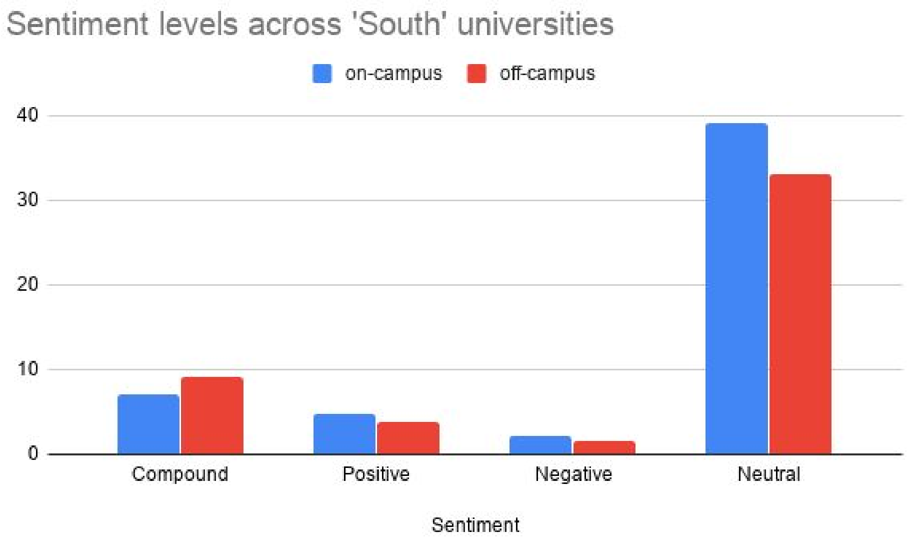
Sentiment Analysis across the entire dataset

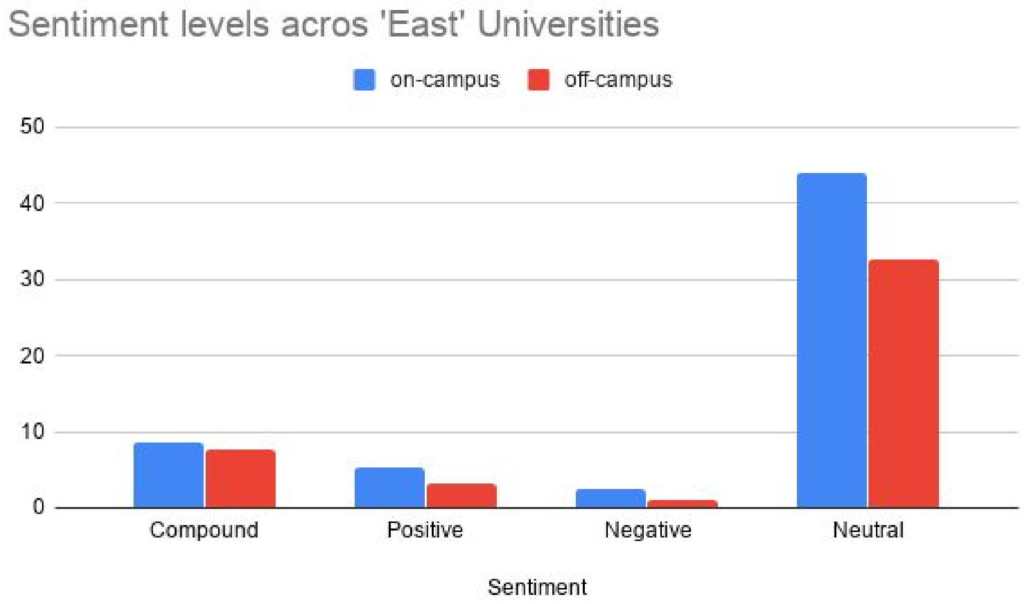


These results come as an average of sentiment analysis across all posts gathered. Compound is the total sentiment. We can see that overall people tend to prefer on-campus housing rather than off-campus housing. The interesting thing in this graph is that every single value for on-campus is higher than off-campus. So even if there is more positive sentiment overall in on-campus housing , there is also an overall more negative sentiment. But all in all on-campus has a higher compound value, so events happening on-campus make people who live there feel better overall.

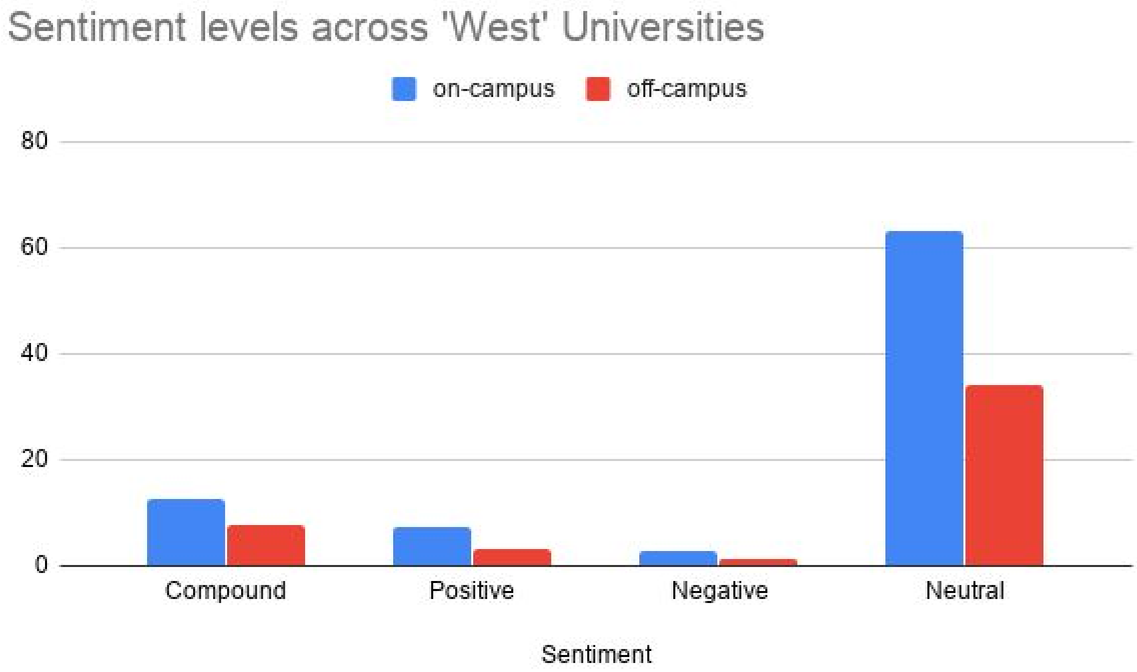
Sentiment Analysis across universities by region







Sentiment levels acros 'East' Universities

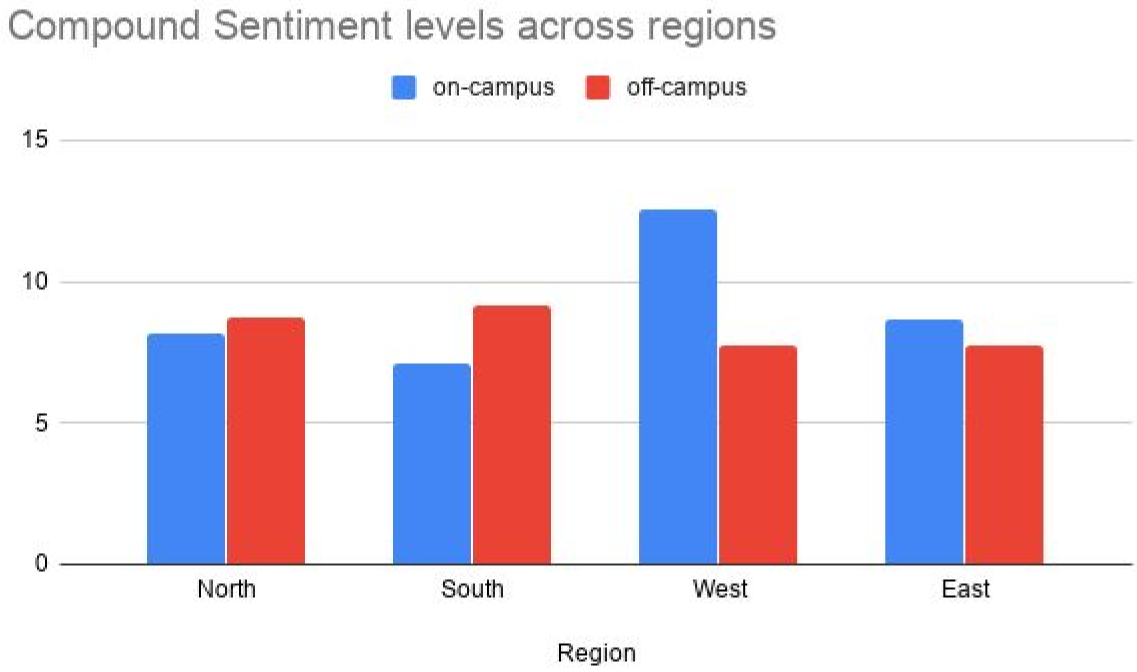


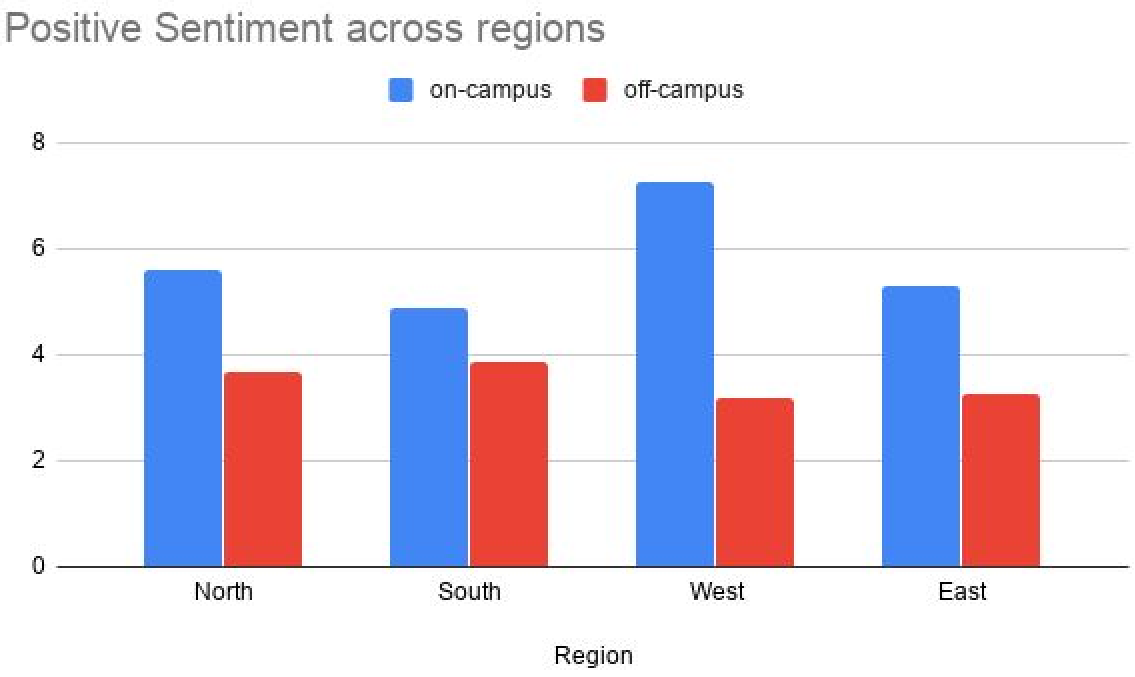
Sentiment levels across 'West' Universities

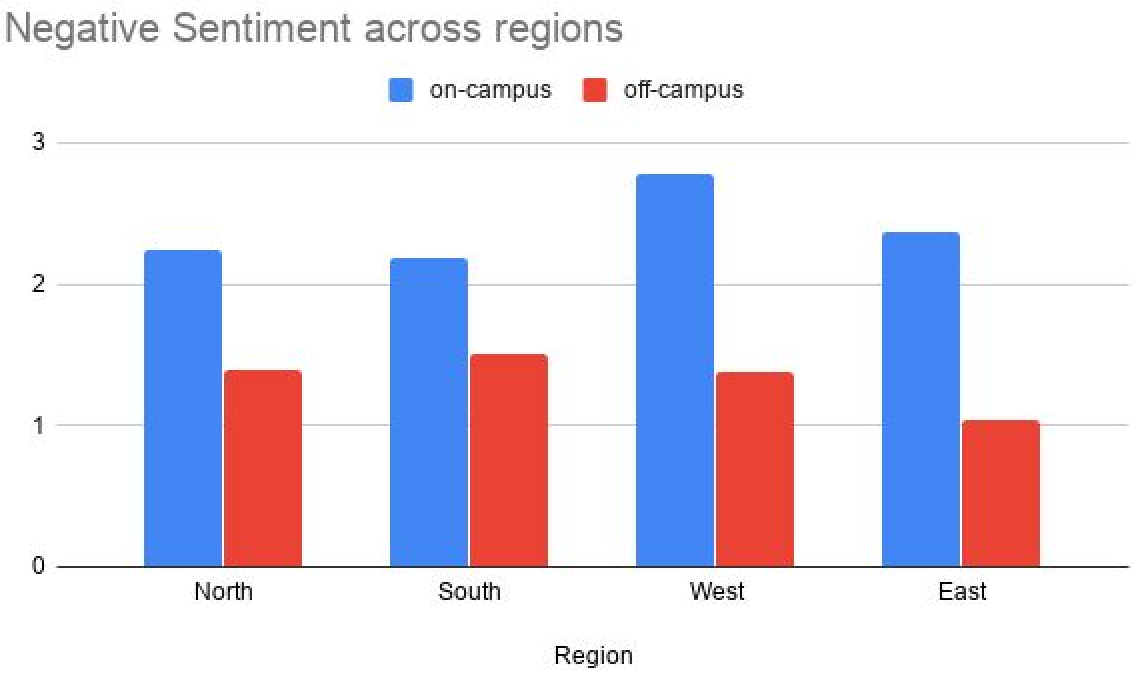
We see a different result for North and South universities here. Here the overall sentiment is higher off-campus than on-campus. That means that a student would be happier if he/she chose an off-campus housing here. The thing to note is, the positive sentiment is still higher for on campus as compared to off campus. It is just that there is much lower negative sentiment off campus that it affects the total sentiment making off campus a better choice.

A comparison between sentiment level in North , South , West and East regions is given below. It will help us decide which region is better than other regions.

Sentiment Analysis among different regions




Positive Sentiment across regions
on-campus
off-campus
8
6
4
2
0
North
South
West
East
Region


Negative Sentiment across regions
on-campus
off-campus
3
2
1
0
North
South
West
East
Region

The above graphs shows us rankings as to which region should a student study in.
If he/she opts for on-campus housing :

- West Region
- North Region
- East Region
- South Region

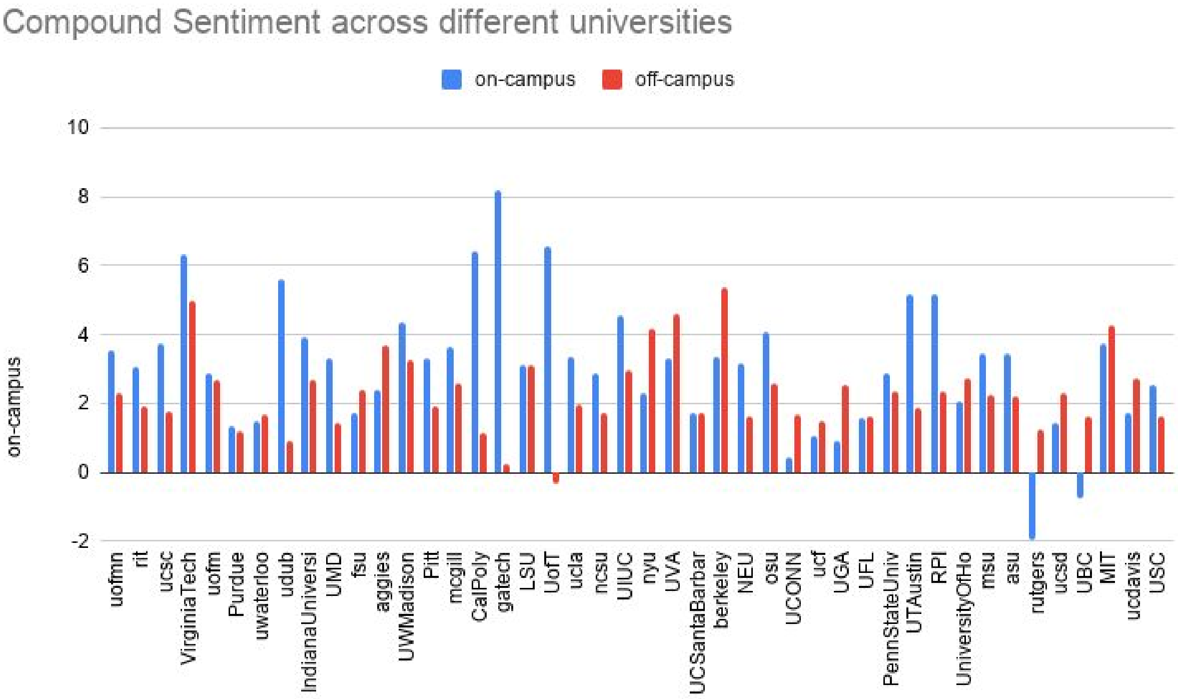
We can see for on- campus housing west region is most preferred. It makes sense too as this area mostly comprises of the California state and it is known to be expensive. So off-campus housing would be too expensive putting on pressure on students and increasing their negative sentiment. Also colleges here are big name colleges , UCLA , Stanford etc. where students would love spending time on campus.

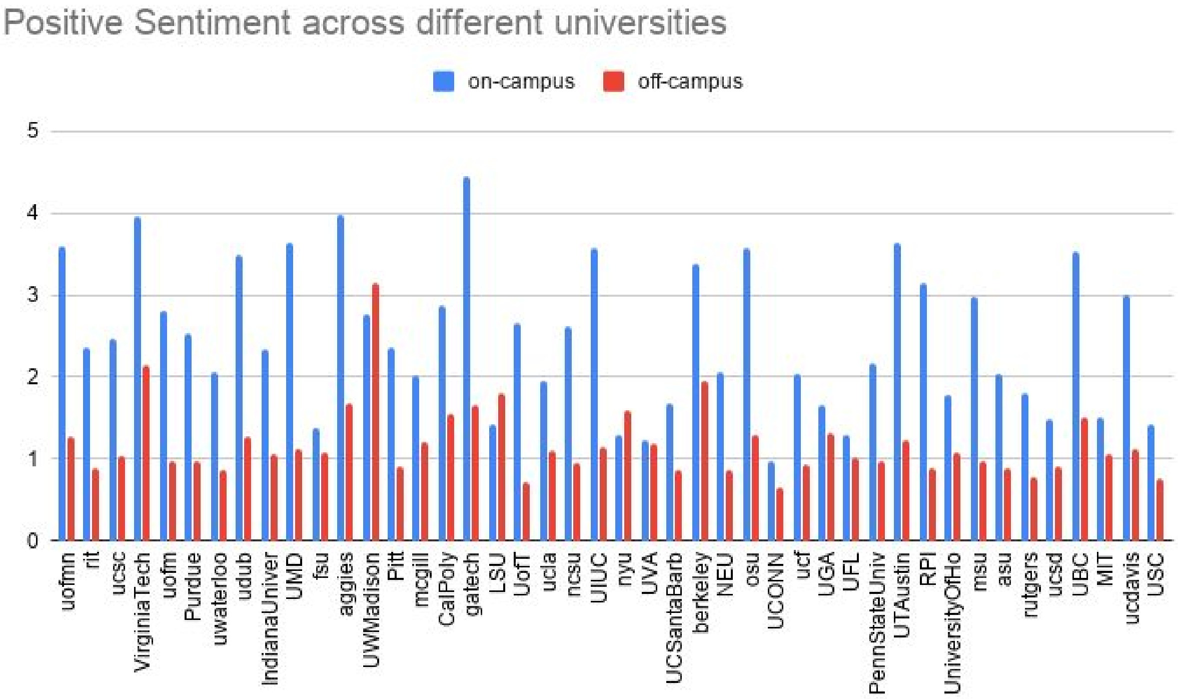
If he/she opts for off-campus housing :

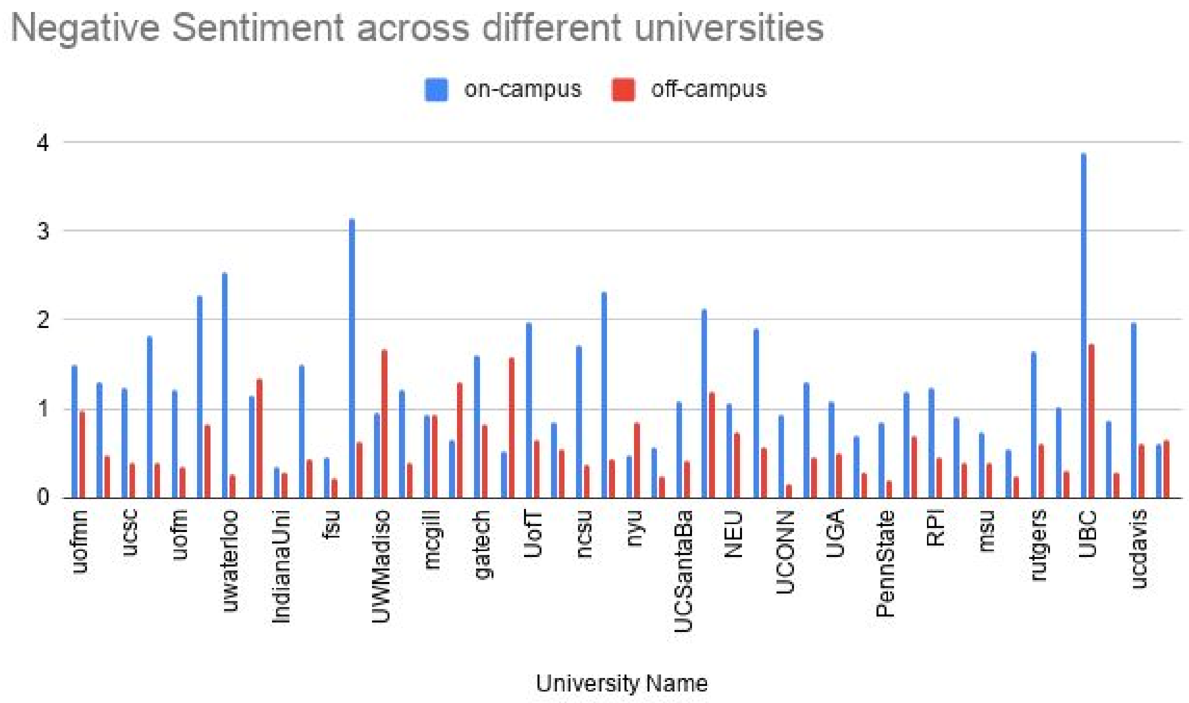
- South Region
- North Region
- East Region
- West Region

If the student chooses off-campus housing then he/she will be happier in the South region. And as discussed above the place where the students will be the least happy would be the west region , most probably because of the high prices in the city.

Sentiment Analysis across Universities

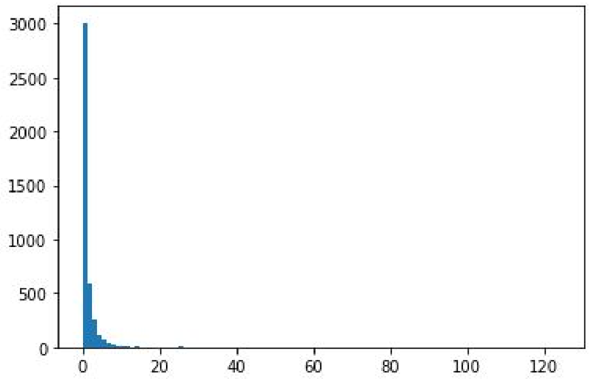




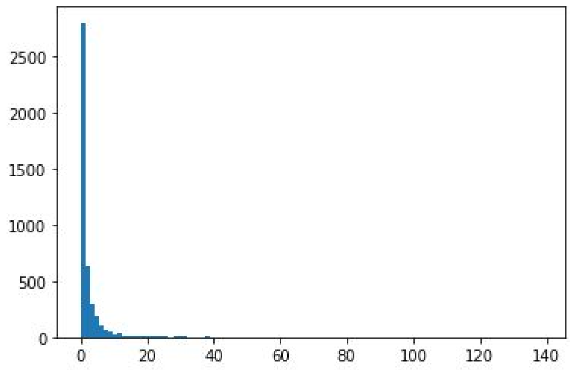


The above graphs give us a wealth of information and allow us to draw so many conclusions. These help us compare happiness levels between on-campus/off-campus among universities.
For example from the above I can say , a student will be happier taking off-campus housing at NYU than on-campus at Purdue.
We note the highest positive sentiment for on campus housing is by Gatech and for off campus it's by UWMadison.
We can make a lot of these types of conclusions , an interesting thing that I find is, the graph tells us that overall a student will be happier off campus in NYU than compared to on campus. And this makes sense as well, NYC is one of the best cities in the country and the student would prefer living off campus here rather than on campus.

If we look at histograms of these sentiment values we notice something. There aren't many happy/sad events i.e there are a lot of posts that give low sentiment values (sentiment is not happiness, a low sentiment does not mean unhappy. A low positive sentiment score just means the post lacks happiness and not that it suggests sadness) In other words, things are pretty much neutral and there are some events that have higher sentiment values.



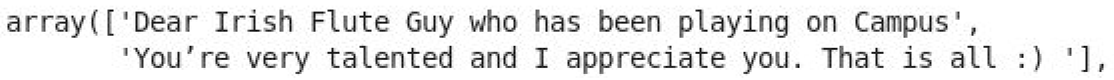
Offcampus , positive sentiment values Histogram



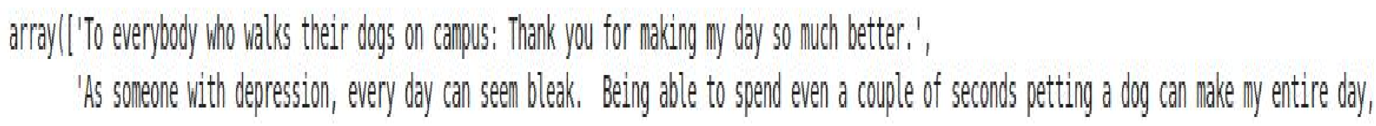
Oncampus , positive sentiment values HIstogram

As we can see from the above graphs that the highest positive sentiment value was from a post from off-campus section but we see that there are a lot of medium values for oncampus which in effect make the overall positive sentiment value of the on campus dataset greater. This also means there are more things that happen frequently on campus by which students get happy whereas off campus events are sparse but when they do happen they make a big splash (perhaps major festivals like Christmas light up the neighbourhood resulting in a huge increase in the happiness levels.)

Some examples of posts with the highest positive sentiment values:

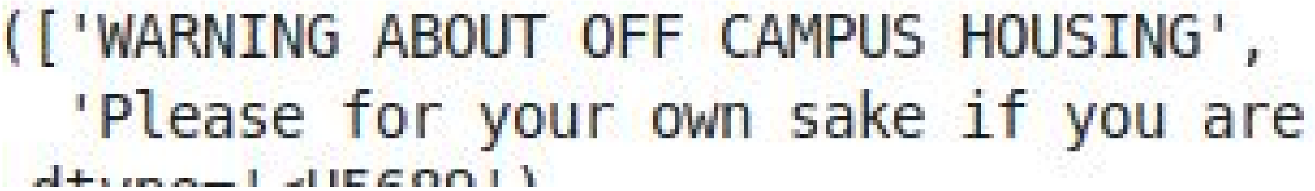


```
array(['Dear Irish Flute Guy who has been playing on Campus',
       'You’re very talented and I appreciate you. That is all :) '],
```

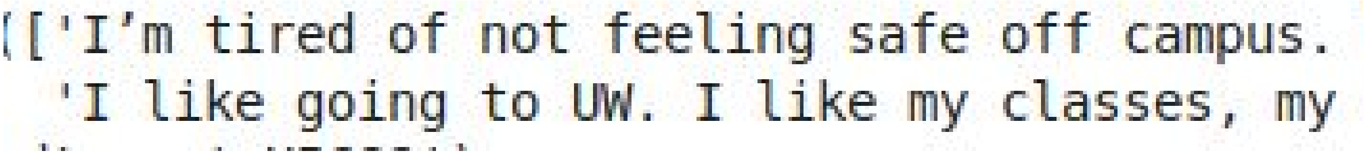


```
array(['To everybody who walks their dogs on campus: Thank you for making my day so much better.',
       'As someone with depression, every day can seem bleak.  Being able to spend even a couple of seconds petting a dog can make my entire day,
```

Some examples with the highest negative sentiment values:



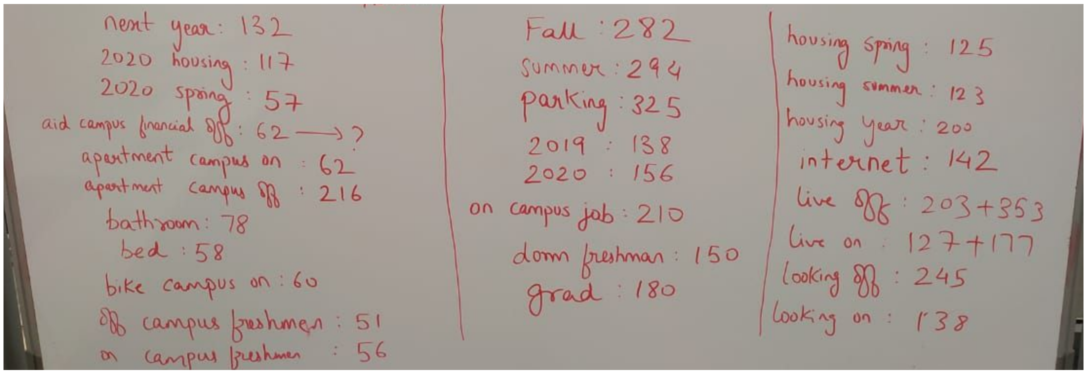
```
(['WARNING ABOUT OFF CAMPUS HOUSING',
  'Please for your own sake if you are
```



```
(['I’m tired of not feeling safe off campus.
  'I like going to UW. I like my classes, my
```

Results from Frequent Itemset Mining using FPGrowth:

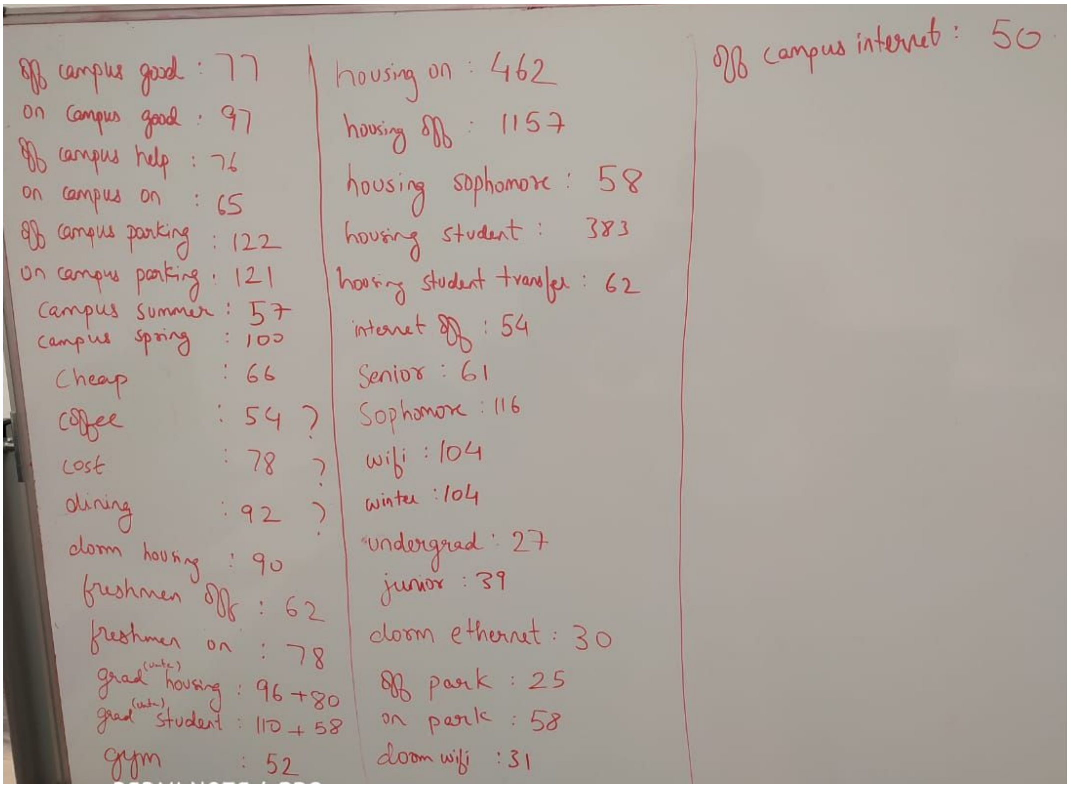
I ran the FP-Growth algorithm multiple times with different thresholds and wrote down results which seemed interesting. This experiment did not give insights into problems faced by students which I thought it would before carrying this out. But it did give some other insights.



next year: 132
2020 housing: 117
2020 spring: 57
aid campus financial off: 62 → ?
apartment campus on: 62
apartment campus off: 216
bathroom: 78
bed: 58
bike campus on: 60
off campus freshmen: 51
on campus freshmen: 56

Fall: 282
Summer: 294
parking: 325
2019: 138
2020: 156
on campus job: 210
dorm freshman: 150
grad: 180

housing spring: 125
housing summer: 123
housing year: 200
internet: 142
live off: 203+353
live on: 127+177
looking off: 245
looking on: 138



off campus good : 77
on Campus good : 97
off campus help : 76
on campus on : 65
off campus parking : 122
on campus parking : 121
Campus Summer : 57
campus spring : 100
Cheap : 66
coffee : 54 ?
cost : 78 ?
dining : 92 ?
dorm housing : 90
freshmen off : 62
freshmen on : 78
grad (unte) housing : 96 + 80
grad (unte) Student : 110 + 58
gym : 52

housing on : 462
housing off : 1157
housing sophomore : 58
housing student : 383
housing student transfer : 62
internet off : 54
Senior : 61
Sophomore : 116
wifi : 104
winter : 104
undergrad : 27
junior : 39
dorm ethernet : 30
off park : 25
on park : 58
doom wifi : 31

off campus internet : 50

These were some interesting results got from frequent itemset mining , but most of the patterns which were frequent didn't give me much information. One thing I learned from doing this is that students really care about 'parking' in deciding where to live. The word 'parking' , 'parking off campus' , 'parking on campus' appeared frequently.

Other information this gives us is about the number of sophomores , freshmen , graduates , seniors , juniors talking about housing.

The string 'apartment off campus' appeared much more than 'apartment on campus' but that is obvious as people need to ask about locations and ratings and other stuff for off campus apartments , for on campus its standard and is known.

People also wrote about 'bike campus' frequently , that does suggest a lot of people do bike and they would like to be as near the campus as possible.

With the word ‘dorm’ the most frequent things were ‘dorm wifi’ and ‘dorm ethernet’. So if a student does decide on going for a dorm , one of the most important things to the student is the internet connection there.

Conclusions and Future Work

The conclusions we can draw are :

- Overall on campus housing is happier than off campus housing.
- If we look according to regions the west region is happiest for on-campus housing.
- If a student wants to go for off-campus housing then the south region is the happiest.
- There are some universities which have higher levels of happiness off campus than on campus. (eg. NYU)
- There are not many happy/sad events occuring in day to day life , the happiest event occur off campus but very sparsely , on campus events which are happy in nature occur often but the value of the sentiment is not as high as the happiest of the off campus events.
- From the graph showing comparisons between universities one can draw other variety of conclusions by comparing and contrasting.
- Parking is important to students in terms of housing.
- If decided on a dorm then the most important thing is the Wifi and the Ethernet services there.

Future Work : -

- The frequent set item mining was done using only the titles of the posts as performing it on the body of the post was time intensive. It would give us better results if we perform it on the body as well.
- Posts with a question type nature can be isolated, of the type , “Which is better , off campus or on campus?” and the top comment on the post can be analysed to know about which is better and why exactly.
- This time I performed clustering manually, but we can run sentiment analysis over posts from different universities and then use clustering over the result to find clusters which naturally arise due to sentiment values and not only by geographic location.

ACKNOWLEDGEMENTS

My earnest thanks to Prof. Jeibo Luo for the support extended to me in numerous ways while doing this study.