Jin Li

5500 South University Ave. Chicago, IL 60637

646 472 6952

jinli7255@gmail.com

Economics and Statistics Class of 2022

[10/3/2018]

Tasks:

Decipher API given to me by Christina’s husband. Essentially read through the code to understand how it works.

Write my own API code. Will do this by creating a tesseract search. More documentation is on: <https://github.com/JinLi711/FAUI-Data-Scraping/tree/master/Data%20From%20Yelp>

Email With TNC on Starting Up:

*Welcome to the Scenes Project! Here are two main memos to start with. Most urgent is to read how to do the LOG file and start for yourself, email the LOG to Terry Clark as attachment regularly.*

*Download these 2 files - you don’t need to have a Dropbox account to download, just any browser like Safari or IExplorer. The two filenames are*

*scenes administrative memos 8*

*Note to new staff on Scenes project 4 doc files*

*These are what you can search for on one drive.*

*if you cannot SEARCH, you can scroll to the bottom:*

*here they are at the bottom of the main root folder: 2 .doc files.*

*You Tube Scenes staff meeting April 25, 2018:*

*<https://www.youtube.com/watch?v=xEOsRpyR1ck>*

*SCENESCAPES BOOK:*

*<https://www.dropbox.com/s/wz0j4rcct6cl378/ScenescapesFreeDownload.pdf?dl=0>*

*OLD, SUPERCEDED:*

*New staff download two files. Here are their filenames.*

[*https://www.dropbox.com/s/8l2tambhwsc69gp/Scenes.administrative%20memo8.docx?dl=0*](https://www.dropbox.com/s/8l2tambhwsc69gp/Scenes.administrative%20memo8.docx?dl=0)

[*https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0*](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)

Email With TNC on Meeting:

*Normally every Wed as it says in the doc. 4:30 in SS but I wrote I think I will email you M or Tu. each week.*

*On Sep 30, 2018, at 5:08 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Are there any upcoming meetings?*

*On Sat, Sep 29, 2018 at 10:36 PM Jin Li <jinli7255@gmail.com> wrote:*

*jinli11@uchicago.edu*

*On Sat, Sep 29, 2018 at 9:51 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:*

*Actually can you access the one drive I will send you a link if you use your university of Chicago email then you will be able to search and if you go in through office 365 it may last more than 30 days otherwise it ends in 30 days..*

*Then if you go to the research assistance large files Christopher Graziul has many files including logs there neatly organized*

*Sent from my iPhone*

*Terry Nichols Clark*

*Professor of Sociology*

*University of Chicago*

*tnclark@uchicago.edu*

*On Sep 29, 2018, at 8:21 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Would a log look like this? Could you provide feedback?*

*I also skimmed the video (should I watch the entire thing?)*

*On Sat, Sep 29, 2018 at 6:46 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:*

*Welcome to the Scenes Project! Here are two main memos to start with. Most urgent is to read how to do the LOG file and start for yourself, email the LOG to Terry Clark as attachment regularly.*

*Download these 2 files - you don’t need to have a Dropbox account to download, just any browser like Safari or IExplorer. The two filenames are*

*scenes administrative memos 8*

*Note to new staff on Scenes project 4 doc files*

*These are what you can search for on one drive.*

*if you cannot SEARCH, you can scroll to the bottom:*

*here they are at the bottom of the main root folder: 2 .doc files.*

*You Tube Scenes staff meeting April 25, 2018:*

*https://www.youtube.com/watch?v=xEOsRpyR1ck*

*SCENESCAPES BOOK:*

*https://www.dropbox.com/s/wz0j4rcct6cl378/ScenescapesFreeDownload.pdf?dl=0*

*OLD, SUPERCEDED:*

*New staff download two files. Here are their filenames.*

*https://www.dropbox.com/s/8l2tambhwsc69gp/Scenes.administrative%20memo8.docx?dl=0*

https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0

<JinLi Log01.docx>

W*e were talking about sending my background information to Ben.*

*My Background in Python:*

*Did a 5 month course in high school (sophomore year) learning Python.*

*Wrote some small scripts for real estate company (can't share my files here, but I wrote scripts that would find and delete duplicate files, empty folders. I wrote another script that took information from an excel sheet, and displayed it on a browser. On the (internal) website, a person would search for the building name and type of information for that building (example, address), and the website would return the appropriate information.*

*Wrote 2 machine learning programs:*

*https://github.com/JinLi711/Simple-RNN-for-Stock-Prediction*

*https://github.com/JinLi711/Predicting-House-Hold-Poverty-Levels-in-Costa-Rica*

*There's documentation in there, but I basically had practice manipulating dataframes, using Python's tensorflow and scikit-learn.*

*Current Courses I'm taking:*

*Honors Analysis in R^n (MATH 20700) (I already have a background in multi-variable, linear algebra, analysis)*

*Honors Computer Science (CMSC 16100) (learning haskell)*

*Reading Cultures (HUMA 14000)*

*Self, Culture, and Society (SOSC 12400), but probably going to switch in biology*

*Below are two research papers I wrote (I worked with a professor from NYC in the Game Theory one. The other was a 3 month paper from an AP Statistics course)*

*If there's anything to add, ask me.*

*Thanks!*

*Jin Li*

*On Sun, Sep 30, 2018 at 7:57 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:*

*sorry I had to go to dinner but we can talk more now if you have decisions that*

*are salient now?*

*I dont see your phone handy.*

*tnc*

*On Sep 30, 2018, at 6:50 PM, tnclark.mbp2 <tnclark@uchicago.edu> wrote:*

*we should talk about your more general courses. can talk or skype now if you like*

*312 842 5169*

*tnc*

*On Sep 30, 2018, at 5:21 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Also, do you recommend if I take a statistic course for my first quarter? (I'm allowed to take STAT 24400)*

*On Sun, Sep 30, 2018 at 5:08 PM Jin Li <jinli7255@gmail.com> wrote:*

*Are there any upcoming meetings?*

*On Sat, Sep 29, 2018 at 10:36 PM Jin Li <jinli7255@gmail.com> wrote:*

*jinli11@uchicago.edu*

*On Sat, Sep 29, 2018 at 9:51 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:*

*Actually can you access the one drive I will send you a link if you use your university of Chicago email then you will be able to search and if you go in through office 365 it may last more than 30 days otherwise it ends in 30 days..*

*Then if you go to the research assistance large files Christopher Graziul has many files including logs there neatly organized*

*Sent from my iPhone*

*Terry Nichols Clark*

*Professor of Sociology*

*University of Chicago*

*tnclark@uchicago.edu*

*On Sep 29, 2018, at 8:21 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Would a log look like this? Could you provide feedback?*

*I also skimmed the video (should I watch the entire thing?)*

*On Sat, Sep 29, 2018 at 6:46 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:*

*Welcome to the Scenes Project! Here are two main memos to start with. Most urgent is to read how to do the LOG file and start for yourself, email the LOG to Terry Clark as attachment regularly.*

*Download these 2 files - you don’t need to have a Dropbox account to download, just any browser like Safari or IExplorer. The two filenames are*

*scenes administrative memos 8*

*Note to new staff on Scenes project 4 doc files*

*These are what you can search for on one drive.*

*if you cannot SEARCH, you can scroll to the bottom:*

*here they are at the bottom of the main root folder: 2 .doc files.*

*You Tube Scenes staff meeting April 25, 2018:*

*https://www.youtube.com/watch?v=xEOsRpyR1ck*

*SCENESCAPES BOOK:*

*https://www.dropbox.com/s/wz0j4rcct6cl378/ScenescapesFreeDownload.pdf?dl=0*

*OLD, SUPERCEDED:*

*New staff download two files. Here are their filenames.*

*https://www.dropbox.com/s/8l2tambhwsc69gp/Scenes.administrative%20memo8.docx?dl=0*

*https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0*

*<JinLi Log01.docx>*

Email with Ben on Web Scraping

*Edit\**

*Wrong book link: should be this one:*

*https://www.amazon.com/Web-Scraping-Python-Collecting-Modern/dp/1491985577/ref=sr\_1\_2?ie=UTF8&qid=1538937516&sr=8-2&keywords=web+scraping+with+python*

*On Sat, Oct 6, 2018 at 6:21 PM Jin Li <jinli7255@gmail.com> wrote:*

*Easiest solution: Google Maps API*

*https://cloud.google.com/maps-platform/*

*But it costs 200 dollars a month.*

*I can web scrape, but I'm an amateur at it (I know how to do basic web scraping (like beautifulsoup, selenium), but I don't have have the experience to troubleshoot possible problems that may be very time-consuming to solve. For example, some websites will block users if they web scrape, and I know it's possible to avoid being blocked, but I just don't know how.)*

*But if web-scraping is something that I'm going to be doing for the long run, I'm willing to learn:*

*https://www.amazon.com/Mining-Social-Web-Facebook-LinkedIn/dp/1449367615/ref=sr\_1\_9?ie=UTF8&qid=1538864923&sr=8-9&keywords=web+scraping+with+python*

*I'll buy that book and teach myself.*

*For web scraping methodology:*

*When searching in Google Maps, let's say Cafe, it actually does turn out that in the google result, Starbucks show up (along with other cafes that don't have cafe in their names.)*

*When collecting the data from the web, I don't need to search "cafe" to extract, I can just extract all possible results. The problem with that is there's huge bias (Google shows the top cafes first, so the bias will be based on Google's algorithms.) But I think Google Maps will eventually exhaust all possible options. Also, I can search google by location. By default, Google searches results partly based on my current location, but I can change my location, and search from the new location.*

*Also, here's another database: https://www.uscompanieslist.com/. Costs $200, but it has 28 Million lists of US Companies, searchable by zip code. Also, the search contains links to the companies website. Also, I could contact the website and they might give some data for free if we say it's for academic purposes.*

*On Fri, Oct 5, 2018 at 2:15 AM Ben Picker <benpicker@gmail.com> wrote:*

*As a starting point, you should look into Hyesun's idea of zip codes. Perhaps people webscrape by identifying the zipcode of the business on the website. Maybe then then check this against other businesses in the area. I don't know. But this would be a good starting point if you don't have any initial ideas.*

*Ben*

*On Fri, Oct 5, 2018 at 1:57 AM Ben Picker <benpicker@gmail.com> wrote:*

*Jin,*

*I would additionally like you to look into web scraping methods. But I am interested in a very specific issue.*

*Suppose I only scrape websites with the word "cafe" in it, Starbucks might be excluded because it doesn't have cafe in its name. But the density of Starbucks obviously varies by region, for instance urban areas might likely a higher density. So if Starbucks isn't in the data set, then that geographical variation isn't either, which leads to bias. If we try to draw conclusions about patterns of cafes across the US or a city, that geographical variation in density won't be present.*

*These are the sorts of issues that are easily solved with the API because it's not done via keywords but instead by geospatial coordinates. But for scraping, we don't have that option. So what do we do?*

*This leads to the critical question*

*Can we systematically scrape data from businesses so that, within any given geographical location I pick, I can be sure I've gotten all the businesses in an area? If not, how do researchers scrape so as to avoid creating biased data sets?*

*I want you to look into these methods.*

*Again, let me know if you have questions.*

*Ben*

Email with TNC on My Role

*[can you and Ben bend what you discussed as a To Do for your in as much detail as possible, then ask for more input where you need in that context?](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[My main point is to help get results comfortably for you and not crate hard tasks. we need your help in advising what is hard or easy.](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[tnc](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[On Oct 7, 2018, at 10:02 PM, Jin Li <jinli7255@gmail.com> wrote:](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[Dear Dr. Clark,](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[I'm going to be web scraping data and using APIs.](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[Is there a particular place you want me to explore? For example, data from Facebook, Yelp, Instagram, etc.](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[Is there any particular type of data that you want? For example, data in the Chicago area, data on individual people vs. data on events.](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[Also, you said you would send me a link to access one drive.](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[Email: jinli11@uchicago.edu](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[From,](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

*[Jin Li](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)*

Problems:

API is unpredictable and very limiting.

API is only willing to do a maximum of 5000 requests a day, but the number is sketchy. For example, making one requests sometimes decreases the request count by 10.

The concentric tesseract search took way too long, produced way to many repetitions, and failed to get the intended results.

Solutions:

Create my own web scraper.

Tasks Completed:

<https://github.com/JinLi711/FAUI-Data-Scraping>

[Note that all Documentation is in the link.](https://www.dropbox.com/s/pl7qpj8sknwkey6/NOTE%20to%20new%20staff%20on%20Scenes%20Project4.docx?dl=0)

[10/17/2018]

Objectives:

Learn how to avoid traps that block me.

Learn parallel scraping

Learn how to use Cloud computing.

Measure time it takes to web scrape.

Correspondence to Hyesun about Collecting Data:

*It's my first year, so haven't decided yet. Probably math and economics.*

*On Sat, Oct 13, 2018 at 11:42 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Ok, thanks!!*

*What is your major?*

*Hyesun Jeong, PhD, AIA Associate*

*Postdoctoral Research Fellow*

*Urban Innovation Analysis, Inc.*

*https://sites.google.com/view/hyesunjeong*

*On Oct 13, 2018, at 11:36 PM, Jin Li <jinli7255@gmail.com> wrote:*

*I'm new here, so I haven't seen much of the available data. I'll check it out though.*

*Python's pandas is basically R (manipulations of numbers, strings; nothing related to GIS).*

*I'll get to more web scraping after my ethernet cable arrives.*

*On Sat, Oct 13, 2018 at 11:21 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*I think it matters in case of making variable that count the number of cafes per zip code.*

*Have you seen Scenes merge data? Most variables are at the zip code level (and also county, city, etc)*

*If we only want to analyze the reviews, then it wouldn’t matter how many cafes in each zip code. However, I believe we will integrate this analysis with Scenes dimension score and other zip code level data in national level eventually.*

*For a short term, we can confine this to Chicago, but we need to have correct number of cafes per zip code for later national study.*

*The last thing you mentioned - Python’s pandas, I don’t know what it looks like. Can you try it and see how it comes out differently?*

*Thanks much!*

*Hyesun Jeong, PhD, AIA Associate*

*Postdoctoral Research Fellow*

*Urban Innovation Analysis, Inc.*

*https://sites.google.com/view/hyesunjeong*

*On Oct 13, 2018, at 11:11 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Hey Hyesun,*

*Just wondering, if data is gathered for a national scale, shouldn't it not matter if a search overlaps zip codes?*

*Another thing that I could do is scrape directly the Zip Code and separate it through Python's pandas, which would be very fast to do.*

*On Sat, Oct 13, 2018 at 10:58 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Hi Jin,*

*I can also trim it with GIS. If it’s for one zipcode, or one city like Chicago, it is not a problem.*

*But, if we want to replicate it for national scale which is about 40,000 zip codes, it may be troublesome. GIS consumes lots of memory, and it will freeze, and it will take a long time.*

*The nature of Yelp is basically search for restaurants/cafes/others “nearby” but not “within” certain place. And that is fundamental issue for counting the number of cafes for each zipcode, city, etc.*

*So, the general thing is..until cutting out the yelp result from GIS boundary, we cannot publicly report the number of cafes for each zipcode… right?*

*What do you think?*

*Hyesun Jeong, PhD, AIA Associate*

*Postdoctoral Research Fellow*

*Urban Innovation Analysis, Inc.*

*https://sites.google.com/view/hyesunjeong*

*On Oct 13, 2018, at 10:43 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Sorry about that. I didn't check.*

*Ben knows something about GIS (Geographical information System), which is a software that trims data based on grid locations. For example, it can trim based on Zip Code 60637, and remove all data points not in 60637.*

*I'll try to keep in mind more possible biases in the future.*

*On Sat, Oct 13, 2018 at 10:37 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Hi Jin,*

*I don’t think there are 60 cafes in zip code 60637. When you think of cafes in Hyde Park area by intuition, it cannot be 60..*

*I’ve re-checked the Yelp page you sent for 60637, and found that there are mix of cafes in 60637 and other zip codes.*

*Is there way we can correct this??? See below.*

*As you see in the screen shot, #21 Carver 47 cafe and #22 Ain’t she sweet cafe are not in 60637. They are 60653.*

*Also, the other screen shot below shows #2 Cafe 53 is in zipcode 60615, not 60637.*

*Hyesun Jeong, PhD, AIA Associate*

*Postdoctoral Research Fellow*

*Urban Innovation Analysis, Inc.*

*https://sites.google.com/view/hyesunjeong*

*On Oct 13, 2018, at 10:11 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Hey Hyesun,*

*TLDR: just see bold.*

*1.*

*I counted 20 seconds for one business. I pretty sure it would be much faster if I had an ethernet cable (and even faster if I could somehow access the campus computers). In Zip Code 60637, there are about 60 cafes, so about 20 minutes. In a city like Chicago, there are 2500 cafes, (so 14 hours!). In the US, there are 32,000 cafes (7.5 days).*

*But the problem with web scraping is that if I scrape continuously for 14 hours straight, servers at Yelp would probably block me (and possibly prevent me from ever accessing Yelp again). But there are ways around that (I heard of a guy who web scraped 32 million users in Facebook before Facebook ever found out, and Facebook is significantly more tight about web scraping). The trick is to bounce IP addresses to prevent servers at Yelp from ever knowing all the requests are coming from my labtop. This would slow down the process but would be safer. I currently do not know how to bounce IP addresses, but I'm learning.*

*2.*

*https://github.com/JinLi711/Simple-RNN-for-Stock-Prediction*

*https://github.com/JinLi711/Predicting-House-Hold-Poverty-Levels-in-Costa-Rica*

*To sum up what's in the links, I've done RNN on stock predictions, and I know how to apply ML algorithms like SVM to do sentimental analysis. I also have a book on neural networks that basically give the code for creating a sentimental analysis.*

*For sentimental analysis, I would just need text (like a review comment) and whatever information you want from it (ex. how satisfied people are with the business, or how educated people seem based on what they write). But depending on the information you want, I would have to label the data in such a way. For example, if you want to measure how educated people seem based on what they write, I would have to rate a few reviews myself (based on some criteria) on how educated the reviewer seems to me. This will allow an algorithm that I write to have a basis for learning.*

*3.*

*The algorithm I wrote excludes the ad. When parsing the HTML, the tags for ads are different from tags for businesses.*

*From,*

*Jin Li*

*On Sat, Oct 13, 2018 at 9:10 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Hi Jin*

*Thanks for update.*

*It looks really good as far as I see the sample output which has information we need for cafe. My question:*

*1. How much time did it take to do this sample output, and how long do you estimate it will take to do one zipcode, city, and whole US? Just want to frame out timeline based on that.*

*2. Can you share me some past example s of sentimental analysis and RNN you did? In order for you to replicate analysis, what information do you need for cafes?*

*3. Is it possible to exclude businesses popped in Advertisement (Ad) section when you scrape cafes? When I click the Yelp link you sent, there was Sushi bar in the Ad section.*

*On Oct 13, 2018, at 8:47 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Just to clarify, I don't actually have web scraped data of all cafes in Yelp.*

*But, I currently am able to find all links to cafes in Yelp, given a starting website (like:*

*https://www.yelp.com/search?find\_desc=cafe&find\_loc=Chicago,+IL+60637)*

*The code that I wrote allows me to gather all links in this page, all links in the second page, and so on.*

*I was also able to scrape for data on one business page. For example:*

*https://www.yelp.com/biz/greenline-coffee-chicago*

*I can gather the basic data:*

*{'Accepts Apple Pay': 'No',*

*'Accepts Credit Cards': 'Yes',*

*'Accepts Google Pay': 'No',*

*'Address': '501 E 61st StChicago, IL 60637',*

*'Bike Parking': 'Yes',*

*'Caters': 'No',*

*'Gender Neutral Restrooms': 'Yes',*

*'Good for Working': 'Yes',*

*'Health Score': '95 out of 100',*

*'Outdoor Seating': 'Yes',*

*'Parking': 'Street, Private Lot',*

*'Price Range': 'Inexpensive',*

*'Take-out': 'Yes',*

*'Wheelchair Accessible': 'Yes',*

*'Wi-Fi': 'Free'}*

*And I can gather all the reviews of that page (see csv).*

*This is just a starter (I can gather any data you want on Yelp, as long as you can see it on the Yelp page).*

*I didn't write a recursion that will gather data for all cafes yet because I was not quite sure if you would want or need this data.*

*And also, I'm quite interested in writing neural networks to do sentimental analysis on review comments (I've written RNNs before, and I know a decent amount of machine learning). If you want or need any help with that, I would love to help.*

*Jin*

*On Sat, Oct 13, 2018 at 8:10 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Jin,*

*Yes. Can you share some example of web scraping output? We can then compare it with API result.*

*Thank you*

*On Oct 13, 2018, at 4:00 PM, Jin Li <jinli7255@gmail.com> wrote:*

*I have 1500 instances of cafes in Chicago (but Yelp review has 2500).*

*Personally, I would rather web scrape data. It takes longer, but it is easier for me to manage, and web scraping allows access to significantly more data.*

*Do you want me to send a sample of what I can get by web scraping?*

*From,*

*Jin Li*

*On Thu, Oct 11, 2018 at 1:47 PM Jin Li <jinli7255@gmail.com> wrote:*

*I have data for businesses in Zip Code 60637 (not necessarily cafes)*

*On Thu, Oct 11, 2018 at 12:51 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Hi Jin,*

*Here is cafe list for Chicago from business license data I have cleaned over time.*

*How is Ben and your work doing with API? Can you share me any sample data you have so far?*

*Thanks!*

*Hyesun Jeong, PhD, AIA Associate*

*Postdoctoral Research Fellow*

*Urban Innovation Analysis, Inc.*

*https://sites.google.com/view/hyesunjeong*

*On Oct 11, 2018, at 8:37 AM, Jin Li <jinli7255@gmail.com> wrote:*

*Hey Hyesun,*

*I'm downloading API data from Yelp, and I am not sure if my method of data collection is correct. I just want to cross-check your data set with mine.*

*Thanks,*

*Jin*

*On Thu, Oct 11, 2018 at 10:04 AM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Hi Jin,*

*I heard you want to acquire the list of all cafes.*

*Can you specify what exactly you need, and what you are using it for, etc?*

*Thanks*

*Hyesun Jeong, PhD, AIA Associate*

*Postdoctoral Research Fellow*

*Urban Innovation Analysis, Inc.*

*https://sites.google.com/view/hyesunjeong*

*<cafe\_in\_chicago.csv>*

*<Greenline Coffee Reviews.csv>*

*<Screen Shot 2018-10-13 at 10.26.01 PM.png><Screen Shot 2018-10-13 at 10.32.13 PM.png><Screen Shot 2018-10-13 at 10.32.13 PM.png><Screen Shot 2018-10-13 at 10.32.13 PM.png>*

Things I Accomplished:

[*https://github.com/JinLi711/FAUI-Data-Scraping*](https://github.com/JinLi711/FAUI-Data-Scraping)

*Note that all documentation is in the code.*

*To sum up in a few lines:*

*I created code for getting data from API and for web scraping*

Goals:

*Figure out how long it takes to gather all data for cafes in yelp*

*Gather data for one region as a starter.*

*Figure out how to sort/ organize my collected data (including based on zip code)*

Things I Accomplished Part 2:

*So I calculated the time it takes to get all the data for Yelp for cafes in the US.*

*The problem with Yelp is that I am not allowed to search the entire U.S. or an entire state at once. The best I could do is search by Zip Codes. But the problem of searching for Zip Codes is that it also includes businesses near the Zip Code, but not in the Zip Code. This just makes scraping slower, but I could easily remove them after collecting the data.*

*I used a random number generator to generate Zip Codes, found the number of cafes in that Zip Code, then took the average. I found that there are about 30 cafes in each Zip Code. That does seem high, but businesses are liberally labeled to garner more search results (like McDonalds is considered a cafe) and the Zip Codes search includes cafes nearby not in Zip Code. From the census data, there are about 42000 Zip Codes in the US, meaning that there are about 1,300,000 search results from Yelp.*

*So calculating the runtime, I get about 10 seconds to gather all the data from one business, meaning about 160 days for the entire U.S. Basically not feasible.*

*At least, on my computer. More bad news: my laptop got blocked from accessing Yelp (not sure if this is for a few days, or forever).*

*But I did some more research, and here are some great news.*

*Apparently, I could scrape in parallel. Basically, I could run like 10 web scrapers at once, theoretically reducing 160 days to 16.*

*Also, I could run scraping programs on Google Cloud, controlled by my laptop. Essentially, it would be much faster and can hold larger amounts of data. Google Cloud is also very cheap (a few pennies per hour of run time!), easy to use, and easy to manage.*

*For Chicago Area: 90 zip codes. 2500 cafe businesses. Time as of now: about 7 hours*

*For Houston Area: 178 zip codes. 1284 cafe business. Time: 3.5 hours.*

[10/24/2018]

Tasks:

Learn how to avoid scraping traps.

Find Yelp data sets in Google

Learn to apply what I have on Yelp to see what I can get for Instagram, Facebook

Meeting Discussion:

* What leads to being blocked: repeated over querying
* 1 hour 18 min, Victoria's log
* What secondary sources can I get?
* Find data zizi papa chairman of department of communication of u Illinois
* Need my expertise

Things Completed

* Learned how to use Google Cloud to avoid being blocked (using Virtual Machines)
* Learned to “appear” human when scraping (adjusting query headers, implementing random time stops, deleting cookies)
* Learned that some web browsers (like Bing, DuckDuckGo) are easier to scrape

Don't worry about the law too much.

But keep note, I should respect robots.txt

But if I do collect the data, DO NOT give it to others.

Just don't sell or advertise that the data is mine.

Don't mess with financial or gov data.

Don't infringe on actual copyright stuff.

Don't enter anything that requires permission, passwords.

Don't scrape emails, usernames.

Don't spam forms.

- \*\* Set up Google Cloud Platform and learned to create Virtual Machines.

This is a big step forward for me because it allows me to run scripts on computers with IP addresses all over the world, meaning I can avoid being blocked. Also, running my programs on Google Cloud is faster.

After some hassle, I was able to run my python script successfully on a virtual machine in the Google Cloud.

However, this introduces another hurdle that I have to do next week: how do I collect the data? I do not know how my data is being stored and where it is going because I've never used Google Cloud Platform before.

- Learned to "look human".

(adjusting query headers, implementing random time stops, deleting cookies)

- Also another big problem: inconsistency.

This was really a low point for me because I would get inconsistent results when running my projects. I triple checked my code, and there seems to be nothing wrong with it. It may just be how my laptop is connecting to Yelp's interface.

But I did address this: I made a "label" to indicate whenever I miss some data, and I'm hoping that given this label, I can go back and get the data that I missed. So it takes longer, but not too much of a problem.

- Extra: Changed my browser. Apparently, Google really does not like web scrapers, I'm planning to scrape through DuckDuckGo (another browser) that pledges itself to maintain anonymity.

Email To Hysesun

*Jin Li <jinli7255@gmail.com>*

*Sun, Oct 28, 10:14 PM (3 days ago)*

*to Hyesun*

*My Task List was slightly different.*

*I apologize for taking so long to give you scraping results. The algorithm is already written, but I need to make adjustments to make sure I don't get blocked.*

*So here's what I accomplished this week:*

*- Did research on what is legal and what isn't.*

*Here's what I placed in my log:*

*Don't worry about the law too much.*

*But keep note, I should respect robots.txt*

*But if I do collect the data, DO NOT give it to others.*

*Just don't sell or advertise that the data is mine.*

*Don't mess with financial or gov data.*

*Don't infringe on actual copyright stuff.*

*Don't enter anything that requires permission, passwords.*

*Don't scrape emails, usernames.*

*Don't spam forms.*

*- \*\* Set up Google Cloud Platform and learned to create Virtual Machines.*

*This is a big step forward for me because it allows me to run scripts on computers with IP addresses all over the world, meaning I can avoid being blocked. Also, running my programs on Google Cloud is faster.*

*After some hassle, I was able to run my python script successfully on a virtual machine in the Google Cloud.*

*However, this introduces another hurdle that I have to do next week: how do I collect the data? I do not know how my data is being stored and where it is going because I've never used Google Cloud Platform before.*

*- Learned to "look human".*

*(adjusting query headers, implementing random time stops, deleting cookies)*

*- Also another big problem: inconsistency.*

*This was really a low point for me because I would get inconsistent results when running my projects. I triple checked my code, and there seems to be nothing wrong with it. It may just be how my laptop is connecting to Yelp's interface.*

*But I did address this: I made a "label" to indicate whenever I miss some data, and I'm hoping that given this label, I can go back and get the data that I missed. So it takes longer, but not too much of a problem.*

*- Extra: Changed my browser. Apparently, Google really does not like web scrapers, I'm planning to scrape through DuckDuckGo (another browser) that pledges itself to maintain anonymity.*

*- Again, sorry for not producing results. But the great thing about coding is the scalability of what I do. The first step is long and annoying, but once I pass this step, there would be a lot more data coming in.*

*Future Goals:*

*- Learn how to use task manager (so I can run scripts when I'm sleeping)*

*- Learn to use Google Cloud Platform to gather data that I collect.*

*More goals to come.*

*On Sun, Oct 28, 2018 at 5:20 PM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Hi Jin*

*Here is what I summarized Wed meeting discussion for next task:*

*Task for Jin Li*

*Web scraping for Yelp reviews and informations for cafe - All zip codes in Chicago first, then move on to larger geographic level*

*We need to see more consistent report on the scraping result.*

*Find a way to utilize Instagram’s hashtag (need to see precedence or reference)*

[10/31/2018]

Tasks:

Check out Octoparse

Finish learning how to use Google Cloud (learn how to run algorithms at night time)

Scrape data (produce results)

Things Accomplished:

Looked through professional web scraping sites. A lot of them are like hundreds of dollars per month, and that’s for the basic plan (high limitations).

Checked out [ScrapingHub](https://scrapinghub.com/), which can be a huge tool for me. However, I would have to learn how to use scrapy, and I do not yet have time for this (I would have to restructure my code). But I’m considering learning this over the Christmas break.

Got 1000 websites.

Tested my code through GCP, turned out to be a failure.

Tried to cross compare what I scraped with census data.

Problems Encountered:

Octoparse requires a Windows operating system, and I have a mac

Only Octoparse seems to provide free access.

For some reason, I can’t gather data from Yelp through Google Cloud Platform.

Also, GCP requires a huge time investment, which I don’t have. I plan to either read up on this on winter break or something.

Link to instance [number 990](https://www.yelp.com/search?find_desc=cafe&find_loc=Chicago,+IL&start=990) works, but [1000](https://www.yelp.com/search?find_desc=cafe&find_loc=Chicago,+IL&start=1000) doesn’t. This is a problem because there are 2500 cafes in Chicago. But I already collected 1000 websites.

Future Goals:

*Learn to use Scrapy (and ScrapingHub) during Christmas break.*

*Somehow get access to Octoparse.*

*Check out* [*https://www.parsehub.com/*](https://www.parsehub.com/)

Email to Hyesun

*Never mind I found the csv file you sent me.*

*On Sat, Nov 3, 2018 at 10:34 AM Hyesun Jeong <j224sun@gmail.com> wrote:*

*Hi Jin,*

*Census data only has the aggregate number for each zipcode.*

*Do you need that?*

*If you need each list of individual cafes, I have business license data I sent you last time (as csv)*

*Let me know if you need census data then.*

*Thanks!*

*Hyesun Jeong, PhD, AIA Associate*

*Postdoctoral Research Fellow*

*Urban Innovation Analysis, Inc.*

*https://sites.google.com/view/hyesunjeong*

*On Nov 2, 2018, at 11:10 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Dear Hyesun,*

*I was able to scrape 1000 business links (from the Yelp search cafe, Chicago). The great thing is is that the names of the businesses are in the link. Do you have census data so I could compare my results?*

Possible People to Contact:

Paper on [Creating](https://mag.uchicago.edu/science-medicine/artificial-intelligence-meets-yelp) Fake Yelp Reviews (UChicago)

[Kevin](http://noithytiff.com/) Yuanshun Yao

[Heather Zheng](http://people.cs.uchicago.edu/~htzheng/index.html)

[Ben Y. Zhao](http://people.cs.uchicago.edu/~ravenben/)

Google Scholar Search: yelp cafe

No relevant results

Google Search: yelp scraping

Paper: [Yelp](https://pdfs.semanticscholar.org/9751/83a458ed6340d9de12d3431fa9da8bf9a912.pdf) Review Manipulation and Social Effects

Check the Methodology Section

(they don’t seem to be experts at web scraping, but they got it down, albeit very slowly)

Elaine Arbaugh (not uchicago)

Zachary Fein

Google Search: uchicago web scraping

\*[Benjamin](http://www.bensoltoff.com/) Soltoff, [lecturer](https://cfss.uchicago.edu/) in computational social science at the University of Chicago. He teaches a course on web scraping for social science.

Check out:

A [Comparative](http://ir.kdu.ac.lk/handle/345/1051) Study on Web Scraping

Web [scraping](http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1119445&dswid=-2452) to monitor the client and the competitor’s brands performance and product offering in an online space

Things Accomplished:

Reorganized / updated my code.

Looked through multiple sites to see how I can possibly access instagram’s API.

Found professors to contact for help on scraping.

Some links:

[Stack](https://stackoverflow.com/questions/41405697/how-to-extract-instagram-data)

[Medium](https://medium.com/@h4t0n/instagram-data-scraping-550c5f2fb6f1)

[Endpoints](https://www.instagram.com/developer/endpoints/)

[Blog](https://later.com/blog/instagram-api/)

[Github API](https://github.com/facebookarchive/python-instagram)

Problems:

I don’t have much to do.

Instagram API is no longer public and requires registration.  
 Instagram API is a bit confusing for me to use.

For some reason, when I tried to register, it won’t allow me.

[11/14/2018]

*Email To Hyesun:*

*Dear Hyesun,*

*I’ll just describe what I did:*

*1) Gather all possible zip codes. (This was not hard to do, I just googled zip codes, and there was a csv file on all zip codes).*

*2) Create another csv file (1 column) for the websites to visit. For example, a row of the csv file would be like https://www.yelp.com/search?find\_desc=cafe&find\_loc=60007.*

*3) Create another csv file that’s just like the one described above, except it only contains zip codes for Cook County (Chicago).*

*4) Scrape websites based on that Cook County csv file. Note, I only did a few zip codes (including 60637), because my scraping is not entirely automated. I still need to discuss with some grad students/professors before taking the next step. If you want to see more zip codes, I could do it, but I would have to do some more manual clicking and waiting.*

*5) Find all categories of Yelp (see attached csv file).*

*6) Pick ones to filter (I chose: 'burgers', 'fast food', 'convenience stores', 'gas', 'bars', 'nightlife', 'delis', 'home'). I tried to be conservative about my choices.*

*7) Now that I have data scraped, I filter the websites. I remove all duplicate websites.*

*8) Then output the result to a (1 column) csv file containing just links of websites to scrape.*

*Note that there are two parts of scraping (first getting the websites, then getting the reviews, addresses, etc for each website.*

*Also, I scraped raw data and filtered it afterward, instead of filtering while scraping. This is because it is not only easier to do, but also because we can keep more data.*

*Any suggestions?*

*Questions for the future:*

*1) When scraping Yelp reviews, do you want all Yelp reviews? Some businesses have thousands of reviews, making the process a lot slower. If we have a cutoff, the process would be a lot quicker.*

*2) Are there any other criteria by which you want me to filter Yelp websites?*

*3) What are your opinions on business chains? (like Starbucks, Dunkin Donuts)?*

*4) Also, could I have your phone number for better communications? Some of my emails have been quarantined before (including ones from UIA).*

*Thanks,*

*Jin Li*

Things Accomplished:

Helped Alexis scrape about 20 businesses for the top 50 reviews.

Organized the data, and used Python to create charts to describe word frequency of reviews in each cafe

Gathered zip codes, created csvs for websites to visit.

Filtered websites

Problems:

I’m a bit slow on the function names because I’m also learning R, and the functions are getting jumbled in my head.

I still have not visited professors (will do this on Monday).

My scraping process is still not incredibly automated.

11/20/2018

Hyesun Email:

*Hi Alexis and Jin,*

*Good collaboration.*

*Charts are cut off in the bottom as you can see. Why don’t you adjust them to make them fully visible?*

*Do you remember we discussed the main categories of key words in the beginning - food, atmosphere, surrounding, people, etc.?*

*The keywords in those charts are more likely separate individual words that you can characterize into groups of those main categories.*

*Most reviews have a higher frequency of food related words, like coffee, lattes, food, etc. However, we are not meant to analyze whether latte or sandwich is more popular in each cafe.*

*The previous work that analyzed social media by others was simply testing correlation of different food menus of cafes, without any implications of how place serves culture.*

*Thus, we also want to see how people engage in surrounding of cafes beyond utilitarian function of the place. This includes non-food key words, like*

*1. Types of (Participatory) cultural activities outside cafe - art, ambience, gallery, festival, streetart, mural, event, community, demonstration, parade, shop, etc.*

*2. Types of patrons - alone, group, friends, couple, family, kid, etc.*

*3. Types of activities inside cafe - laptop, meeting, work, study, talk, chat, date, space, etc.*

*These can be still rough category, but why don’t you play around more possible keywords and show them with main or sub categories?*

*This is not exactly equivalent to Scenes coding amenities, but we want to have more distinct and subtle keywords to be measured alternatively through social media.*

*How is analyzing photos on Instagram going? That would potentially have similar method as this text scraping, although most people would post photos on food in general on the pages of cafe. What do you think?*

TNC Email:

*Jin Li,*

*the two main things for you to read and comment on work for you that are important is first Hyesun’s big LOG and second my effort to build propositions which can lead to regression models in the Cafe Propositions paper draft of a week or so back.*

*the context for all this is the Scenescanes book, which will help you see what we are doing more generally.*

*I dont think I have seen any reactions to these from you?*

*We want to challenge you to engage both with the programming and the testing of ideas specifically so you understand and are creative in all you do.*

*If you try to work from a Cookbook you wont learn chemistry, analogously.*

*Glad to skype or email though the Thanksgiving holiday if you read and react to these.*

*TNC*

*On Nov 20, 2018, at 11:25 PM, Jin Li <jinli7255@gmail.com> wrote:*

*I plan to do a skype chat with Alexis (or meet in person) before I start testing some ideas out. I will write everything up in a log.*

*On Tue, Nov 20, 2018 at 11:20 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:*

*i dont have his email. Dan should reply in a few days about how far along he is.*

*Our MAIN jobs are as I wrote; we need your help to prepare THEM now. Look for analysis models that we can adapt, types of statical / tables, restuls. Simple or complex. Start with the*

*Simple. jUst list in your log as possible models to adapt to our work.*

*will let you know when I hear more.*

*tnc*

*On Nov 20, 2018, at 11:11 PM, Jin Li <jinli7255@gmail.com> wrote:*

*Is there any way I could communicate with Zack? I think I can help speed up the process.*

*On Tue, Nov 20, 2018 at 5:50 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:*

*BIG NEWS*

*He will give us all the US for all categories like bars and cafes etc and comments by visitors up to 200? or so comments. IN a few days hopes to know when done. using many many spiders. Zack I think is the assistant doing this.*

*So lets shift to HOW to analyze when we have the data and to iclukde e.g. how to join / filter by key words, linking Alexis’s work to more general modeling of key words and in turn to*

*other variables.*

*I have recording on recent work by Dan on OneDrive /Scenes staff recording*

*15 min. video. mp4 Nov 20 2018*

Email to Alexis:

*Alexis Pearson*

*Wed, Nov 21, 7:52 PM (4 hours ago)*

*to me*

*Let's do tomorrow midday like noon?*

*Alexis Pearson*

*UChicago- Harris School- MPP Student*

*\*please excuse the potential errors, this was sent from a cell phone\**

*From: Jin Li <jinli7255@gmail.com>*

*Sent: Wednesday, November 21, 2018 8:12:54 PM*

*To: Alexis Pearson*

*Subject: Re: FAUI Cafe*

*I'm sorry I completely forgot about our call. Could we plan another time? Either today or tomorrow morning?*

*On Wed, Nov 21, 2018 at 1:47 AM Alexis Pearson <pearsona@uchicago.edu> wrote:*

*Yes that's fine. My number is 734-674-8741. Plesse call me if I forget*

*Alexis Pearson*

*UChicago- Harris School- MPP Student*

*\*please excuse the potential errors, this was sent from a cell phone\**

*From: Jin Li <jinli7255@gmail.com>*

*Sent: Wednesday, November 21, 2018 12:25:19 AM*

*To: Alexis Pearson*

*Subject: Re: FAUI Cafe*

*How about 4:30? I have class until 4:20.*

*On Tue, Nov 20, 2018 at 11:36 PM Alexis Pearson <pearsona@uchicago.edu> wrote:*

*I can call around 4 central time?*

*Alexis Pearson*

*UChicago- Harris School- MPP Student*

*\*please excuse the potential errors, this was sent from a cell phone\**

*From: Jin Li <jinli7255@gmail.com>*

*Sent: Tuesday, November 20, 2018 11:35:05 PM*

*To: Alexis Pearson*

*Subject: Re: FAUI Cafe*

*My phone number is 646 472 6952*

*On Tue, Nov 20, 2018 at 11:34 PM Jin Li <jinli7255@gmail.com> wrote:*

*Time?*

*On Tue, Nov 20, 2018 at 11:30 PM Alexis Pearson <pearsona@uchicago.edu> wrote:*

*I'm traveling tomorrow but we can talk over the phone*

*Alexis Pearson*

*UChicago- Harris School- MPP Student*

*\*please excuse the potential errors, this was sent from a cell phone\**

*From: Jin Li <jinli7255@gmail.com>*

*Sent: Tuesday, November 20, 2018 11:26:58 PM*

*To: Alexis Pearson*

*Subject: FAUI Cafe*

*Hey Alexis,*

*Could we meet sometime tomorrow? Skype is also fine.*

*Planned discussion:*

*Ways to represent frequency of words.*

*Best,*

*Jin Li*

Goals:

Help Alexis document her qualitative analysis

Discuss with Alexis better ways to represent word frequency (and implement them)

Compare with Bizzip the results of Dan’s scrape

Read through Hyesun’s log and Clark’s files.

Looking at categories (trying to determine type of cafes), at artistic relevant words (I know some cafe are very art linked). Look at words to determine political connections (look at locations too) Think of a way to contrast with bars. Look at pictures.

Skype Discussion With TNC (Main Points and My Goals)

Spend some time looking at the scenes book

Learned about multilevel modeling and Geographically weighted regression (GWR)

Look at the verbal theory of math model (see the blalock paper and the models). My goal is to make pictures like this.

Douglas Noonan (look at how he analyzes historic (artistic) distribution and monuments and their effects on economic growth)

I have do do literature reviews before doing data analysis.

Look back at the 4 paragraphs TNC sent me and work backwards from the last paragraph

I need to be able to “link multiple discussions” (like how keywords from Yelp link to politics (Trump, Clinton, black lives matter). How this can be related to voting, job decline, unemployment, migration. Even small mentions of keywords can be key.) and build regressions

Find a way to compare cafes with bars (how are they linked, how do they tell different stories about the political atmospheres)

I need access to the Yelp data set that Christina has (get it from Ben or Hyesun)

Things Accomplished

Looked through the scenes book and made highlights/notes (read through the important points of the chapters, but I focused in on the Science of Scenes to get an idea of how I can find ways to fit Yelp data into Bizzip and Yellow Pages)

Took notes on multilevel modeling, Geographically weighted regression (GWR), and the modeling of Blalock’s box graphs

[Came](https://docs.google.com/document/d/1b1pPOKbBz7f7ZUgyhyAfj9MO_rBtjk5aw-zjCrJPsfE/edit) up with preliminary ideas that I could possibly investigate.

[11/28/2018]

Email from Alexis

*Sorry I was not very good at communicating the past couple of days. I was at home and did not have much time to sit in front of a computer. I was thinking for graphical representation of the word frequency we could try and do the following:*

*Search through the comments for these specific words:*

*Key words: espresso, coffee, latte, mocha, bakery, food ,decoration, style, seating, picture, lighting, community, quietness, music, space, tables, outlets, wifi, “vibe”, event, mural, art, signage, relationship to public transit /Metra/ L station / bus stop/ subway, graffiti, neighborhood/community festival, event, farmers market/flea market/ seasonal market, friends, studying, meetings, children, hang out, dating, meeting, working*

*Make bar graphs (like you did before) to show the frequency of these words per cafe*

*Then make a graph that would show the frequency of these words per neighborhood (bohemian, corporate, neighborly)*

*I hope this is helpful. Let me know if you have any questions*

Some People To Contact for Help on Linguistics:

[Jason Riggle](https://linguistics.uchicago.edu/faculty/riggle)

[Ming Xiang](https://lucian.uchicago.edu/blogs/mingxiang/)

[John Goldsmith](https://linguistics.uchicago.edu/faculty/goldsmith)

Things Accomplished:

Add more specific details to my notes

Set up a meeting with linguistic professor

researched how to do text analysis

Added in quotes from scenes that I would like to address

Expanded on Alexis’s task on representing frequencies of words

Goals:

Learn to understand how to “algorithmically” understand the human language

[12/19/2018]

Goals:

Create word embedding for yelp data

Gather more data for myself (since we don’t have the data yet)

Represent the data in a chart

Things Accomplished:

Researched word embedding

Wrote a program to create Word2Vec for a few thousand reviews

Documented my code

Found and created lists of words to filter

tried different dimension reductions (TSNE, ISOMAP, PCA)

improved my pipeline for filtering and normalizing text

tried and compared cbow and skip gram

thought about how to create a visual for zip codes

thought about ways of measuring clusters

Created a write-up.

[12/26/2018]

Hyesun, TNC, Alexis Correspondence:

Are you suggesting that I should drop the scatter plot and just do a graph like yours?

Sorry for not clarifying: x and y doesn't mean anything useful in terms of presentation. What should be focused on is the location of the dot with respect to each other, and the x and y measures that (but what you want is a visual, so that measurement doesn't matter).

For example, if the words donut, bagel, friends, and social were to be the scatter plot,

donut would be close to bagel, and friends would be close to social.

Dots that are closer together have similar meaning, so when a person looks at the scatter plot, they immediately see the separation of words under groups of words, and they get a feel of frequency by looking at the size of the dot.

If you want to compare a lot of words, a bar graph is hard to look at, and there is no intuition behind the ordering of the x-axis. Example: the attached picture.

Open to suggestions / counterarguments.



On Mon, Dec 24, 2018 at 11:34 AM Hyesun Jeong <j224sun@gmail.com> wrote:

Hi Jin

I agree that it is hard to know what x and y axis mean.

I get that darker red means more frequency, but what is y axis for?

Can you simplify the chart like the one I made up below? X means key word, Y means frequency. Color bars mean 3 neighborhood types.

This is just a quick example I made up in excel, but you see my point?

Hyesun Jeong, PhD, AIA Associate

Postdoctoral Research Fellow

Urban Innovation Analysis, Inc.

<https://sites.google.com/view/hyesunjeong>

On Dec 23, 2018, at 4:46 PM, Jin Li <jinli7255@gmail.com> wrote:

I did it for 3 zip codes by doing some small amount of scraping. 3 businesses for each zip code. I'll do small amounts right now so the python program runs faster.

The size of the dot means the same thing as the color (which is the frequency of the word).

In the future, I hope to color code the dots based on the word category (social interaction, transportation, art/culture).

Also, I can do some machine learning labeling so you don't have to manually label zip codes (or cafes) as bohemian, corporate, tradition, etc. (Do you think this is worth doing?)

I'm still working on filtering words (which I have to do manually), so the plots I sent only have your keywords. But if you want to see a plot will all "interesting" words, I can send you one.

I'm also trying to fix the words on the graph so they don't overlap.

Suggestions?

On Sat, Dec 22, 2018 at 9:28 PM Jin Li <jinli7255@gmail.com> wrote:

What’s the progress on dan silver’s scraped data?

Sent from my iPhone

On Dec 22, 2018, at 7:30 PM, Terry Nichols Clark <tnclark@uchicago.edu> wrote:

Sounds clear and doable I hope?

Sent from my iPhone

Terry Nichols Clark

Professor of Sociology

University of Chicago

tnclark@uchicago.edu

On Dec 22, 2018, at 5:46 PM, Hyesun Jeong <j224sun@gmail.com> wrote:

Hi Jin,

It is hard to guess how it looks like without actual sample of results.

As we have discussed in meeting many times, we are not expecting thousands zip codes at once. We want to see more step-by-step result from zip codes in Chicago, to Cook County, and to larger region. How is it feasible?

How about starting with 55 zip codes in Chicago?

Visualization does not need to be fancy or technical, but just need to show 1) frequency of key words of cafes, and 2) types of neighborhood (bohemian, corporate/utilitarian, traditional/local/neighborly)

A simple basic question to pursue in your work as an example: What are the key words of cafes in each type of neighborhood that have more frequency? Do bohemian cafes have more key words related to community and cultural activities of surrounding? If you use machine learning, I believe it will automatically generate all key words by itself without setting a list of keywords like we did at first. Right?

Please share me some actual sample of zipcode, so I can take a look and advise.

Thank you

Hyesun Jeong, PhD, AIA Associate

Postdoctoral Research Fellow

Urban Innovation Analysis, Inc.

<https://sites.google.com/view/hyesunjeong>

On Dec 21, 2018, at 5:04 PM, Jin Li <jinli7255@gmail.com> wrote:

I was able to perform the same thing but split into different neighborhoods (Bohemian, corporate, etc). The code is generalized so that I can do different categories (so in the future, I can just replace the current categories I'm using with zip codes instead).

I was also able to get and create lists of words to filter out. I used google's csv that has 20,000 most common words, and I created my own list of words to filter out (mostly generic food words like muffins, cupcakes, etc). My thinking was that it would be interesting to look at uncommon words, which may say something about the community.

Filtering out the top 20,000 words may reduce the information that I can analyze, but the only other option would be for me to filter words manually, which would take a long time.

Also I noticed that when I produced the models for each category, the models look very different, and that's simply because of the randomness involved in language and the lack of data that I have. My solution to this would be to use a baseline model (like the one created by google, where they used a similar algorithm but on hundreds of trillion of words.) Then using that baseline, train the yelp data on top of it for each category. Then for each interesting word, measure the deviation from the Google model.

Also, I'm having trouble producing a visual that can generalize well. Yea, I can produce thousands of charts or graphs, one for each zip code, but no one's going to look at that.

So how can we visualize the distribution of words in a way that is scalable (to tens of thousands of zip codes)?

Also, I think I'm having trouble in terms of finding what exactly to measure that can be meaningful. Plotting frequencies of words is easy, but do you have any other suggestions about what I should be measuring? I guess this is the problem of trying to apply quantitative methods to something that is very qualitative.

Suggestions?

On Wed, Dec 19, 2018 at 6:31 PM Hyesun Jeong <j224sun@gmail.com> wrote:

Hi Jin

Thanks for sending the progress.

In general, machine learning test seems good strategy. However, what’s missing here is the relationship with geographic locations of cafes.

Which cafes have more cluster of what kinds of key words and how are they different by location (or zipcode) of cafes?

Are there any comments on the neighborhood in reviews? Not the general term of neighborhood, but cultural dimensions specific to different neighborhoods around cafe.

Can you probably use this to advance to differentiate zip codes of cafe?

Hyesun Jeong, PhD, AIA Associate

Postdoctoral Research Fellow

Urban Innovation Analysis, Inc.

<https://sites.google.com/view/hyesunjeong>

On Dec 19, 2018, at 5:25 PM, Jin Li <jinli7255@gmail.com> wrote:

Here's a really rough plot that I was able to do.

I can technically plot all the words, but there are thousands, so if I plot them all you would not be able to see anything. So I just plotted Alexis's words for now just for a visual reference.

If you think this is interesting to consider, I will spend more time on this.

Also, the results aren't that great because I was only able to do this on thousands of reviews (usually people would do this on millions of sentences).

But even so, we can still see some interesting things going on, like:

on the right side: vibe, expresso, mocha, music are clustering together, which indicates that that in Yelp reviews, these words are connected more so than other words.

On Mon, Dec 17, 2018 at 9:22 PM Jin Li <jinli7255@gmail.com> wrote:

I'll get working on it, but just note that I've never done something like this before, so it might take some time. Also, NLP only works well if done on a lot of data (and I mean like a lot).

For scraping, the task has been given to Dan Silver, and I'm not sure what the progress is.

On Sun, Dec 16, 2018 at 10:12 PM Hyesun Jeong <j224sun@gmail.com> wrote:

Hi Jin,

I think machine learning would be more systematic way to generate a series of words by itself.

Can you do test run it for cafes in Chicago?

Also, I have been lost of what is the latest progress of scraping. Do we already have result of all cafes in Chicago or what is the state of work?

Hyesun Jeong, PhD, AIA Associate

Postdoctoral Research Fellow

Urban Innovation Analysis, Inc.

<https://sites.google.com/view/hyesunjeong>

On Dec 16, 2018, at 8:40 PM, Jin Li <jinli7255@gmail.com> wrote:

And just to be more specific, this is what I visualize:

I think the problem with the current approach is that we come up with words and then find them in Yelp (what if there are useful words or categories of words to analyze that we haven't come up with). But using machine learning and natural language processing (NLP), we go the opposite direction (find words on Yelp first).

The photo below uses the embedding technique to group words together based on their similarities: the closer the red dots are to each other, the more related they are.

- We can apply the same technique to separate Yelp keywords into categories that Hyesun described (social interaction, transportation, art, and culture) without actually being explicit about it (and maybe the algorithm will create clusters of words that we did not expect).

- We can plot the word frequencies using scatter plots instead of bar graphs (like the graph below, but also have a bigger red dot to indicate greater usage).

- We can also create two different graphs (one for cafe, one for bars) and compare them.

<image.png>

Here's another picture that has the same premise, but is in 3D and the words are color-coded based on categories.

<image.png>

On Sun, Dec 16, 2018 at 4:10 PM Jin Li <jinli7255@gmail.com> wrote:

Alexis, could you actually send me your methodology write up so I could add to it?

Also, I think we will have enough data for us to do more advanced analysis (like natural language processing) and not just searching for keywords. What do you think?

On Wed, Dec 12, 2018 at 3:00 PM Terry Nichols Clark <tnclark@uchicago.edu> wrote:

Hyesun thanks for making this more explicit. It may be that initial focus only on one ZIP Code in Chicago lead to a low proportion of frequency of mansions of the more meaningful term example is on our list. But as we have more included with a larger number of cases across the whole country this should be less of an issue. So let’s not drop a bigger concepts just because we have so few cases are no cases in Hyde Park.Paragraph new paragraph we’re seeing more and more creative work along these lines which I’ll forward it’s up to you and we can talk about and explore more specific options about how to apply based of what you’re doing. Like the BA thesis by Tristan.

I just created two new folders on Dropbox for us to share these kinds of documents. We also need the files from Alexis and Jun to be there so you and others can see how they relate to work in China and other places outside Hyde park

Sent from my iPhone

Terry Nichols Clark

Professor of Sociology

University of Chicago

tnclark@uchicago.edu

On Dec 11, 2018, at 8:50 PM, Hyesun Jeong <j224sun@gmail.com> wrote:

Hi Alexis and Jin,

You said these words:

1. coffee
2. friends
3. latte
4. meetings
5. studying
6. etc.

I think these key words are too generic and oversimplified, and have no relationship to other urban surroundings.

The initial set of key words we discussed is:

* 1. Key words: espresso, coffee, latte, mocha, bakery, food ,decoration, style, seating, picture, lighting, community, quietness, music, space, tables, outlets, wifi, “vibe”, event, mural, art, signage, relationship to public transit /Metra/ L station / bus stop/ subway, graffiti, neighborhood/community festival, event, farmers market/flea market/ seasonal market, friends, studying, meetings, children, hang out, dating, meeting, working

How come all others related to the arts and cultural elements are gone?

Key words like latte and coffee have no further interpretations other than knowing what people order.

I suggest have both big categories and subcategories in key words.

For example, key words like “meeting”, “group study”, “talking”, “friends”, “chatting” should be categorized as SOCIAL INTERACTION

Other key words like “L”,”bus stop”, “Metra”, “subway” can be under TRANSPORTATION

“Graffiti”, “streetart”, “mural”, “gallery”, “festival”, etc can be under ART & CULTURE

How about re-generating this kind of hierarchy with more specific key words under some types of categories?

Hyesun Jeong, PhD, AIA Associate

Postdoctoral Research Fellow

Urban Innovation Analysis, Inc.

<https://sites.google.com/view/hyesunjeong>

On Dec 5, 2018, at 5:00 PM, Alexis Pearson <pearsona@uchicago.edu> wrote:

This is an update on the visuals that have been created.

Hey Jin,

Thank you for classifying the graphs by neighborhood. I think the next thing we need to work on is trying to reduce the words we have visualized sticking to the ones that Hye Sun has suggested as the high frequency words we should be looking for. For example words like:

1. coffee
2. friends
3. latte
4. meetings
5. studying
6. etc.

It was mentioned in the email before I will send it again if you need those words.

Also can you email me a list of what you did methods wise so I can add it to a total methodology list.

Hye Sun,

Any ideas or suggestions?

**Alexis Pearson**

University of Chicago-Harris School| MPP Student

**From:** Jin Li <jinli7255@gmail.com>

**Sent:** Thursday, November 29, 2018 7:36 PM

**To:** Alexis Pearson

**Subject:** Re: FAUI Cafe

I'm going to start thinking about more complex analysis of words. But since finals week is approaching, I'm probably not going to do much for the next two weeks. But I'll do a lot during the Break.

On Mon, Nov 26, 2018 at 10:29 AM Alexis Pearson <pearsona@uchicago.edu> wrote:

Hey,

Sorry I was not very good at communicating the past couple of days. I was at home and did not have much time to sit in front of a computer. I was thinking for graphical representation of the word frequency we could try and do the following:

1. Search through the comments for these specific words:
   1. Key words: espresso, coffee, latte, mocha, bakery, food ,decoration, style, seating, picture, lighting, community, quietness, music, space, tables, outlets, wifi, “vibe”, event, mural, art, signage, relationship to public transit /Metra/ L station / bus stop/ subway, graffiti, neighborhood/community festival, event, farmers market/flea market/ seasonal market, friends, studying, meetings, children, hang out, dating, meeting, working
2. Make bar graphs (like you did before) to show the frequency of these words per cafe
3. Then make a graph that would show the frequency of these words per neighborhood (bohemian, corporate, neighborly)

I hope this is helpful. Let me know if you have any questions

**Alexis Pearson**

University of Chicago-Harris School| MPP Student

**From:** Alexis Pearson

**Sent:** Wednesday, November 21, 2018 7:51:59 PM

**To:** Jin Li

**Subject:** Re: FAUI Cafe

Let's do tomorrow midday like noon?

Alexis Pearson

UChicago- Harris School- MPP Student

\*please excuse the potential errors, this was sent from a cell phone\*

**From:** Jin Li <jinli7255@gmail.com>

**Sent:** Wednesday, November 21, 2018 8:12:54 PM

**To:** Alexis Pearson

**Subject:** Re: FAUI Cafe

I'm sorry I completely forgot about our call. Could we plan another time? Either today or tomorrow morning?

On Wed, Nov 21, 2018 at 1:47 AM Alexis Pearson <pearsona@uchicago.edu> wrote:

Yes that's fine. My number is 734-674-8741. Plesse call me if I forget

Alexis Pearson

UChicago- Harris School- MPP Student

\*please excuse the potential errors, this was sent from a cell phone\*

**From:** Jin Li <jinli7255@gmail.com>

**Sent:** Wednesday, November 21, 2018 12:25:19 AM

**To:** Alexis Pearson

**Subject:** Re: FAUI Cafe

How about 4:30? I have class until 4:20.

On Tue, Nov 20, 2018 at 11:36 PM Alexis Pearson <pearsona@uchicago.edu> wrote:

I can call around 4 central time?

Alexis Pearson

UChicago- Harris School- MPP Student

\*please excuse the potential errors, this was sent from a cell phone\*

**From:** Jin Li <jinli7255@gmail.com>

**Sent:** Tuesday, November 20, 2018 11:35:05 PM

**To:** Alexis Pearson

**Subject:** Re: FAUI Cafe

My phone number is 646 472 6952

On Tue, Nov 20, 2018 at 11:34 PM Jin Li <jinli7255@gmail.com> wrote:

Time?

On Tue, Nov 20, 2018 at 11:30 PM Alexis Pearson <pearsona@uchicago.edu> wrote:

I'm traveling tomorrow but we can talk over the phone

Alexis Pearson

UChicago- Harris School- MPP Student

\*please excuse the potential errors, this was sent from a cell phone\*

**From:** Jin Li <jinli7255@gmail.com>

**Sent:** Tuesday, November 20, 2018 11:26:58 PM

**To:** Alexis Pearson

**Subject:** FAUI Cafe

Hey Alexis,

Could we meet sometime tomorrow? Skype is also fine.

Planned discussion:

Ways to represent frequency of words.

Best,

Jin Li

<plots.zip>

<Word\_Embedding\_Plot.jpg>

<60603\_Word\_Embedding\_Plot.jpg><60637\_Word\_Embedding\_Plot.jpg><60626\_Word\_Embedding\_Plot.jpg>

Things Accomplished:

read through Hyesun's Retail Conference Slides

created many graphing ideas

Goals:

Discuss art project with Ben

What I Wrote Down For Graphing Ideas:

Graphing Ideas

**Purpose:**

Create a graph (or graphs) that is not only representative of the data, but also scalable with the data. This means that if the data available grows very large, the usefulness of the graph will not diminish when it tries to represent the data.

Inspired by: Handbook of Data Visualization

**Things to Consider:**

* The words that we are comparing will only increase, and become more niche
* The starting words that we have now are very broad.
* Some words may be specific to only one group (Bohemian, corporate, traditional)
* Is there a way to seperate words into categories by machine learning?
  + Doing it by hand may be tedious.
  + Also, there may be other groups of words that are significant
* We can use 2 or more graphs jointly
* The graph should be as simple as possible, but also has to convey as much meaning as possible. The reader should not have to strain their brain to understand a graph because the whole purpose of a graph is to easily convey information to the audience.

**Example of category and words:**

Social Interaction

* Meeting
* Group study
* Talking
* Friends
* Chatting

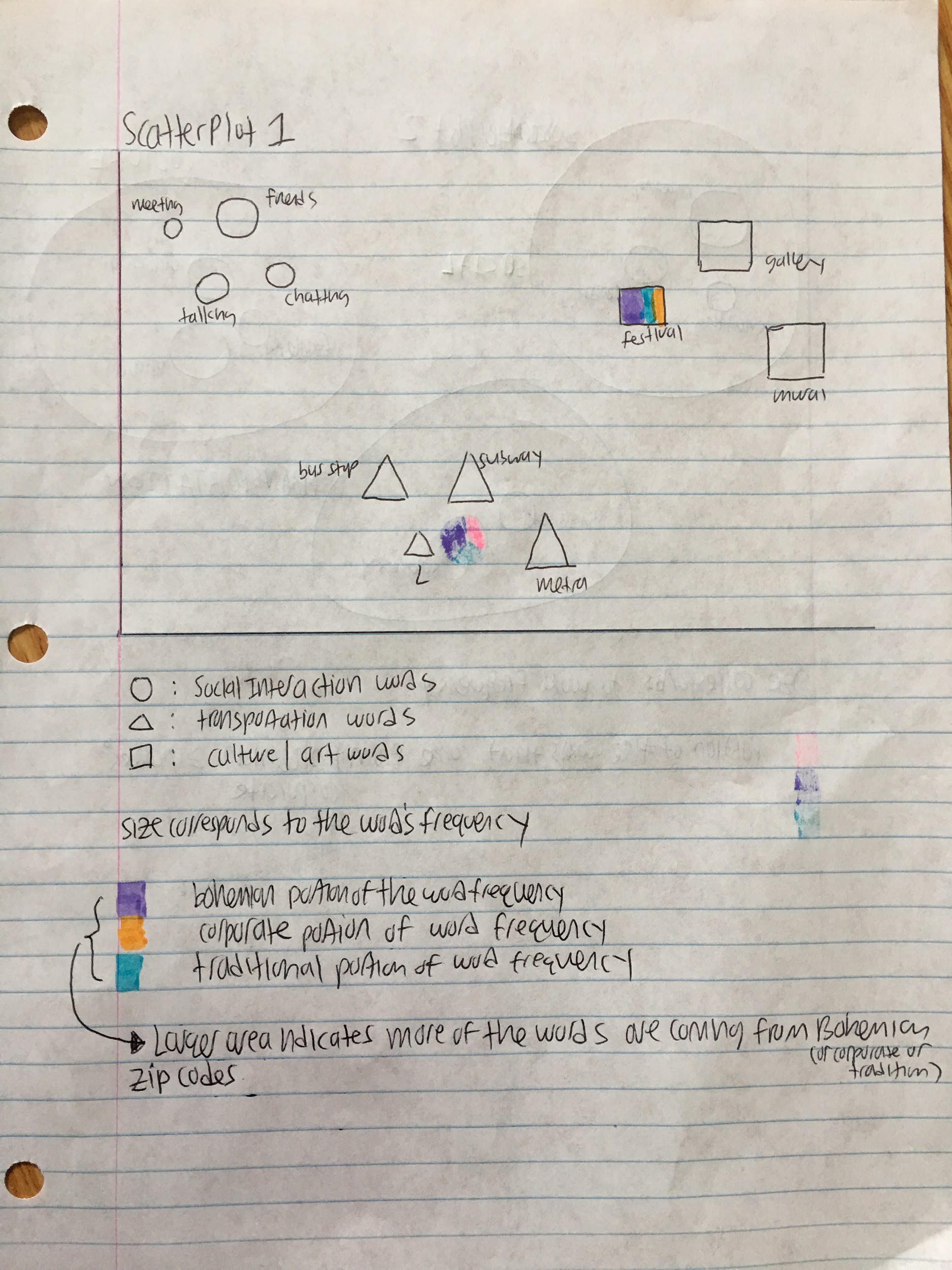
Transportation

* Bus stop
* L
* Subway
* Metra

Culture

* Graffitti
* Street art
* Mural
* Gallery
* Festival

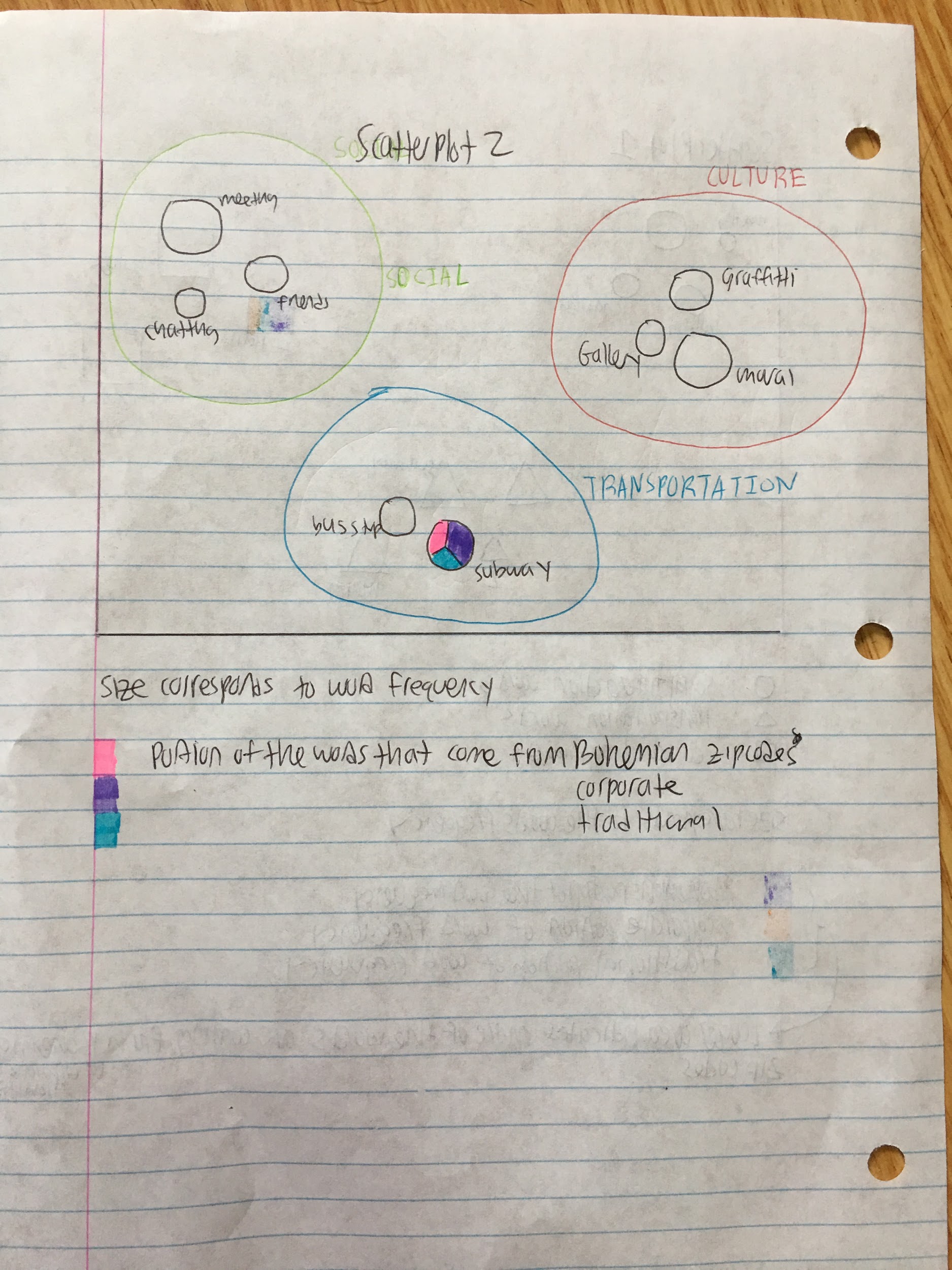
**Plots:**

**Scatterplot 1:  
**

Cons:

* the different shapes confuses the sizes (how do we compare the size of the triangle to the size of the circle)?
* X, y axis don’t have any significant meaning

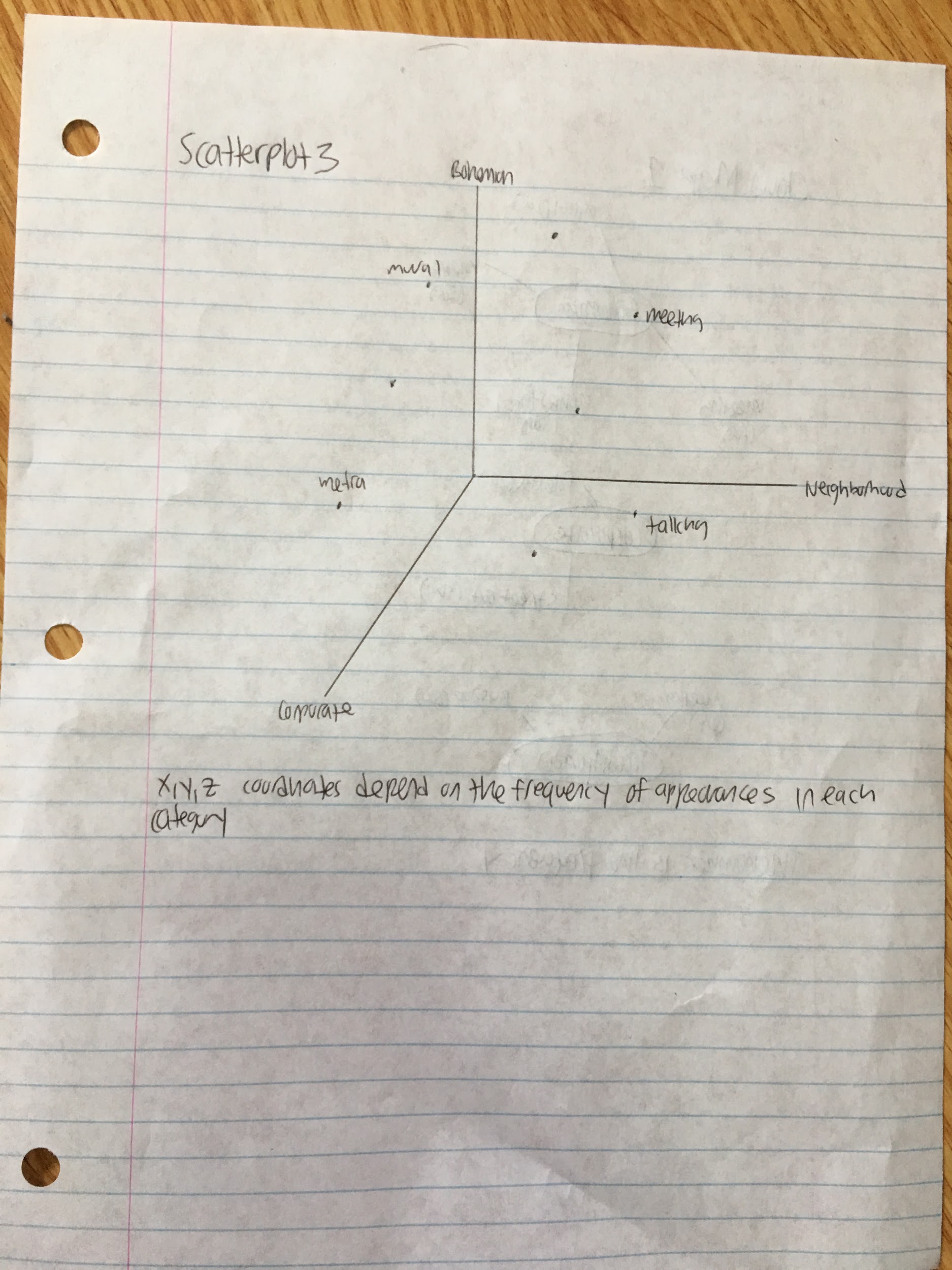
**Scatterplot 2:**

****

Cons:

* Larger circles may overlap
* X,y axis don’t have significant meaning

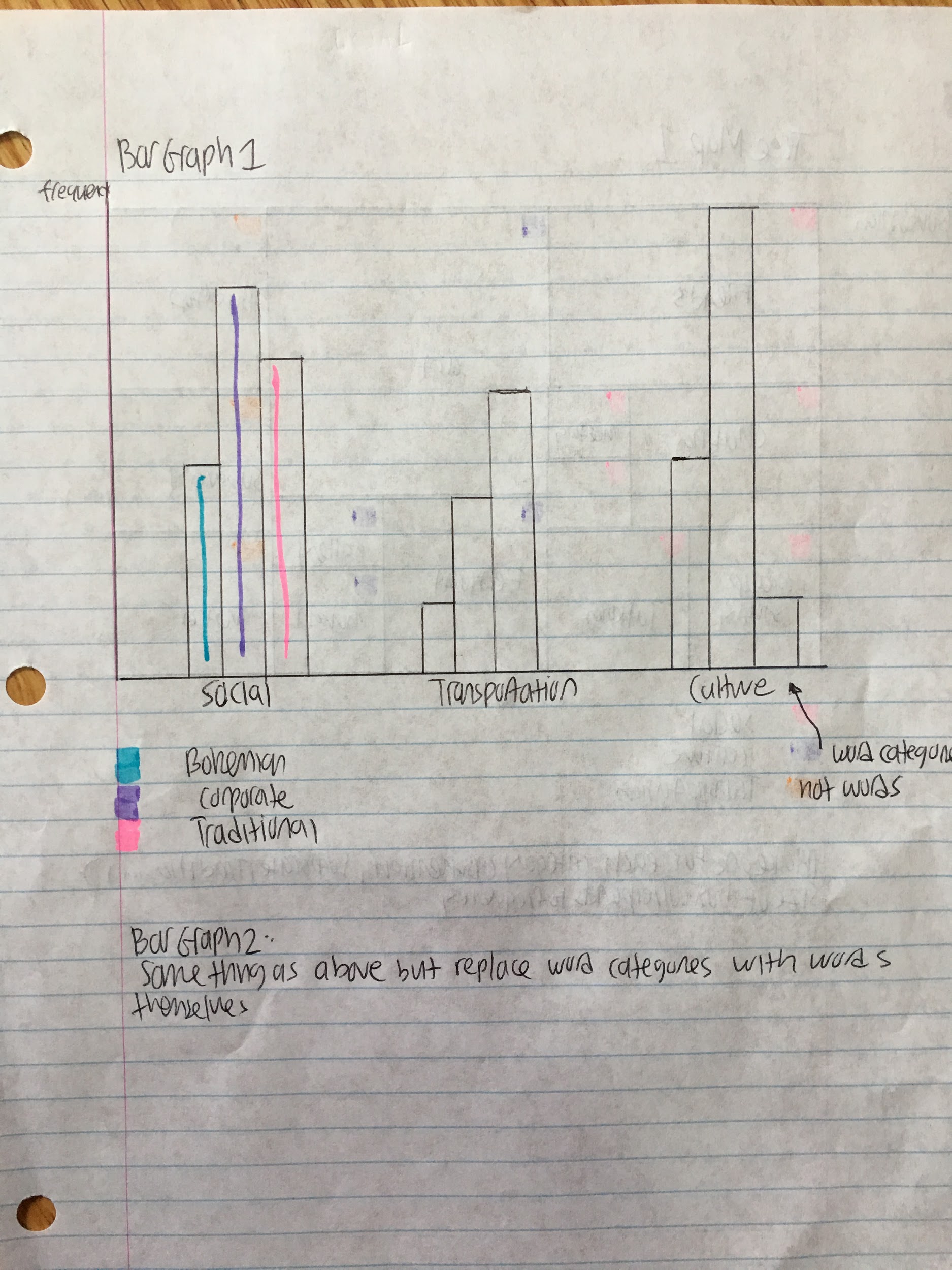
**Scatterplot 3:**



Cons:

* 3-dim on 2-dim paper is problematic (hard to see where the points really are)
* No logical grouping of words

**Bar graph 1:**

****

Cons:

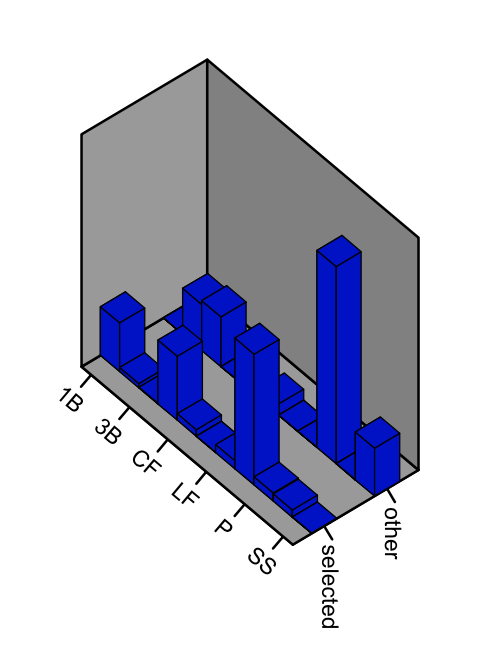
* Doesn’t say much about individual words

**Bar graph 2:**

Cons:

* Unscalable with many words
* Ordering of x axis may be a problem

**Bar graph 3:**



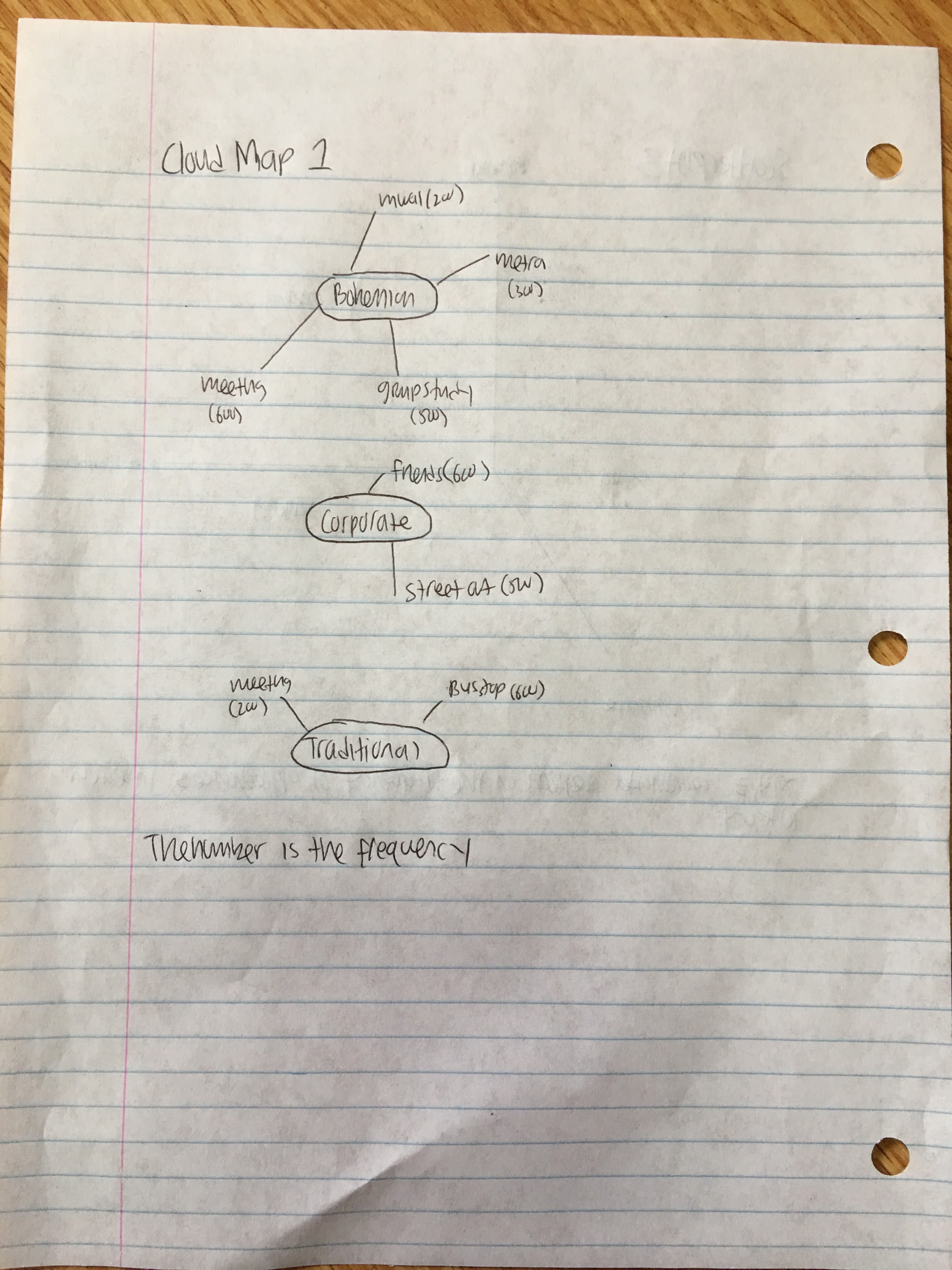
Source:

* page 223 from the textbook

Something like this, but where it says “selected”, “other”, replace it by “Bohemian”, “Corporate”, “Traditional”.

Replace the other axis with words.

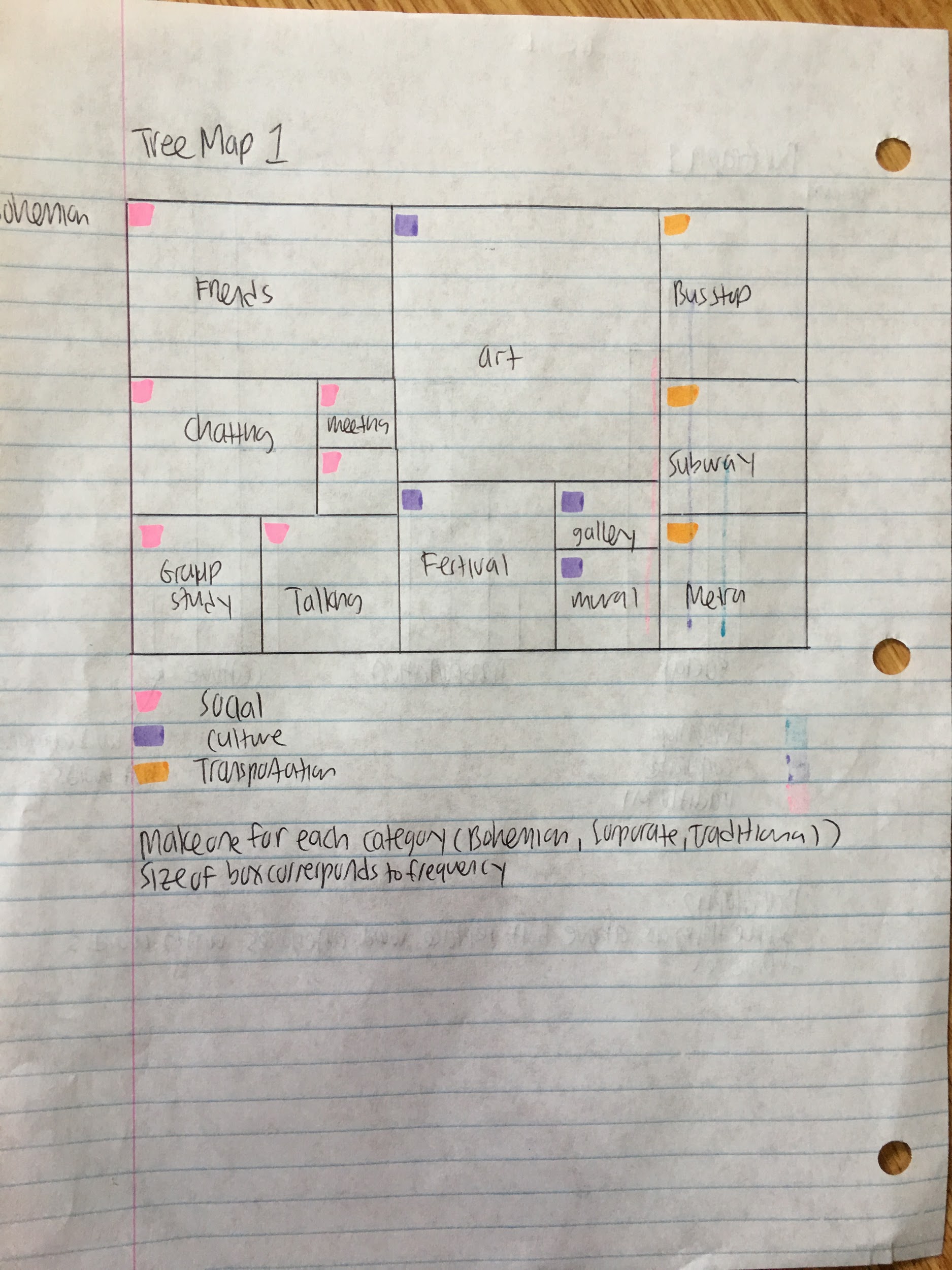
**Cloud Map 1:**



Cons:

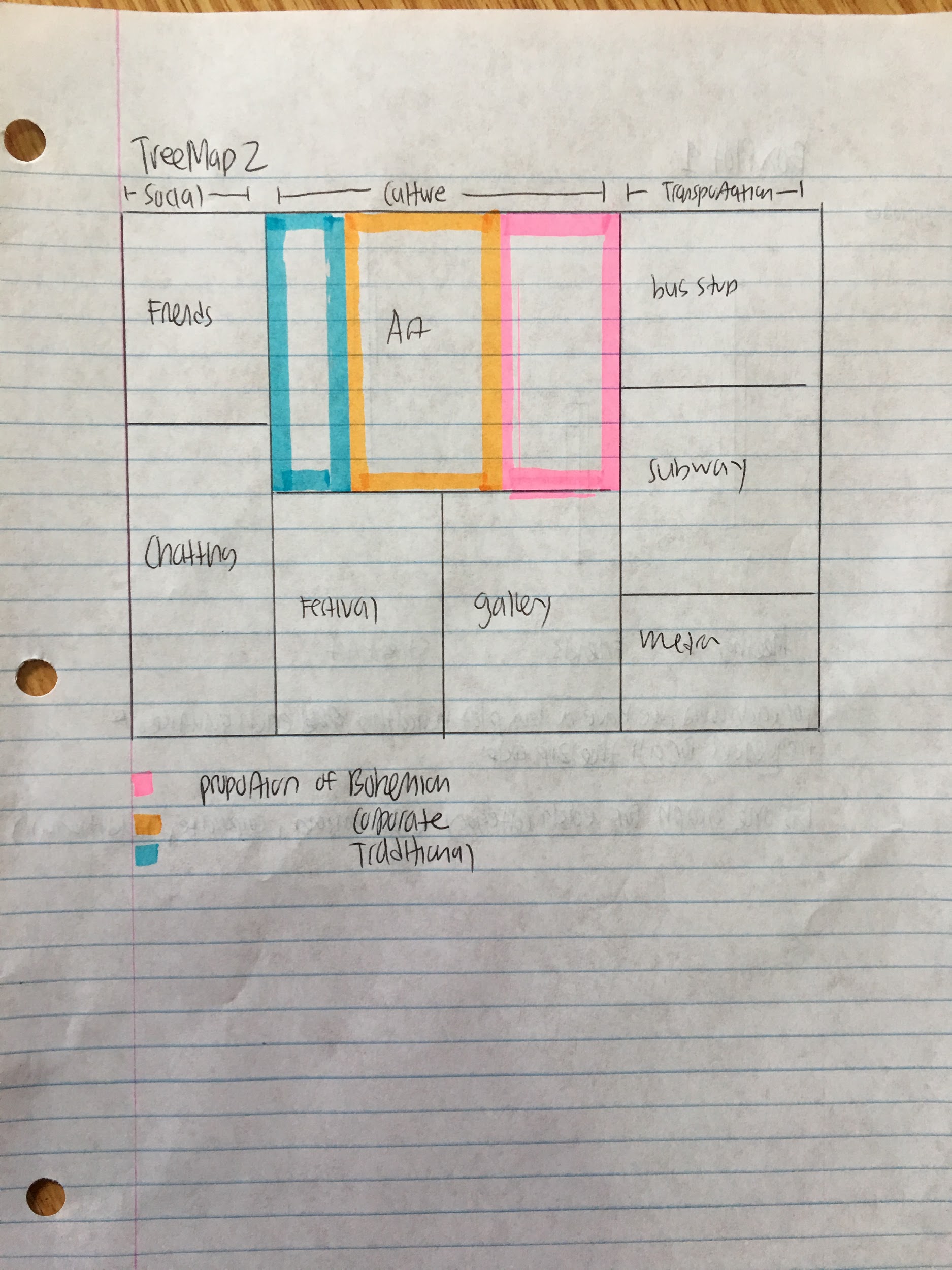
* very difficult to explicitly compare words from one category to another

**Tree Map 1:**

****

Cons:

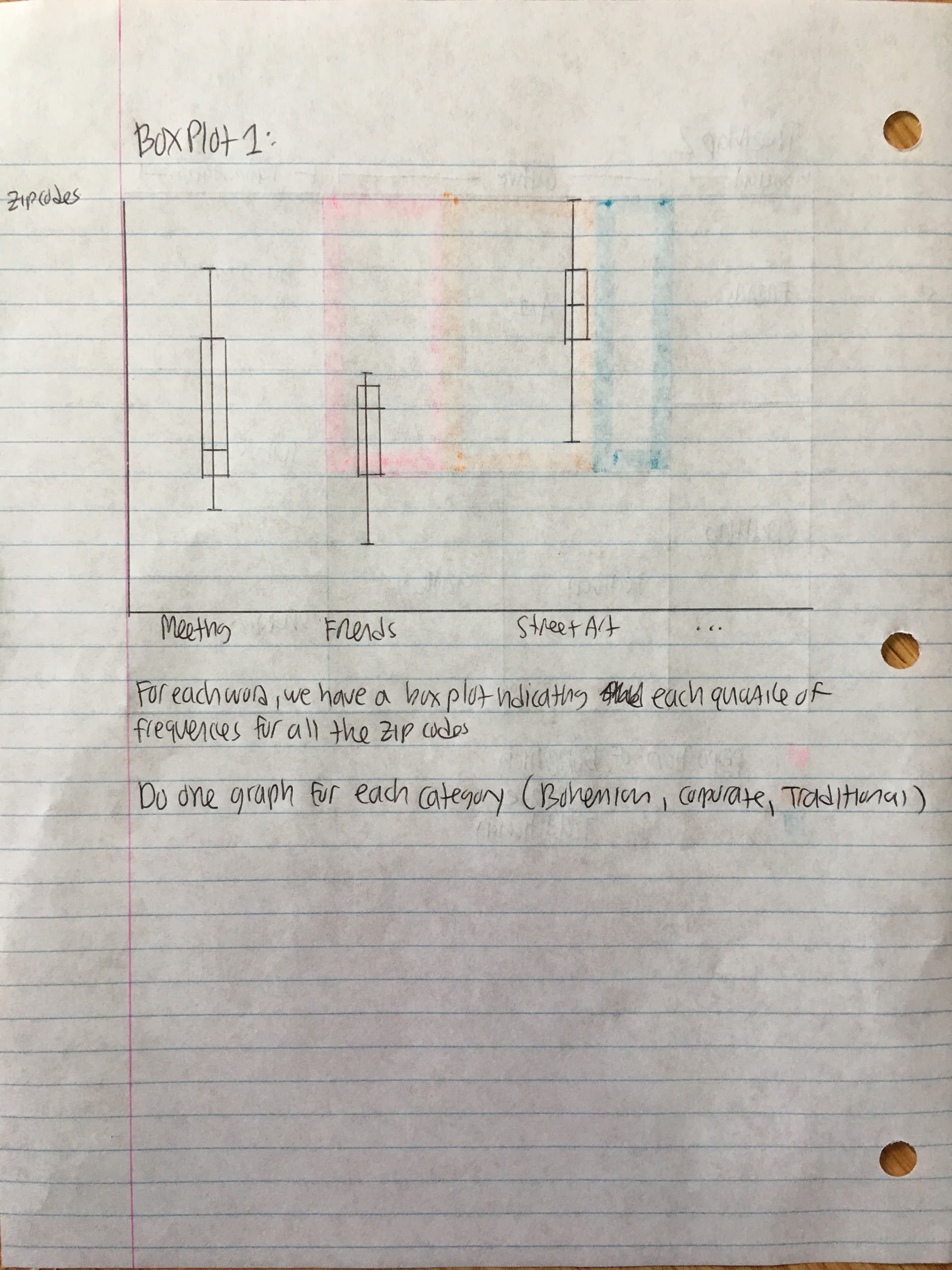
* Only really useful for proportions and not absolute amounts. The different categories may have different numbers of words, making comparisons more difficult.

**Tree Map 2:**

Cons:

* same as tree map 1

**Box Plot 1:**

****

Cons:

* not very scalable

**Plots that may be useful to consider:**

* Hierchal trees (see page 129 of the book)
* Box plots
* Panel scatterplots
* Linked plots

Things To Work On:

* Dynamic plot with time series, but still need the data and waiting for hyesun’s response

**Things to check out again:**

* Page 325 of the textbook

Google Common Words:

<https://github.com/JinLi711/google-10000-english>

Word Lists:

[Word Lists In Github](https://github.com/JinLi711/FAUI-Data-Scraping/tree/master/Yelp/NLP%20for%20Yelp/Word_Lists)

Multilevel Modeling

* [Wiki](https://en.wikipedia.org/wiki/Multilevel_model)
* Multilevel modeling is like linear regressions, but it is extended to different hierarchies.
* Attributes that we are comparing can be nested in other attributes that we are comparing. For example, we can be comparing political engagement under the hierarchy of geographical locations.
* We use this method when we decide that a current model does not contain independent results, we want to compare groupings

Geographically Weighted Regression

* [Spatial Analysis](https://en.wikipedia.org/wiki/Spatial_analysis#Spatial_regression)
* [A spatial](https://www.mailman.columbia.edu/research/population-health-methods/geographically-weighted-regression) analysis technique that models local relationships between variables and outcomes
* Is basically ordinary least square regression (OLS) but extended to encompass locality (meaning it creates an OLS for each location

Blalock Paper Notes

* Pictoral Models uses arrows and boxes to imply causation
* Variables in boxes and lines were labeled with either (+) or (-)
* I can’t really tell why some lines are dotted while other lines are dark (maybe darker lines imply higher correlation?)
* Arrows can be multi-directional (one box can connect to multiple boxes)
* We can even have loops
* Can label the difference between the main theory and the auxiliary theory

Ideas About How To Interpret Text Language (and what I can do with NLTK)

* Think about how words are grouped together (split sentences into bigrams).
* Look at uncommon words, not just common ones
* Create frequency tables for different hierarchies (frequency under zip codes, businesses, individual reviews)
* Think about different ways to split the string of texts. By sentences? Paragraphs?
* How can I interpret ambiguity in the text? For example, how do I deal with pronouns?
* How can I normalize text? And what are the drawbacks of doing so? Lowercase all the text (but ALL CAPS indicate emotions). Convert all words to the same tense (flew to fly). Remove plurals (cats to cat).
* Tag parts of speech (but not sure how this would help yet)
* Convert words that have the same meaning to the same word (like happy and joyful to happy)
* How to deal with numbers?

Research Questions:

* How do cafes create a scene?
* How can we know what “scene” is in the cafe?
* What values do certain cafes in an area display (like creativity, cosmopolitanism)?
* How can we use that to relate to politics?
* How can I add on to or contradict what has been said in the scenes book?

More Specific Questions:

* Is there a way to distinguish which cafes/bars are more conservative/ liberal?
* How does time intervals of activity (cafes in afternoon vs bars at night) affect how people voice their opinions?
* How can I know what central themes are in certain businesses? (feminism, egalitarianism, peace, gay rights)
* How can location (proximity) of cafes and bars to cultural places affect political expression?
* How can I find cleavages (age, race, gender, religion, class) inside different bars/cafes based on Yelp data?
* Are there connections between what people decide to eat and their political agendas?

Quotes From Scenes Book:

These quotes are for ideas I want to address using the Yelp data. Note that page number is according to PDF.

“Ambitious political programs have emerged around the world, with the character of the scene sometimes taking center stage as a topic of political controversy and target of public policy.” (page 177, Scenes)

Major Claim: “Scenes grow more politically salient in general with (1) the rise of culture, (2) the rediscovery of the urbane, and (3) the new political culture.” (177)

“Democratic votes were correlated with more urbane dimensions like transgression and self-expression; Republican votes, with more communitarian dimensions, like neighborliness and tradition.” (178)

“Styles of political leaders and urban governance have in some cities dramatically changed, adding new cultural and aesthetic sensitivity to their past repertoires” (179)

“Scenes accordingly become a topic of political contestation and a source of political authority” (179)

“Scenes provide cues about the character of a place, which some people find welcoming and others find alien and strange, sorting themselves accordingly.” (182)

“New Social Movements Are Typically Located in Dense, Walkable Areas with Self- Expressive Scenes and Many Artists” (195)

“Where do NSMs thrive? They are usually present in high rent, high crime counties, and there are more of them in Democratic counties. Neighborhoods with any (and many) NSMs are usually in dense, lower rent zip codes with strong cultural employment concentrations, nonwhite residents, and college graduates.” (198)

“When walking and self-expression come together, the result is quite likely to be organizations advocating for human rights, social justice, and the environment.” (202)

“self-expressive scene, for instance, likely indicates a broadly culturally liberal environment” (206)

Scenes have been rising.

Politicians are supporting more arts and scenes. They want to create buzz.

What I Want to Measure (For Both Bars and Cafes)

This is going to based on what the quotes say and what I think might help.

* Conservatism vs liberalism
* Democratic vs Republican
* Rural vs Urban
* Political agendas (feminism, egalitarianism, peace, gay rights)
* Social cleavages (age, race, gender, religion, class)
* The sense of scene (people’s identifications with a particular place, or multiple places)
* “Buzz”

How I Am Going to Measure It

Below, I will indicate the data available on Yelp, and possible ways of using them.

Explicit Political Comments:

Explicit political comments are very rare, but that does not mean that they do not exist. I would need to find some before I decide how to analyze it. Still thinking about this.

The way people write:

It is possible to draw a connection between how reviewers structure their reviews and the type of person they are. And we can look at the connection between the types of people who reviewed and the businesses they review in. For example, maybe a certain cafe have reviewers who sound more educated than the reviewers in other cafes. And we can draw connections between cafes and politics. There are many ways to figure out attributes about people based on what they say. For example, look at how often they refer to themselves in the first person singular (using I), first-person plural (we), or third person. People who use “I” are more self-centered and individualistic. Reviewers who use “we” imply that they are visiting a bar with friends. We can also look at how educated a person sounds. Some people use very technical, possibly pretentious, words to describe their meal, while others use slang. Others are very careful about their punctuation and grammar while others aren’t.

Or we can directly relate the words that users use to the business itself. For example, look at punctuation, especially (!). (!) hints at excitement, passion. Other ways reviewers display their feelings about a business is through the use of ALL CAPS.

Connection Between Business Information and the Cafe’s culture:

Look at the price range, hours opened, days opened, and specific business information. Price range can indicate class demographics. Hours opened can indicate age demographics (businesses that stay open late until at night are geared more for younger people) and culture (night life culture).

|  |  |
| --- | --- |
| **More Business Information Labels (note that this is not comprehensive)** | **What this Can Indicate** |
| Takes Reservations |  |
| Accepts Credit Cards |  |
| Accepts Apple Pay |  |
| Parking |  |
| Bike Parking |  |
| Good for Groups | Social gatherings for people to discuss ideas |
| Noise Level |  |
| Alcohol |  |
| Outdoor Seating |  |
| Has TV |  |
| Take-out |  |
| Attire | More formal attire can indicate that a certain place is for those with higher classes |
| Good for Kids |  |
| Ambience |  |
| Wi-Fi | More millenials |
| Good for Working |  |
|  |  |
|  |  |
|  |  |
|  |  |

Special Features of Yelp:

There are features that users can use (but are optional). This includes the number of check-ins (number of times that Yelp reviewer says to have visited the business) and the number of people who voted for a specific review (users can vote useful, funny, or cool). Possibly, reviews with more votes should be weighted more because more people agree with that review. We can also think about why a certain person would check in multiple times.

[Note that some reviews are filtered by Yelp](https://www.youtube.com/watch?time_continue=71&v=PniMEnM89iY). Basically, only ¼ of the reviews are not recommended (this does not mean one can not see the reviews, it is just not in plain sight). The reviews are filtered by Yelp’s algorithms, and their intent is to filter fake reviews, rants, and unknown people without credibility. The reviews that are filtered do not affect review count or the average stars of a business. The algorithm is also constantly running, so a review that has been filtered can be unfiltered, and visa versa.

It is possible to comment on a Yelp review. Many business owners do this to address issues that reviewers have with a business (many times, the business owner apologizes and promises better service).

Who is reviewing:

We can look at information relating to who is reviewing. Many of the people who review in cafes and bars have a lot of friends in Yelp (over 100) and have reviewed other things. For example, look at [Christopher](https://www.yelp.com/user_details?userid=hT2Mw5m_SXBDVBuvA3KY-w) V. , who I found since he reviewed [Plein](https://www.yelp.com/biz/plein-air-cafe-and-eatery-chicago-2?osq=cafe) Air Cafe & Eatery, a cafe in zip code 60637. Looking at his profile, he seems to love reviewing bars, cafes, restaurants. These reviewers can be a great asset for deciding how to connect different businesses.

Some reviewers take pictures of things other than food (like a painting on the wall). Other reviewers also comment about things other than food (like how a cafe feels like the 80’s.)

Cafes vs. Bars

Look at the time ranges of reviews. Are there certain time periods where the number of new reviews increases/ decreases? This may correspond to times of major political movements, like an election of a new council or president.

Compare words that are tied to emotions.

Look into other reasons why people say they attend a bar or cafe other than the food. For example, what do people say about their purpose of being in a bar or cafe? What emotions do they feel (relaxed, at ease, “high”, happy)?

What cultures do bars/cafes provide? What do the cafe/bars express in terms of art, music, etc. How does this say anything about politics?

Look at the number of new bars/cafes that appear every year. This may indicate something about the new political atmosphere.

Are there any indications on political issues (like acceptance of Africans, gays, feminists,)

What are cafes and bars surrounded by (what restaurants, other bars, or museums, etc surround this location)?

These rhetorical questions will be answered when I can get a larger dataset.

Problems That I’ve Encountered or Expect to Encounter

How to normalize everything:

Dealing with plurals: (cats vs cat)

Dealing with contractions (I’ve vs I have)

Spelling errors

What information am I losing when normalizing?

How to deal with word groups:

If we only consider words by themselves, “I do not like this place” would be interpreted in the same way as “I like this place.”

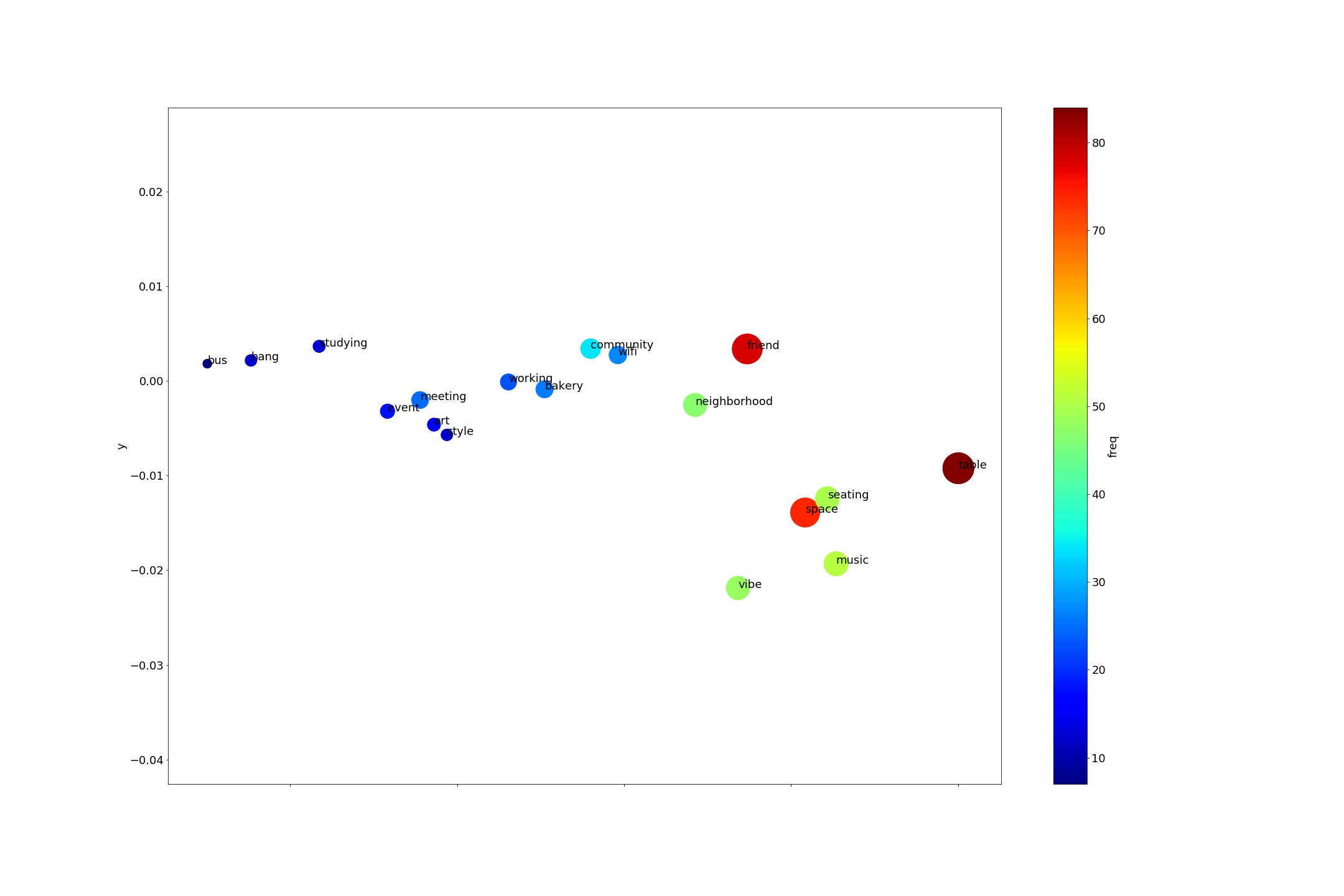
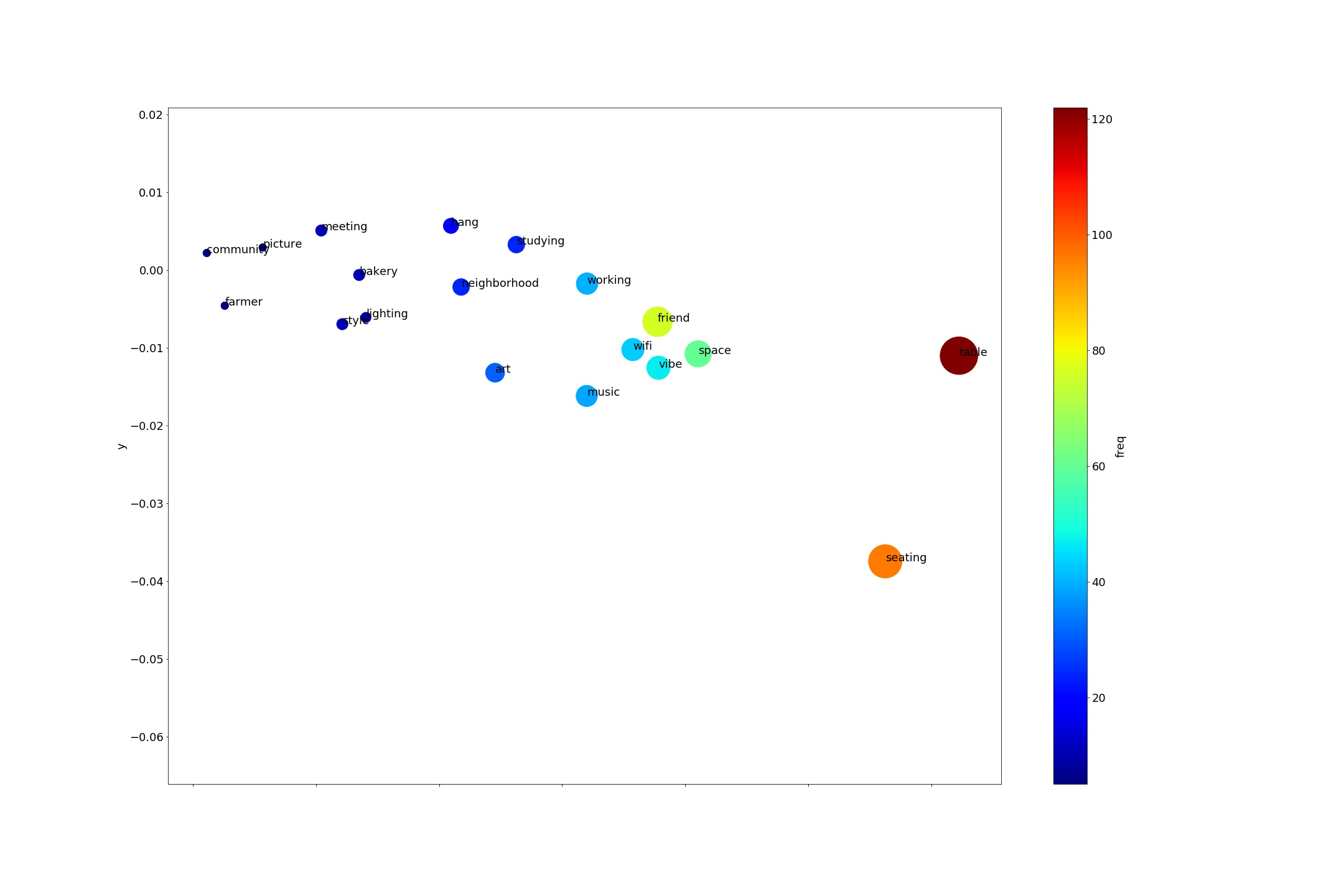
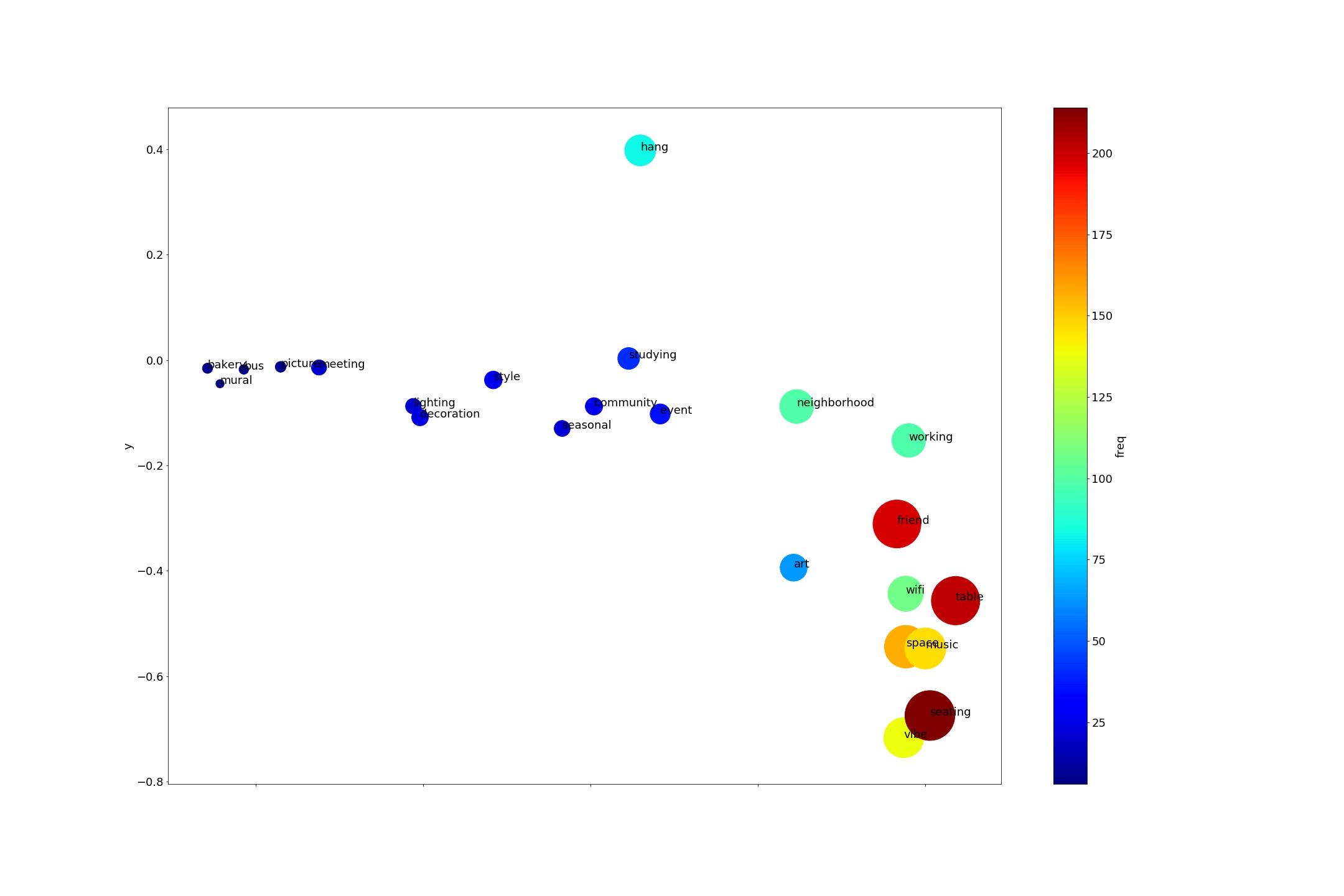
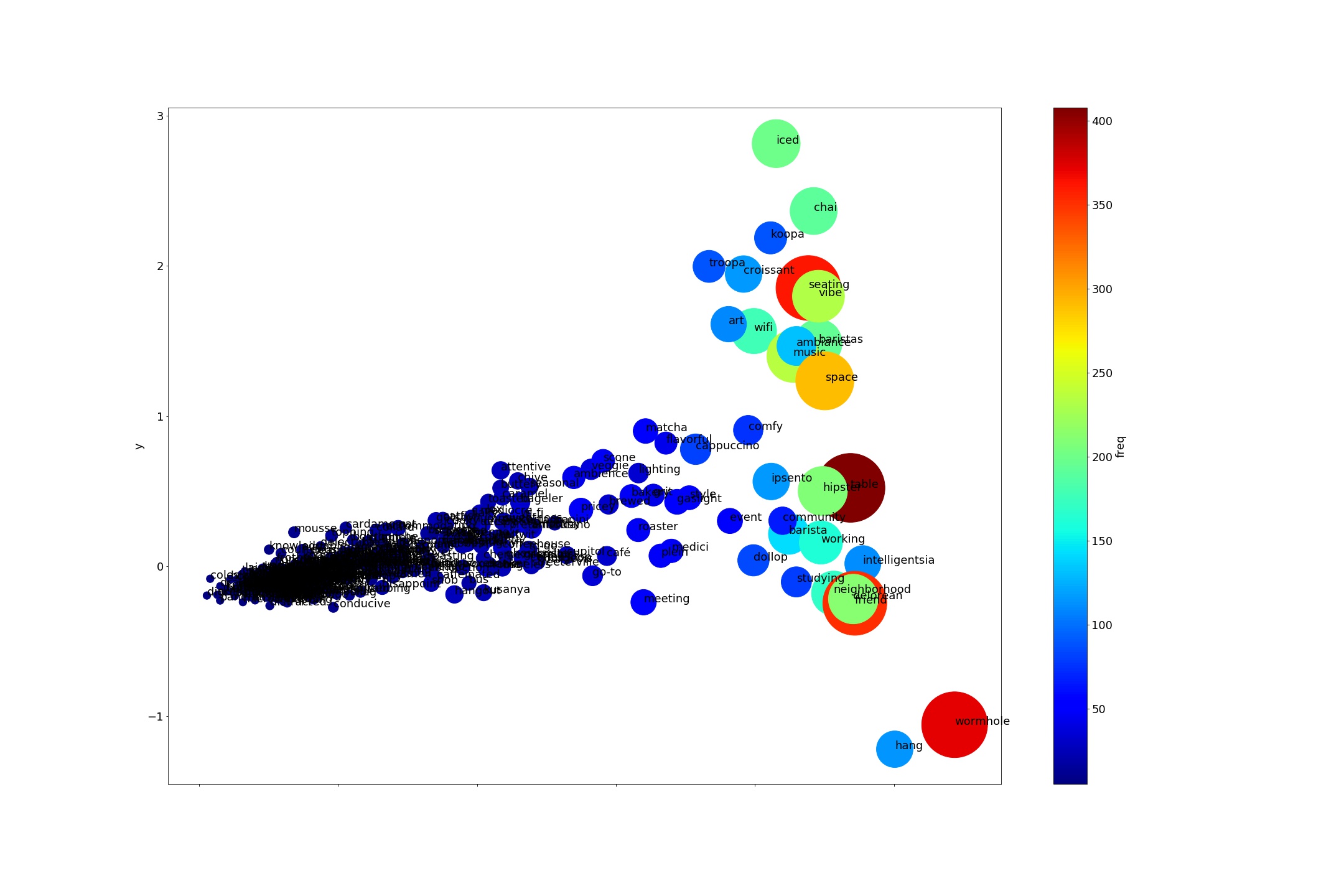
Embedding

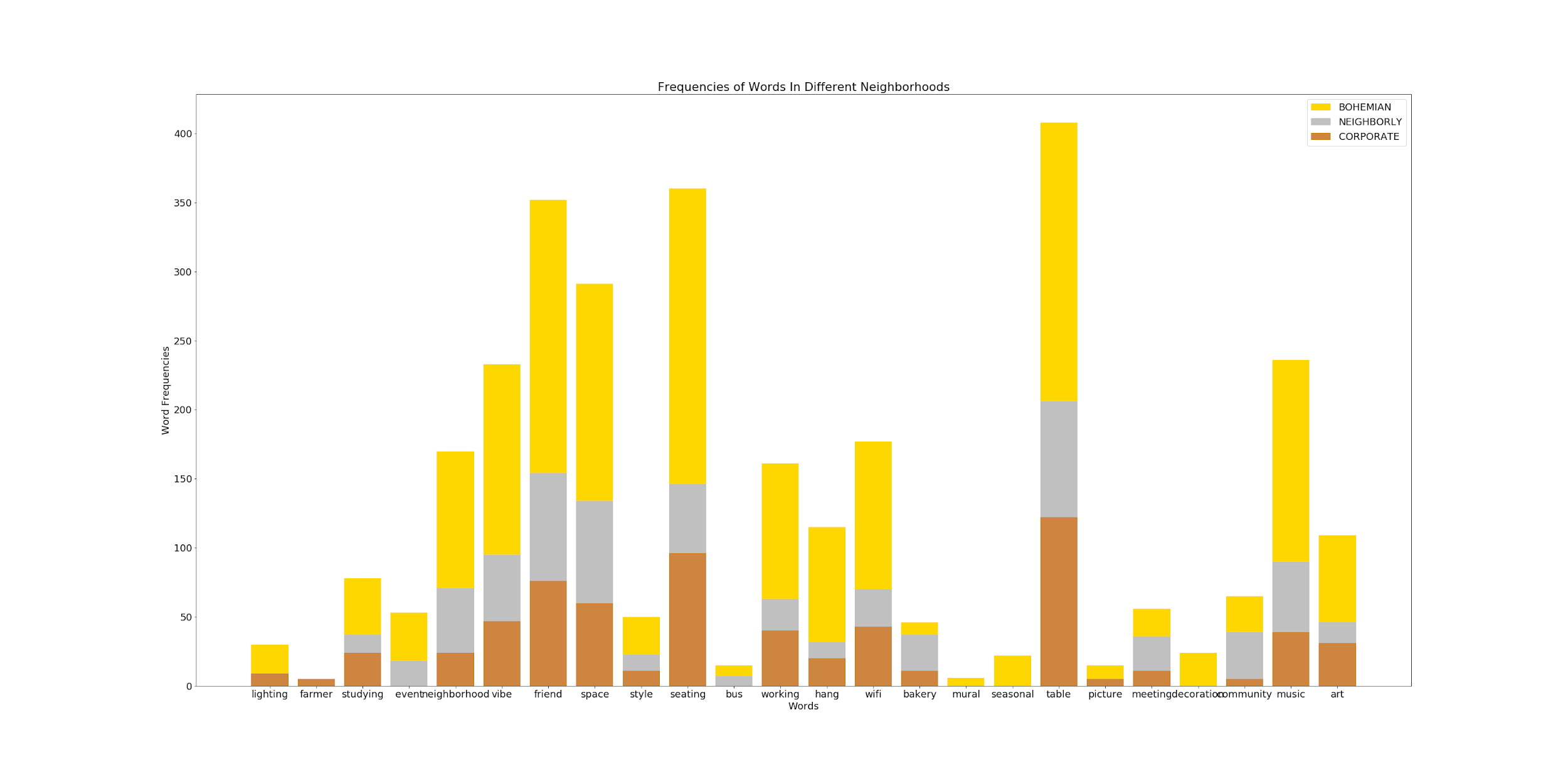
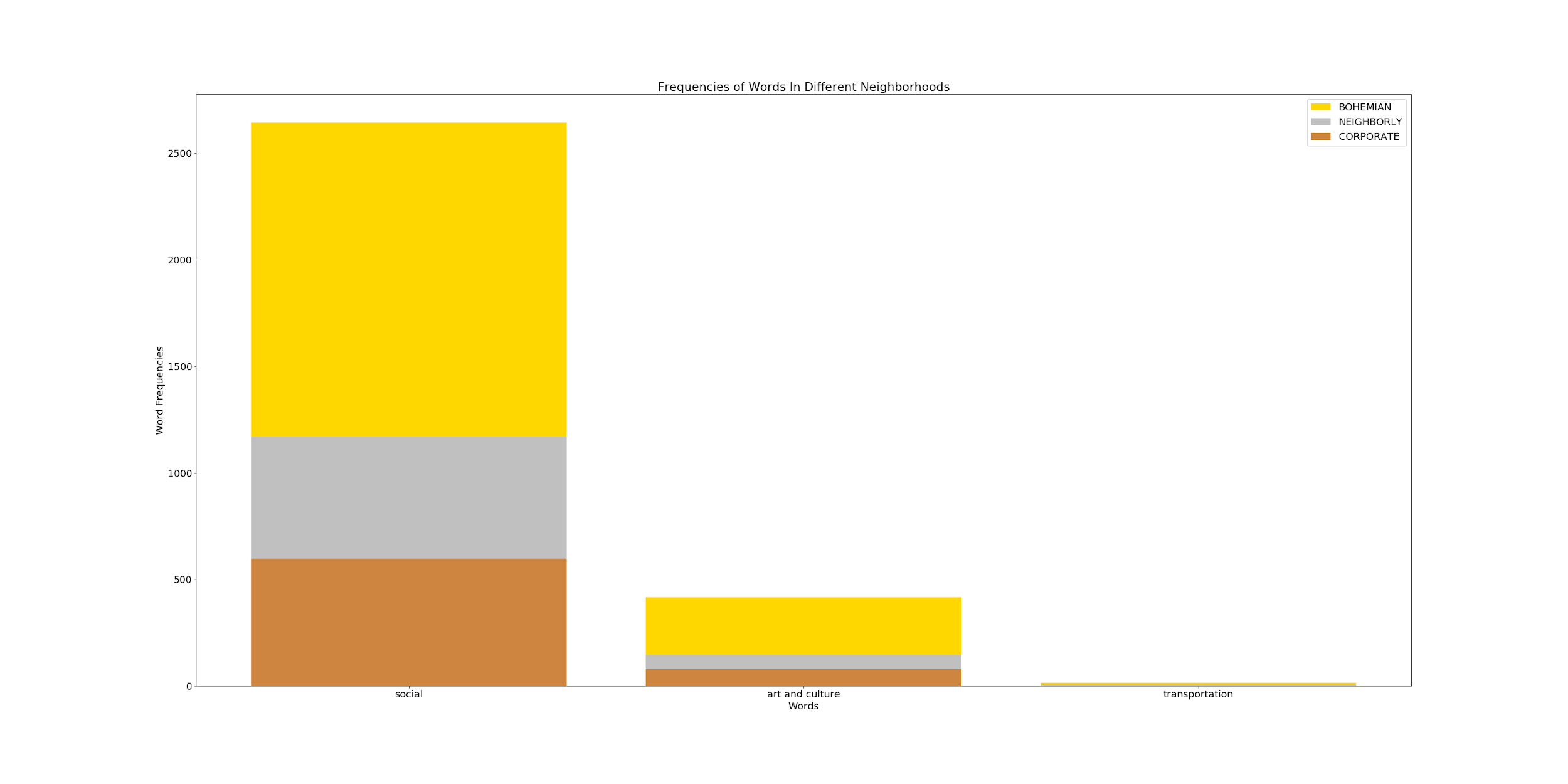
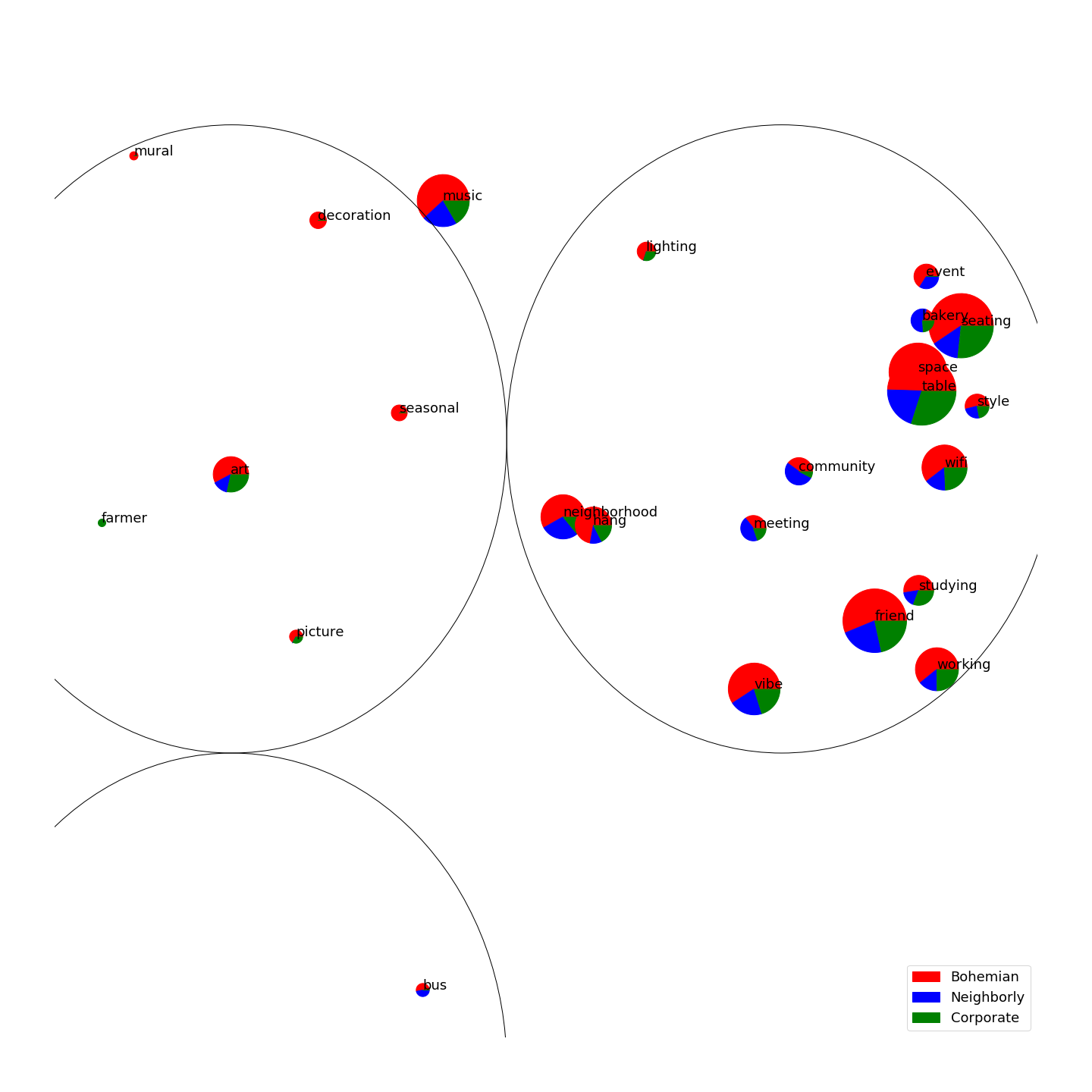
Possible Ideas for Analysis

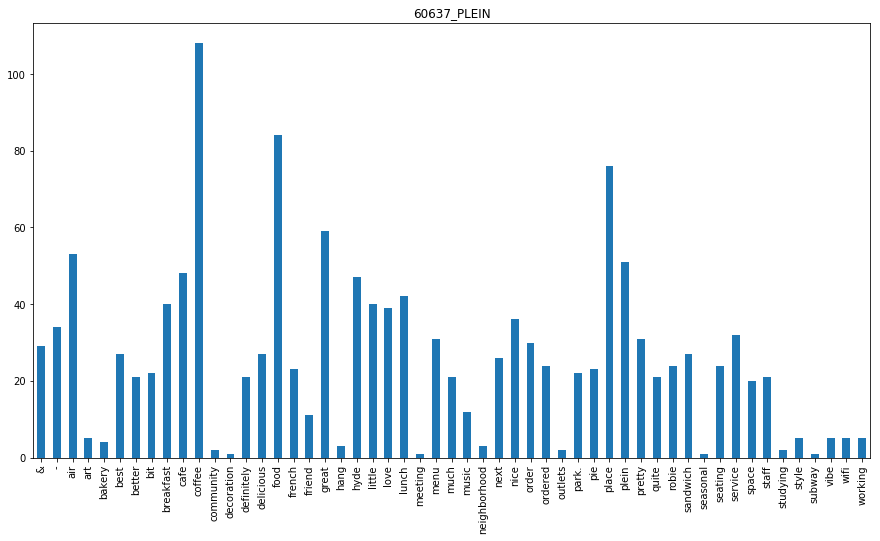
Compare clusters of words to the standard cluster (the one made by google).

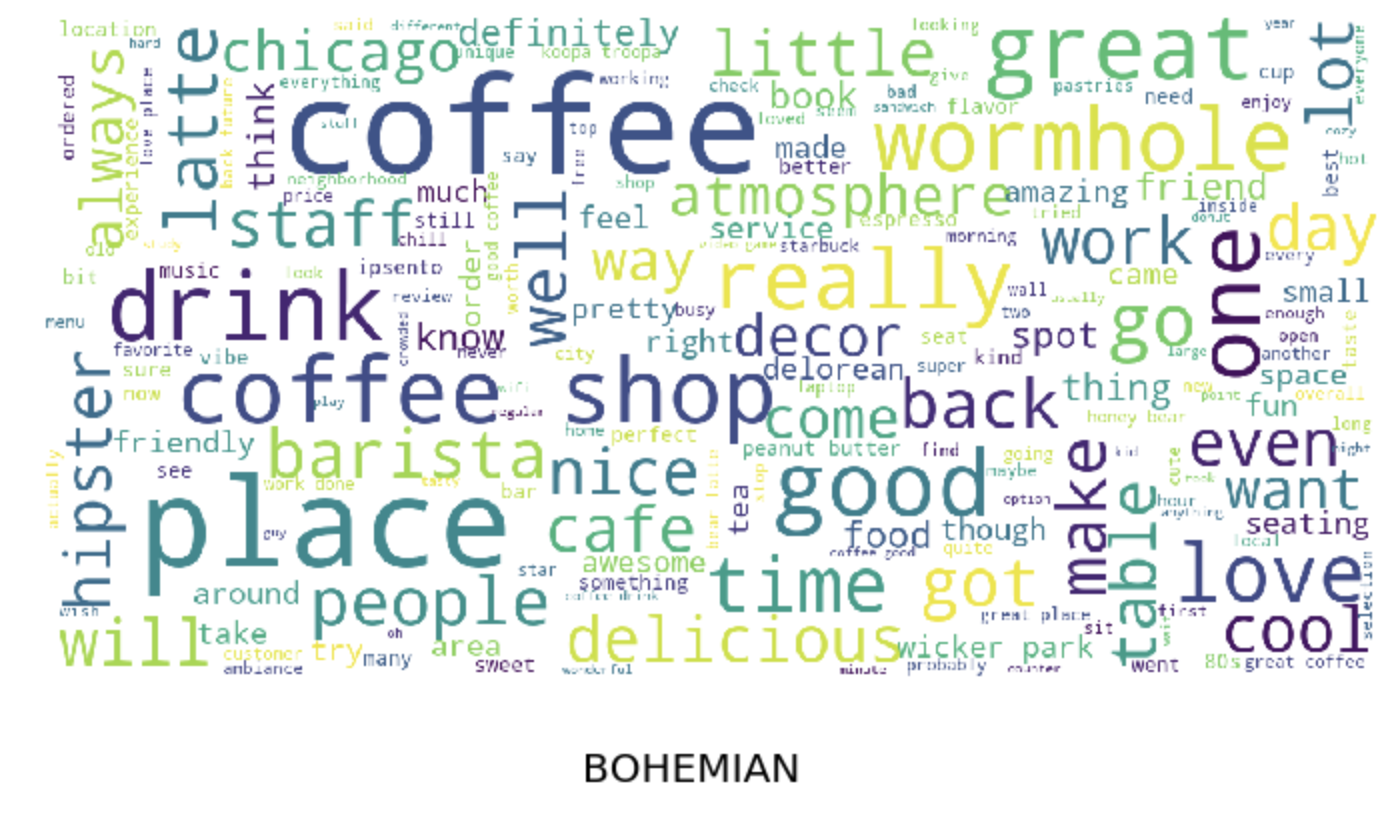
Plots Generated:

Examples of What I’ve Created (note that a lot of these are just drafts, their quality will improve as I start to finalize):









Yelp Text To Data

**Objective:**

Measure word count frequencies under certain variables (which include geographical location, business type)

**NOTE:** This write up will exclude in-depth details regarding Python code. Check out my github if interested (I documented the code). Also, I tried to be quick to the point for my description; just ask me and I’ll try to explain further.

**Steps:**

**Step 1: Obtain the data.**

We expect Dan Silver to provide a large amount of web scraped data for Yelp soon, but we don’t have that right now.

So for now, I use my own custom web scraper to gather pieces of data just to get started.

**NOTE:** Right now, I am not worrying about sampling biases and other statistical issues. I want to build a flexible pipeline that achieves my objective. After I finish, I will make adjustments.

Substeps:

1. Go to [Yelp.com](https://www.yelp.com/), and on the search tab, search “cafe” in the “Find” tab and “ZIPCODE” in the “near” tab, where ZIPCODE is replaced the zip code that we want to analyze.
   1. **NOTE:** the businesses that we get from the search result are not necessarily in the Zip Code, but may just be near the ZIPCODE.
   2. **SOLUTIONS:** the future scraped data should have each business be labeled by its zip code, and we can do some necessary filtering.
2. Scrape the data.
3. Download results as “ZIPCODE\_NAMEOFBUSINESS.csv”

See:

[github](https://github.com/JinLi711/FAUI-Data-Scraping/tree/master/Yelp/NLP%20for%20Yelp) for an example of the content.

**Step 2: Extract relevant data from raw data.**

The scraper picks up a lot of data that we do not need at this point, but we may need for later. Right now, we are only concerned with the reviews.

Substeps:

1. Extract only the reviews from the scraped data from each csv file.
2. Combine the reviews based on zip codes.
   1. For example, if we have 60603\_FAIRGROUNDS.csv, 60603\_GODDESS.csv, and 60626\_CHARMERS.csv, we combine the reviews of 60603\_FAIRGROUNDS.csv with 60603\_GODDESS.csv.

**Step 3:Preprocess the text.**

Take the data, and normalize the words to make them suitable for comparison. Also count the word frequencies

Example:

NOTE: [“X”, “Y”, “Z”] indicates a list, “XYZ” indicates a string.

“This food is amazing! Why did it only cost me 5 dollars?”

Substeps:

1. Lowercase all the words.
   1. “this food is amazing! why did it only cost me 5 dollars?”
2. Replace numbers with [NUM].
   1. “this food is amazing! why did it only cost me [NUM] dollars?”
3. Tokenize into sentences.
   1. [ [“this”, “food”, “is”, “amazing!”], [“why”, “did”, “it”, “only”, “cost”, “me”, “[NUM]”, “dollars?”] ]
4. Replace punctuations and other non-letter characters: ()!@%^&-+\$.,?\*"#
   1. [ [“this”, “food”, “is”, “amazing”], [“why”, “did”, “it”, “only”, “cost”, “me”, “[NUM]”, “dollars”] ]
5. Lemmatize the words.
   1. [ [“this”, “food”, “is”, “amazing”], [“why”, “did”, “it”, “only”, “cost”, “me”, “[NUM]”, “dollar”] ]
   2. The only change here is from dollars to dollar. Lemmatization just refers to removing tenses. For example, “cars” becomes “car” and “flew” becomes “fly”.
6. Count word frequencies.
   1. In this case, the frequency for each word would be one.

**NOTE:** Text is very unstructured data, and we are losing a lot of information by processing it.

Examples:

1. Say we want to search for keyword “art” and its frequency
   1. Sentence: “This art is amazing. It is very inspiring.”
   2. Note that the word “it” references “art”. But our current method will only result in a word count of “art” as 1, instead of 2.
2. Misspelled words

**Step 4: Create an Embedding.**

This is probably the trickiest step, as it is going to be the hardest to understand.

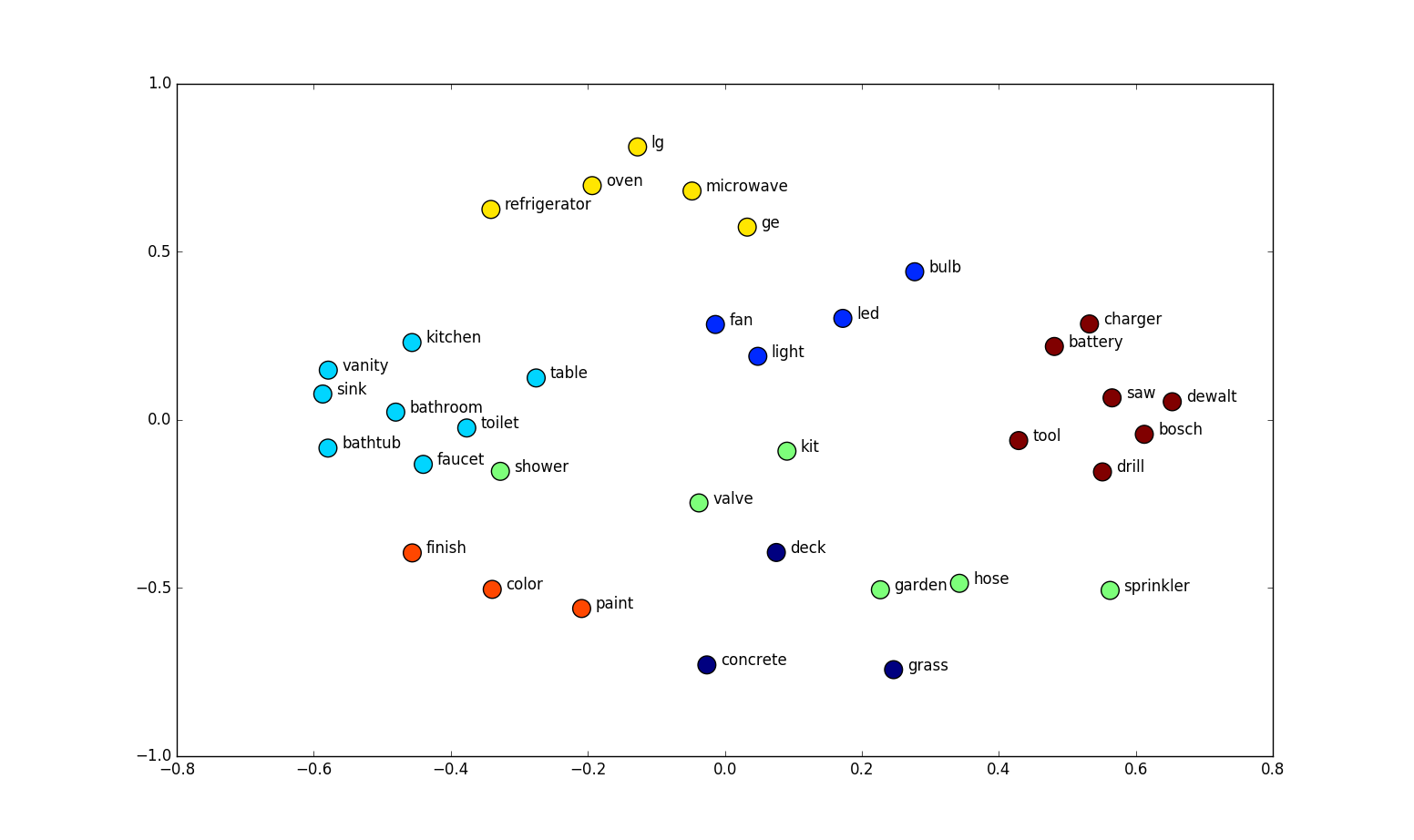
What is word embedding?:

It is essentially a way to map words to vectors in some high dimensional space. The vectors are not static; they can move around in space, but the word tied to that vector remains tied.

Why am I creating an embedding?:

Word embedding is a way to quantify words in the sense of relatedness. If done right, words that have similar meaning will have vectors that are close together. For example, the vector representation of “cheese” will be closer to the vector representation of “butter” than it would to the vector representation of “car”. Using this process, we can plot words on a scatter plot, and viewers can more easily see clustering of words.

Example:



Understanding this plot is not particularly difficult, but people do seem to have a hard time understanding that the x and y are the coordinates of the word’s vector representation. In terms of presentation, the x and y coordinate means nothing, what we want to focus on is how the words are grouped together. As is the picture, “charger”, “battery”, “tool” are clustered together.

Why not a bar graph?:

* Bar graphs are ugly to look at and hard to make any meaning out of when the x column (the categories) becomes large. We have about 30 words that are interesting to compare, and possibly even more later on. You’ll just end up with a huge bar graph that can barely fit in a page.
* Also, I dislike the fact that there is no intuition behind the ordering of the x-axis, and ordering of anything becomes more important the more there is to order.

**NOTE:** I used the [GENSIM](https://radimrehurek.com/gensim/models/word2vec.html) Word2Vec model for embedding words. I choose this because it is fast, easy to use, and reliable.

How It Works:

1. Start with assigning each word to a vector. This can be done randomly, or we can use a pre-trained model that already assigned each word to a vector.
2. Move the vector around.
   1. This is done by machine learning/ deep learning.
   2. Example sentence: “I love to pet my dog.”
      1. Take one word at a time, say the word “pet” (we’ll call this the center word).
      2. The nearby words, aka context words, would be the words around “pet”. They would be “to” and “my”. The window size, in this case, is 1, but we can change this to 2, and the new context words would be “love”, “to”, “my”, and “dog”
      3. Calculate the probability of a context word given a centre word or visa versa.
      4. Based on this probability, adjust the word vectors to maximize the probability.

Two models:

* Continuous Bag-of-Words model (CBOW)
  + predict the center word based on context words
* Skip-Gram model.
  + Predict context words based on the center word.

**NOTE:** the choice of model doesn’t really matter that much right now. I included this because it may end mattering when we have a lot more data.

My parameters for word embedding:

* CBOW
* Minimum word count of 5 (so ignore all words less than the word count)

**Step 5: Dimension Reduction.**

Word embedding usually maps words to very large dimensions (I did 100 dimensions). More dimensions make computations slower, but produces better word embedding results. But we can’t visualize 100 dimensions, so we have to reduce the dimensions to 3 or 2.

**NOTE:** Dimension reduction obviously decreases information. We want to reduce the dimension while preserving as much variance as possible (as much infomation as possible)

Algorithms for dimension reduction:

* Principled Component Analysis (PCA)
  + This is the most common, and it runs pretty quickly.
  + It basically identifies the hyperplane that lies closest to the data, and then it projects the data onto it.
* Locally Linear Embedding (LLE)
  + Measures how each point linearly relates to the closest neighbors and looks for a low-dimension representation of the points where these local relationships are best preserved.
* t-Distributed Stochastic Neighbor Embedding (t-SNE)
  + Best for visual representation, but runs slowly
  + Reduces dimensions by trying to keep similar points close and not similar points far from each other.

I chose PCA for now, but plans may change, so I included other options.

**Step 6: Filtering Words:**

We can now visually represent the words, but we have thousands of words to choose from (and that number is going to rise to tens of thousands when we have more data).

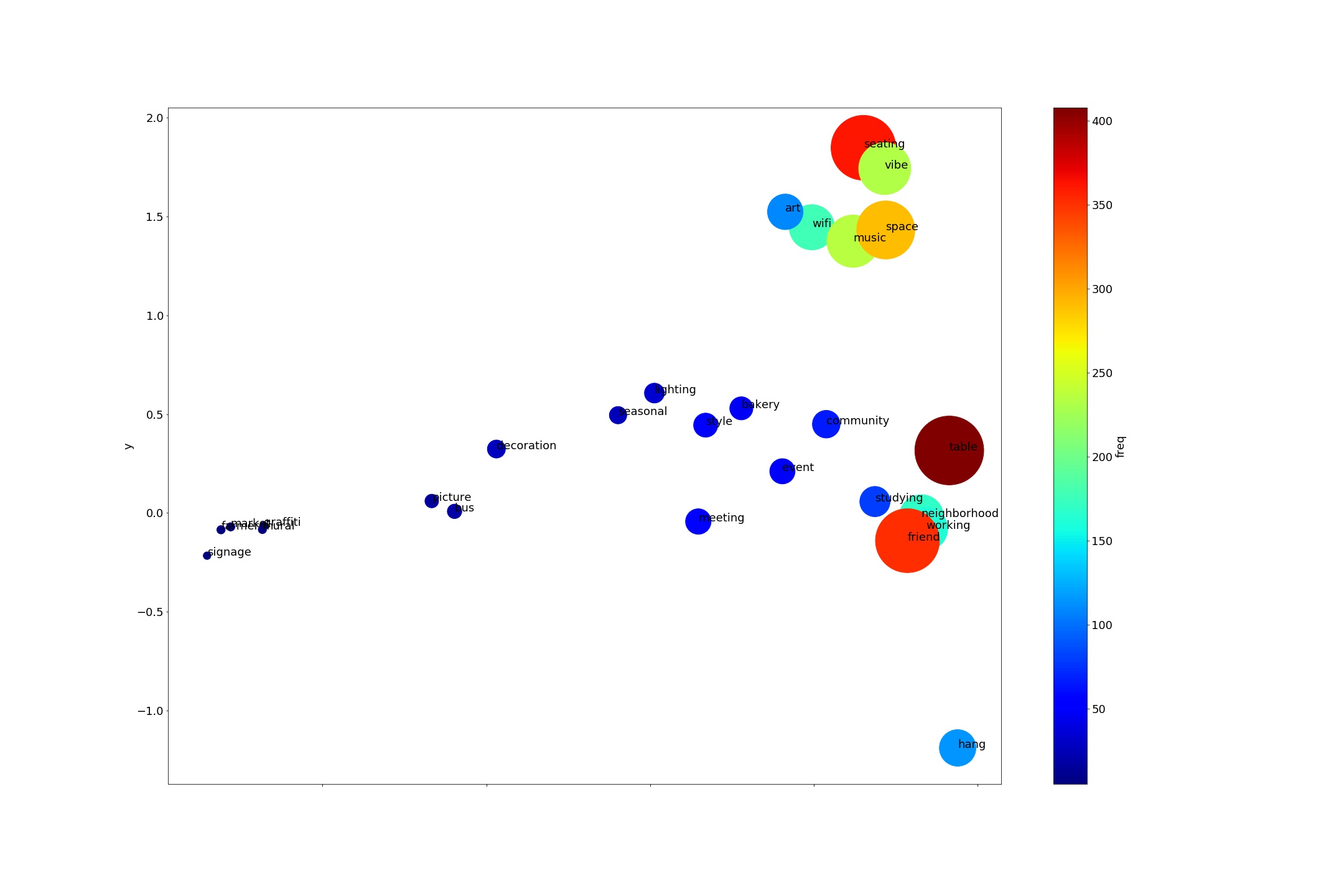
So, lets create some csv files for filtering (send me an email for the csv files, I’m not going to place them here because I’m continuously updating them).

Substeps:

1. Create a list of words to keep (immunity from filtering).
   1. bakery, decoration, style, seating, picture, lighting, community, quietness, music, space, table, outlets, wifi, vibe, event, mural, art, signage, metra, l station, bus, subway, graffiti, neighborhood, community, festival, event, farmer, market, flea  
      Seasonal, friend, studying, meeting, children, hang, dating, working
   2. This is just the list Hyesun made.
2. Use Google’s 20,000 most common words.
   1. I know we may be losing a lot of information by using this as a filter, but I don’t know any other option except hand filtering (which would take a long time)
   2. But if someone comes up with a better way, I’ll take this filter off.
3. Create my own filter list to account for words that are not interesting to consider and are not in the google list.
   1. This includes food items.

**Step 7: Create a scatter plot:**

Self explanatory. I’ll just put an example:



**Things to Work On/ Consider:**

Consider word density instead of absolute word count.

* Some places will just have a lot less words in the reviews.

Download Stanfrod GloVe

Combine the zip codes by type of neighborhood

* Should different zip codes be weighed differently?

Message With Ben and Michal

;

Facebook

Jin

Home

Michał and Ben

Participants: Michał Dzitko, Ben Picker and Jin Li

Ben Picker

Ok

Jan 4, 2019, 11:12 PM

Jin Li

Any time basically

Jan 4, 2019, 11:11 PM

Ben Picker

Whats ur availability Saturday?

Jan 4, 2019, 11:11 PM

Ben Picker

Um ok

Jan 4, 2019, 11:11 PM

Jin Li

You: 󰀀

Jan 4, 2019, 11:11 PM

Ben Picker

Fyi

Jan 4, 2019, 11:11 PM

Ben Picker

Also Michal goes by Mike

Jan 4, 2019, 11:11 PM

Jin Li

Yea

Jan 4, 2019, 11:10 PM

Ben Picker

Would tomorrow be better?

Jan 4, 2019, 11:10 PM

Ben Picker

But it sounds like ull be traveling

Jan 4, 2019, 11:10 PM

Ben Picker

So mike had talked about organizing a call sunday

Jan 4, 2019, 11:10 PM

Ben Picker

Yo jin

Jan 4, 2019, 11:10 PM

Ben Picker

Super

Jan 4, 2019, 10:30 PM

Jin Li

Great

Jan 4, 2019, 10:29 PM

Ben Picker

So first floor at 4:45

Jan 4, 2019, 10:29 PM

Ben Picker

Ok cool

Jan 4, 2019, 10:29 PM

Ben Picker

We dont need a room lol

Jan 4, 2019, 10:29 PM

Jin Li

That’s fine

Jan 4, 2019, 10:29 PM

Ben Picker

Cristina always wants a room

Jan 4, 2019, 10:29 PM

Ben Picker

Wtf lol

Jan 4, 2019, 10:29 PM

Ben Picker

Actually we can just do firsr floor

Jan 4, 2019, 10:29 PM

Jin Li

How do you book a room at the reg?

Jan 4, 2019, 10:28 PM

Ben Picker

I dont know if my cnet will let me

Jan 4, 2019, 10:27 PM

Ben Picker

Hmm actually can u do it

Jan 4, 2019, 10:27 PM

Jin Li

Sure

Jan 4, 2019, 10:26 PM

Ben Picker

I can book a room at the reg

Jan 4, 2019, 10:26 PM

Ben Picker

So u want to do like 4:45?

Jan 4, 2019, 10:26 PM

Ben Picker

Ok great

Jan 4, 2019, 10:26 PM

Ben Picker

Basically free after 3:30

Jan 4, 2019, 10:26 PM

Jin Li

I’m free after 4:30 basically

Jan 4, 2019, 10:26 PM

Ben Picker

Ok why dont u figure it out and lmk. Ill then book a conference room. Let me check my schedule

Jan 4, 2019, 10:23 PM

Jin Li

I havent checked my schedule though

Jan 4, 2019, 10:23 PM

Jin Li

I think so

Jan 4, 2019, 10:23 PM

Ben Picker

Do u have time to meet Monday evening?

Jan 4, 2019, 10:23 PM

Jin Li

We also have to make sure our code runs in parallel

Jan 4, 2019, 10:23 PM

Ben Picker

We’ll have clark to talk to Clark.

Jan 4, 2019, 10:22 PM

Jin Li

Not particularly difficult to access but I think we have to pay for it

Jan 4, 2019, 10:22 PM

Ben Picker

Ok great

Jan 4, 2019, 10:22 PM

Jin Li

I’ve used it before

Jan 4, 2019, 10:22 PM

Ben Picker

No my friend does all the time tho and i assume u have ones who do too

Jan 4, 2019, 10:21 PM

Jin Li

Have you used the supercomputer at midway?

Jan 4, 2019, 10:21 PM

Jin Li

Google platform has free 300 dollars credit

Jan 4, 2019, 10:21 PM

Ben Picker

Ben: 󰀀

Jan 4, 2019, 10:21 PM

Jin Li

Ok cool

Jan 4, 2019, 10:20 PM

Ben Picker

Locally i mean

Jan 4, 2019, 10:20 PM

Ben Picker

Like store it

Jan 4, 2019, 10:20 PM

Ben Picker

So i dont have to access it

Jan 4, 2019, 10:20 PM

Ben Picker

In those cases, ive coded on cloud platforms

Jan 4, 2019, 10:20 PM

Jin Li

I mean 400000 photos is probably terabytes of data

Jan 4, 2019, 10:20 PM

Jin Li

U know how to deal with huge data sets?

Jan 4, 2019, 10:20 PM

Ben Picker

Hmmm

Jan 4, 2019, 10:19 PM

Ben Picker

cool

Jan 4, 2019, 10:19 PM

Jin Li

NYC

Jan 4, 2019, 10:19 PM

Ben Picker

Where are u coming from?

Jan 4, 2019, 10:19 PM

Jin Li

I forgot a.m. is at night

Jan 4, 2019, 10:18 PM

Ben Picker

Ok nvm

Jan 4, 2019, 10:18 PM

Ben Picker

Ohhhh

Jan 4, 2019, 10:18 PM

Jin Li

I mean 12:00 am actually

Jan 4, 2019, 10:18 PM

Ben Picker

Or 5

Jan 4, 2019, 10:18 PM

Ben Picker

Like idk 4

Jan 4, 2019, 10:18 PM

Ben Picker

Ok r u free later in the day?

Jan 4, 2019, 10:18 PM

Jin Li

I’m coming back at like 12:00 pm lol

Jan 4, 2019, 10:15 PM

Ben Picker

Jin

Jan 4, 2019, 9:58 PM

Ben Picker

Are u gonna be on campus sunday?

Jan 4, 2019, 9:58 PM

Ben Picker

Hmmm ok yes i think this is the place to start

Jan 4, 2019, 9:58 PM

Ben Picker

Ok

Jan 4, 2019, 9:56 PM

Ben Picker

Yeah

Jan 4, 2019, 9:56 PM

Jin Li

The second one

Jan 4, 2019, 9:56 PM

Ben Picker

Links?

Jan 4, 2019, 9:56 PM

Ben Picker

Is that in one of the links,

Jan 4, 2019, 9:56 PM

Jin Li

Google cloud has 400000 images from the met

Jan 4, 2019, 9:56 PM

Jin Li

Have you guys looked at the datasets?

Jan 4, 2019, 9:55 PM

Ben Picker

Ok cool. I’ll check them out after my meeting

Jan 4, 2019, 3:49 PM

Jin Li

yea there probably is no need to scrape any data

Jan 4, 2019, 3:49 PM

Jin Li

https://console.cloud.google.com/marketplace/details/the-metropolitan-museum-of-art/the-met-public-domain-art-works?pli=1

Jan 4, 2019, 3:48 PM

Jin Li

https://www.kaggle.com/datasets?sortBy=relevance&group=public&search=art&page=1&pageSize=20&size=sizeAll&filetype=fileTypeAll&license=licenseAll

Jan 4, 2019, 3:45 PM

Jin Li

So i sent a link above

Jan 4, 2019, 3:37 PM

Jin Li

I can’t explain as well as a well written article

Jan 4, 2019, 3:37 PM

Jin Li

Also for michal, measuring style is tricky without understand how a cnn works

Jan 4, 2019, 3:36 PM

Jin Li

Ok

Jan 4, 2019, 3:36 PM

Ben Picker

Ok why dont u check and let me know

Jan 4, 2019, 3:36 PM

Jin Li

But I’ll go looking

Jan 4, 2019, 3:36 PM

Jin Li

Nope

Jan 4, 2019, 3:36 PM

Ben Picker

Ok do u know of any?

Jan 4, 2019, 3:36 PM

Jin Li

There probably are existing datasets we can use

Jan 4, 2019, 3:35 PM

Ben Picker

I guess we need to lay out some parameters

Jan 4, 2019, 3:35 PM

Ben Picker

So like

Jan 4, 2019, 3:35 PM

Ben Picker

Yeah

Jan 4, 2019, 3:35 PM

Jin Li

I’m not sure how easy moma is to scrape

Jan 4, 2019, 3:35 PM

Ben Picker

What do u think of this idea?

Jan 4, 2019, 3:34 PM

Ben Picker

Types of art we would include

Jan 4, 2019, 3:34 PM

Ben Picker

For the like

Jan 4, 2019, 3:34 PM

Ben Picker

Mike has suggested looking at moma’s website and using that as a starting point

Jan 4, 2019, 3:34 PM

Jin Li

You sent a link.

http://algobeans.com/2016/01/26/introduction-to-convolutional-neural-network/

Jan 4, 2019, 3:34 PM

Ben Picker

We will start with the similarity stuff

Jan 4, 2019, 3:34 PM

Ben Picker

i am going to discuss with Mike about other objectives first

Jan 4, 2019, 3:34 PM

Ben Picker

Jin, I want to start working on the similarity thing we mentioned

Jan 4, 2019, 3:33 PM

Michał Dzitko

Both these components (matching and then suggesting) are very important to my goal with this project. Perhaps we could recconnect about this during our call, but would you guys reccommend any particular ordering of which one to start with?

Jan 4, 2019, 3:12 PM

Ben Picker

^ jin isnt this what we said yesterday?

Jan 4, 2019, 3:10 PM

Michał Dzitko

Ah - I was just in touch with Ben and I realize I might have misunderstood one thing.

It seems you guys are thinking that the exact match of a photo to an image in a data base isnt an ML task, though the process of suggesting visually/aesthetically similar images is.

Jan 4, 2019, 3:10 PM

Michał Dzitko

Also if you guys have a chunk of time this weekend, I think it would be very beneficial for us all to get on a phone call together to catalyze clarifications and make sure we are all on the same page to start a first pass attempt

Jan 4, 2019, 2:53 PM

Michał Dzitko

Sorry for the wall of text naturally

Jan 4, 2019, 2:52 PM

Michał Dzitko

Hey guys back w comments re: your discussion yesterday:

1. The objective is to create an algorithm that takes a picture and recognizes the piece of art that is in it (i.e. This is the Mona Lisa).

1.1. It is my understanding that we could tie additional data to what we understand as the unique object "mona lisa." By that I mean that as soon as we know that a given picture is the mona lisa, we also know what year its from, what period its from, who the painter is, the materials. That's just because we know these facts about the mona lisa (and almost any other painting since this data is scrape-able from Wikipedia). Am I incorrect in thinking that as soon as we tie a given photo to a photo in our image database, it should be relatively easy to pull up any other data that we tied to that image in our database?

2. I think I'm too much of a noob in ML to know the benefits of supervised/unsupervised approaches. If you'd find it beneficial to explain that to me, I'm all ears.

3. The thought of mathematizing the "distance" between two pieces of visual data as it pertains to art is such a thrilling idea!! My questiosn would be how you defined "style" in your previous approach Jin and whether you think it would be possible to compute any meaningful "distance" between two pieces of visual art?

Jan 4, 2019, 2:52 PM

Ben Picker

Cool

Jan 4, 2019, 3:34 AM

Ben Picker

Ben: 󰀀

Jan 4, 2019, 3:34 AM

Michał Dzitko

hey all sorry for the delay but my evening filled up unexpectedly. Will reply more thoroughly tomorrow

Jan 4, 2019, 3:33 AM

Ben Picker

hmm give me a bit of time to think

Jan 3, 2019, 10:52 PM

Jin Li

Ideas about what we want in it?

Jan 3, 2019, 10:50 PM

Ben Picker

what will we want in it

Jan 3, 2019, 10:48 PM

Jin Li

yea

Jan 3, 2019, 10:48 PM

Ben Picker

So the issue is more then, what data set should we start with

Jan 3, 2019, 10:48 PM

Ben Picker

yeah i agree ok

Jan 3, 2019, 10:48 PM

Ben Picker

ah clever

Jan 3, 2019, 10:47 PM

Jin Li

and we can use some clustering algorithm using that distance measure

Jan 3, 2019, 10:47 PM

Jin Li

essentially, i think we can measure distances from one image to another image by comparing the style

Jan 3, 2019, 10:46 PM

Jin Li

and that mimimizing technique can be used to cluster images

Jan 3, 2019, 10:46 PM

Jin Li

anyways, to create the image, we minimized the "style" difference (how different the styles were for the generated image and for the style image)

Jan 3, 2019, 10:45 PM

Jin Li

since all a cnn is just a bunch of layers with filters that represent different shapes and features

Jan 3, 2019, 10:45 PM

Jin Li

the similarity in "style" can be mathematically defined

Jan 3, 2019, 10:44 PM

Jin Li

so this project was made to generate images, and the way to generate the image was to combine the starting two images (a style image and a base image) and then measure the similarities between the style of the generated image and the style of the style image

Jan 3, 2019, 10:44 PM

Ben Picker

ok sure my b

Jan 3, 2019, 10:43 PM

Jin Li

wait lemme finish explaining

Jan 3, 2019, 10:43 PM

Ben Picker

like 5?

Jan 3, 2019, 10:42 PM

Ben Picker

So these examples have two images, what if you wanted more?

Jan 3, 2019, 10:42 PM

Ben Picker

ideal

Jan 3, 2019, 10:42 PM

Ben Picker

perfect

Jan 3, 2019, 10:42 PM

Jin Li

by comparing the fluidity of images

Jan 3, 2019, 10:42 PM

Jin Li

but i created this

Jan 3, 2019, 10:42 PM

Jin Li

just look at the pictures

Jan 3, 2019, 10:41 PM

Jin Li

https://github.com/JinLi711/Neural-Style-Transfer/blob/master/Examples.ipynb

Jan 3, 2019, 10:41 PM

Jin Li

so like for example

Jan 3, 2019, 10:41 PM

Jin Li

And what it basically does is combine the style of one image with another

Jan 3, 2019, 10:40 PM

Jin Li

I did this project on neural style transfer

Jan 3, 2019, 10:40 PM

Ben Picker

super go for it

Jan 3, 2019, 10:39 PM

Ben Picker

great

Jan 3, 2019, 10:39 PM

Jin Li

It might be helpful

Jan 3, 2019, 10:39 PM

Jin Li

Ok let me explain this

Jan 3, 2019, 10:39 PM

Ben Picker

but art is more fluid

Jan 3, 2019, 10:39 PM

Ben Picker

like a t-shirt is not like a pair of shoes

Jan 3, 2019, 10:39 PM

Jin Li

In an cnn, you “open” the image into its layers and filters

Jan 3, 2019, 10:39 PM

Ben Picker

very definite

Jan 3, 2019, 10:39 PM

Ben Picker

the images we had for ML were like

Jan 3, 2019, 10:39 PM

Ben Picker

i guess im having trouble cuz like

Jan 3, 2019, 10:39 PM

Ben Picker

How do you judge similarity between images?

Jan 3, 2019, 10:38 PM

Jin Li

U suggesting to do the same with pictures?

Jan 3, 2019, 10:38 PM

Ben Picker

When I did word similarity, I used word2vec to create a semantic similarlity space between words.

Jan 3, 2019, 10:38 PM

Jin Li

But we have to make sure that Wikipedia doesn’t have biases that may mess with the data

Jan 3, 2019, 10:33 PM

Jin Li

No clue

Jan 3, 2019, 10:33 PM

Ben Picker

i wonder how they do that

Jan 3, 2019, 10:32 PM

Ben Picker

interesting though cuz presumably they have to handle the traffic

Jan 3, 2019, 10:32 PM

Ben Picker

yeah ok

Jan 3, 2019, 10:32 PM

Jin Li

Because wikipedia literally doesn’t care if you scrape it

Jan 3, 2019, 10:32 PM

Ben Picker

cuz they have stuff to combat it?

Jan 3, 2019, 10:32 PM

Ben Picker

why?

Jan 3, 2019, 10:32 PM

Jin Li

Wikipedia is easy to scrape, sites like moma are a lot harder

Jan 3, 2019, 10:32 PM

Ben Picker

supervised its more reliable

Jan 3, 2019, 10:31 PM

Ben Picker

yeah i agree

Jan 3, 2019, 10:31 PM

Ben Picker

yeah that might be more feasible

Jan 3, 2019, 10:31 PM

Jin Li

Unsupervised learning is usually more difficult to do and less promising

Jan 3, 2019, 10:31 PM

Jin Li

So we do the clustering afterwards

Jan 3, 2019, 10:30 PM

Jin Li

Another suggestion is to do a supervised learning for the training set, apply what the algorithm learned onto pictures in the neighborhood, and then look at patterns in neighborhoods

Jan 3, 2019, 10:30 PM

Ben Picker

ok me too

Jan 3, 2019, 10:27 PM

Jin Li

Seems like a clustering problem so I would say unsupervised is a good idea

Jan 3, 2019, 10:26 PM

Ben Picker

an ability to decipher what kinds of art people like

Jan 3, 2019, 10:26 PM

Ben Picker

that's the question i think we are after

Jan 3, 2019, 10:25 PM

Ben Picker

So I think I'm saying something different than what we talked about earlier. You have correctly stated that the first part of what were discussing doesn't involve machine learning. I think that wasn't quite apparent until I laid out the facts. I am now asking about a separate issue.

We want to be able to take art and find patterns in them. For instance, do certain neighborhoods have preferences? Do they favor certain types of art?

But how are we going to categorize the art in a neighborhood? How can we classify pictures we might pull from ig or whatever into categories that we can then use to look at neighborhoods preferences for particular types of visual art

Jan 3, 2019, 10:25 PM

Jin Li

Or is it to say something about it, like it was made in the renaissance

Jan 3, 2019, 10:24 PM

Jin Li

Like for the painting you just sent, is the goal to label that as Mona Lisa

👍Michał Dzitko

Jan 3, 2019, 10:24 PM

Jin Li

Or label an attribute of the item?

Jan 3, 2019, 10:23 PM

Jin Li

Is the goal just to label precisely what the item is?

Jan 3, 2019, 10:23 PM

Ben Picker

so idk I would think that would make like a supervised classification task harder

Jan 3, 2019, 10:22 PM

Ben Picker

There will presumably be a diversity of different kinds of Art on those murals or styles

Jan 3, 2019, 10:22 PM

Ben Picker

Well, lets suppose we have a bunch of paintings. And we want to create categories to be able to recognize them as part of a set of categories. While we could use traditional ones, don't you think like unsupervised methods would be better? Like for instance, with the mural art Clark has talked about

Jan 3, 2019, 10:22 PM

Ben Picker

give me a second. i have to think how to phrase it

Jan 3, 2019, 10:19 PM

Ben Picker

The second part is like

Jan 3, 2019, 10:19 PM

Ben Picker

So I feel like this task is largely like

1. Have a database of photos with metadata (like time period, etc)

2. Use some kind of image processing that's very simple to match it. Cuz all that's really changing is the lighting and what not

Jan 3, 2019, 10:19 PM

Ben Picker

let me layout what im saying

Jan 3, 2019, 10:18 PM

Ben Picker

give me some time to talk for a second

Jan 3, 2019, 10:18 PM

Ben Picker

i know

Jan 3, 2019, 10:18 PM

Jin Li

Then this isn’t really a machine learning project

Jan 3, 2019, 10:18 PM

Ben Picker

One of Mike's interests is to basically, if given a photo like this, be able to recognize the photo.

Jan 3, 2019, 10:17 PM

Ben Picker

Ben sent a photo.

Jan 3, 2019, 10:17 PM

Ben Picker

Suppose I take a photo of a piece of art

Jan 3, 2019, 10:16 PM

Ben Picker

So let me add a few things to the discussion.

Jan 3, 2019, 10:15 PM

Ben Picker

Alright I think we need a bit more discussion of ideas

Jan 3, 2019, 10:15 PM

Michał Dzitko

Very fair point thank you for raising that. I think the most important categories are [painting name, artist name, year]. I’m happy to relax year to be a general time period for now but for one that’d be a temporary solution as far as I see it.

Wikipedia is a good resource that is my top centralized place where I’d get info data for now. Example page (mobile) https://en.m.wikipedia.org/wiki/Les\_Demoiselles\_d%27Avignon

On top of that there are many repositories of visual art data worth considering. Example page: https://www.moma.org/collection/

I’m not aware of which one of those sets of data would be easier to scrape for the purposes of running an image recognition algo like this. If you do have any ideas on this front they are of course most welcome :)

https://en.wikipedia.org/wiki/Les\_Demoiselles\_d%27Avignon

Jan 2, 2019, 7:23 PM

Jin Li

could you share the wiki pages where we would get the art?

Jan 2, 2019, 7:16 PM

Jin Li

so i wouldnt say its impossible, but it might not be the best idea for getting started

Jan 2, 2019, 7:15 PM

Jin Li

usually the more categories, the lower the accuracy

👍Michał Dzitko

Jan 2, 2019, 7:14 PM

Jin Li

(i misread)

Jan 2, 2019, 7:14 PM

Jin Li

well it would be hard

Jan 2, 2019, 7:14 PM

Jin Li

what do by that?

Jan 2, 2019, 7:14 PM

Michał Dzitko

Just wondering. I’m probably less experienced in ML than you so lmk if I sound completely wrong lol

Jan 2, 2019, 7:14 PM

Michał Dzitko

Ok that could work too. Say we scrape the data from Wikipedia though - these listings do have the year attached to them. Would it be impossibly hard to attach a year to that piece of visual data?

Jan 2, 2019, 7:13 PM

Jin Li

so instead of predicting the exact year the artist painted a painting, it would be more managable to predict a certain time period (say like in the rennaisance)

Jan 2, 2019, 7:12 PM

Michał Dzitko

Could you expand as to what you mean by broader categories?

Jan 2, 2019, 7:11 PM

Jin Li

you too

Jan 2, 2019, 7:11 PM

Michał Dzitko

Hey guys! Pleasure to meet you Jin :)

Jan 2, 2019, 7:11 PM

Jin Li

i think we should start out with broader categories

Jan 2, 2019, 7:10 PM

Jin Li

yea, so I was just saying that it might be difficult to use machine learning to predict the exact artist and exact year

Jan 2, 2019, 7:10 PM

Ben Picker

Hey guys. Starting the group chat so we can get this going

Jan 2, 2019, 7:08 PM

Jin Li

You waved hello to the group.

Jan 2, 2019, 7:08 PM

Ben Picker

Ben created the group.

Jan 2, 2019, 7:08 PM

Generated by Jin Li on Saturday, January 5, 2019 at 12:59 AM EST

[1/2/2019]

Things Accomplished:

Created word clouds

Was able to make plots based on groups

Objectives:

“Add some comments in your log about implications for what you, Hyesun and any others you have read have done, including Scencescapes?”

“list some options that you could adapt from the book, pasting some empales in your log and discussing how to adapt. such COMParing the types of neighbhborhoods in the same picture, such as 3 lines on a line graph, not

each type on a separate PAGE with no comparisons.

Then write out at least ONE example of a good NOTE of the sort we have in Scenescapes and spent many hours refining. ONLY you know what you have nonspecifically and don't communicate it at all.

3D plot

Split into 3 different plots, one plot for each type of word