INF2209: Human Topic Models Midterm

Introduction:

For this assignment, I choose to select a dataset from yelp containing around 7 million reviews from the platform. This dataset was chosen because topic modelling is widely used by private companies to better understand user reviews. One example of this is on google maps, where reviews show key topics such as a certain food offered by the restaurant or the service. I found this update on google maps interesting as it is another way to make user reviews more impactful for the viewer. There is a hidden potential behind topic modelling and user reviews, and I wanted to explore it further in this assignment.

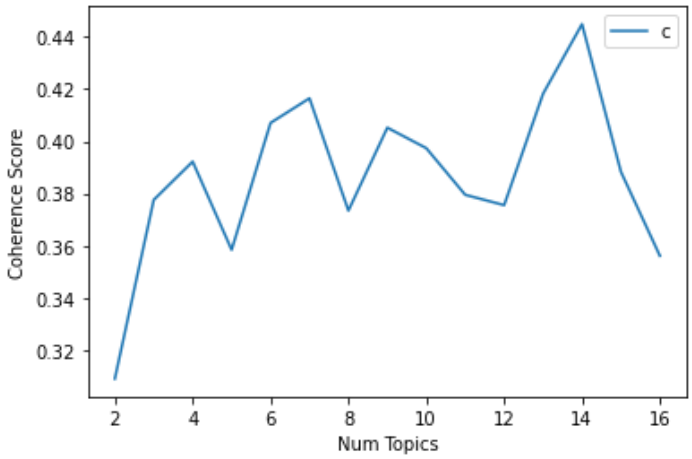
Preprocessing:

Before any models are created, I wanted to explore the dataset to figure out what where the most popular words through WordCloud. However, the dataset seemed to have a "positive" sentiment based on words such as good, great, well and nice being some of the biggest worlds. I decided instead of looking at the entire dataset, I wanted to dive deep on the three-star reviews. It is hard to describe mediocre compared to a really bad place or a really good place. Furthermore, mediocre is subjective from person to person but is it possible to find common topics within these three-star reviews to build a standard three-star review? This is the goal of the assignment.

In the model period, I also noticed that the stopwords from nltk did not apply well to the dataset. There were words such as:" youre, ive, us, im, now, got" that did not really give any insights onto the topic, so I had to add them after the preprocessing phase.

LDA:

When building the LDA model, I found that the number of optimal topics to be 14 based on plotting the number of topics by the coherence score. When the number of topics is 13, the coherence score is 0.45. In addition to the coherence score, the perplexity is -8.2 which determines a good model.



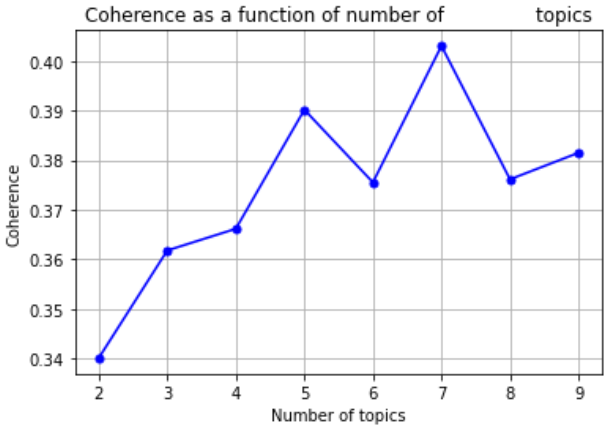
The following keywords from each topic are:

Topic one: Music, fun, far, wings, name, beers, red, twice, afternoon, charge
Topic two: good, food, place, great, service, would, really, nice, pretty, back
Topic three: burger, fires, anything, fast, expensive, burgers, potato, return, aren't, run
Topic four: cream, ice, buffet, making, cook, interior, using, reservation, lamb, setting
Topic five: room, hotel, old, front, someone, walk, manager, floor, stay, received
Topic six: pm, bartender, option, burnt, original, polite, key, liquor, reason, downside
Topic seven: amazing, oysters, slightly, medium, forever, south, rare, complain, sample, oyster
Topic eight: elsewhere, pie, effort, airport, reno, gift, produce, prompt, boring, goods
Topic nine: pizza, coffee, week, chips, tacos, counter, mind, bite, taco, diner
Topic ten: time, order, didn't, came, first, table, another, said, wait, server
Topic eleven: call, giving, add, black, dressing, case, establishment, totally, filling, read
Topic twelve: issues, nola, roast, children, muses, lines, fact, uncomfortable, fb, rude
Topic thirteen: line, recommend, sit, family, Saturday, pancakes, crowded, sign, cocktails, gets

Each review does mention something that standouts such as a particular food or service. To describe this service, each topic also follows with an adjective such as nice or old. Based on these topics, the most interesting observation is that many of these topics start off positive with words such has good, fun, great, nice, and pretty. However, once the topics continue, they get increasingly more negative with words such as complain, burnt, old, uncomfortable, rude, and issues. Based on LDA, it seems that three-star reviews contain mostly positive traits but contain one negative trait within them too.

NMF:

When building the NMF model, I used a similar method of finding the optimal number of topics by plotting the number of topics with coherence. I found that 7 topics are the optimal amount with the coherence score of 0.42.



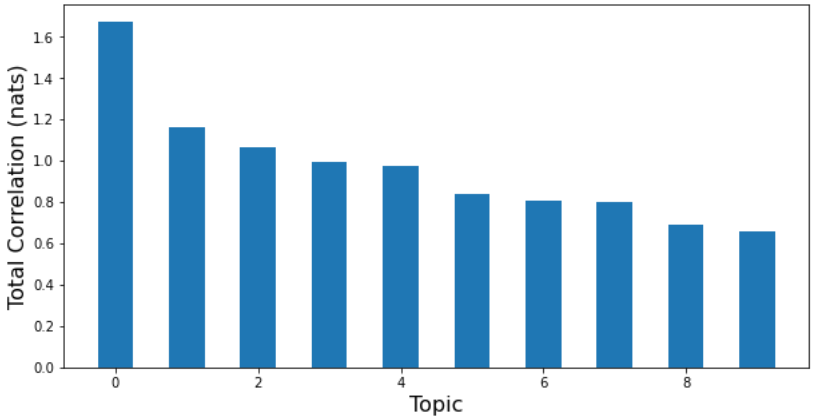
After creating the model with 7 topics, these are the following top 20 words for each topic:

Topic #1	Topic #2	Topic #3	Topic #4	Topic #5	Topic #6	Topic #7
Placing	gotta	pjs	football	chilis	oldest	greeting
Goodbut	prevent	crunchy	Serve	buses	speak	behavior
tilapia	placing	delight	slider	organization	noticed	placing
gin	called	awesome	awesome	fruit	placing	attempt
light	Pricier	tooth	response	fries	write	Bargain
onsite	Serve	chef	death	Said	nieces	Seek
squid	Realizing	place	ole	Sampler	honesty	Dress
badly	seek	wings	froyo	satisfaction	ticket	gambling
Worthy	overhead	organic	Quality	flight	brewery	nieces
dj	thinking	slab	listen	listen	goodbut	mushroom

Just like the LDA model, there seems to be a balance between negative traits and positive traits of the service/product. This can be seen where topics 1 and 6 share the same word “goodbut” which proves that three-star reviews will include a positive and a negative.

CorEx

Producing a CorEx model, I was able to find that after 10 topics, the total correlation metric does level off. Based on this, I choose the total number of topics to be 10. For this model, I didn’t choose any anchor words since I didn’t want to limit the types of responses I would get.



Based on the 10 topics, these are the following top 10 words associated with it.

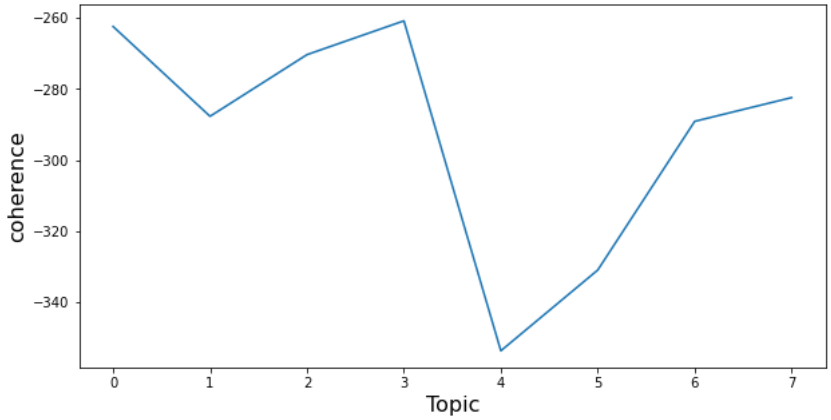
Topic 1: sauce, chicken, cheese, fried, flavor, salad, meat, rice, pork, cooked
Topic 2: minutes, time, took, asked, told, waited, said, wait, arrived, waiting
Topic 3: room, hotel, rooms, stay, stayed, clean, parking, pool, desk, floor
Topic 4: like, dont, know, going, people, want, way, feel, thing, thats

Topic 5: experience, make, right, review, let, new, getting, work, times, started
Topic 6: bar, beer, night, drinks, drink, music, beers, outside, place, friends
Topic 7: ordered, came, table, menu, meal, server, seated, order, food, restaurant
Topic 8: really, didnt, try, wasnt, think, maybe, little, pretty, probably, thought
Topic 9: store, nice, items, need, pros, cons, day, shop, stores, space
Topic 10: cream, taste, chocolate, ice, sweet, dessert, cake, flavors, butter, crust

In this topic model, we can see that the model separated topics based on what the place is, the actual thoughts on the place and products/services. The topics that contain actual items such as topic 10 doesn't really help with understanding three-star-reviews. Furthermore, the topics that do contain thoughts have similar words such as topic 4 where it is all negative.

BTM

Out of all the models, BTM has been the most successful so far. Creating the model, I was able to find that coherence would suddenly drop in after the fourth topic. Using this information, I set the number of topics to be 4 with the coherence as -260.



Based on the topics below, it seems that the model categorized it based on the following: restaurant reviews, services, hotel, and bad reviews. In topic 1, we can see that there are reviews for restaurants with positive words such as favourite, recommended, charm, looooove and great. In topic two, had reviews for archery, a club in Vegas, plumber and a school which all had mixed reviews. In topic three, there are also mixed reviews for hotels such as noisy, uncomfortable but there are also reviews where it would say nice, clean, and comfy. In topic four, there seems to be just general reviews that contain negative words such as mehnothing, mediocoreservers, average, disappointed, tasteless.

This model wasn't perfect as in some topics, it would have an anomaly such as topic three contain a review on a salon when it mainly consists of hotels.

INF2209: Human Topic Models Midterm

Topic 1	Topic 2	Topic 3	Topic 4
[cigarettes, one, favorite, movies, dirty, franks, might, well, bar, featured, movie, id, expect, see, courtney, love, drunk, corner, booth, ben, affleck, pouring, pints, slaughtering, pick, lines, janeane, garofalo, wearing, black, discussing, beatnik, culture, fellow, intellectuals, kate, hudson, sitting, bar, dress, lamenting, lost, love, paul, rudd, consoling, christina, ricci, chainsmoking, outside, dave, chapelle, pulls, disco, cab, franks, melting, pot, local, flavor, like, movie, gets, personality]	[dig, guys, friendly, help, leased, skis, fam, last, years, heat, mold, ski, boot, liners, heat, blower, machine, liners, dont, like, oven, think, guy, heat, mold, could, done, bit, detailed, job, ill, go, back, next, time, need, mold, dont, purchase, kayak, equipment, know, good, selection	[missed, ourconnecting, flight, weather, stayed, rooms, clean, bed, comfy, would, stay]	[mehnothing, special, ok]
[report, card, steak, tacos, guacamoleb, tortilla, chipsc, serviceb, cleanliness, price]	[megaplex, sometimes, helpful, staff]	[hotel, livable, dated, exhaust, bathroom, air, cond, noisy, effective, rate, paid, stayed, sheraton, canal, hotel, staff, friendly, location, good]	[presentation, nicetaste, mediocreservers, clue, roll, whatmakes, wonder]
[say, head, right, back, staircase, go, hurricanes, go, peoplewatch, provoke, beadseeking, tourists, balcony, quite, fun]	[perhaps, bit, jaded, right, applying, grad, schools, mba, iupui, charges, send, transcripts, schools, every, school, want, apply, pay, application, fee, give, iupui, every, transcript, didnt, give, enough, money, throughout, years, undergrad, want, gouge, much, money, humanly, possible, many, colleagues, applying, grad, schools, issue, colleges, send, transcripts, free, charge, must, nice	[nice, neighborhood, bar, smoking, secure, locked, front, door, uniformed, door]	[local, grassfeed, beef, appeals, middleclass, whiteness, said, beef, dryness, cardboard, big, mac, container, like, chardonnay, dry, sustainably, sourced, meat]
[decent, cocktails, recommended, large, groups, bar, tries, salonstyle, furniture, decor, look, ad, hoc, curb, steals, salon, speakeasy, charm, much, prefer, ranstead, room]	[estimate, another, plumber, highest, quote, plumbers, interviewed, job, prompt, communication, though, reconsider, future, jobs, seems, reliable, often, get, pay]	[hella, uncomfortable, seats, equivalent, plastic, stools, outlets, wifi, slow, af, cuties]	[like, olive, garden, salad, tasty, enjoyed, stuffed, rigatoni, sausage, flavorful, also, stuffed, chicken, marsala, chicken, asiago, pasta, dish, appetizer, raviolis, calamari, stuffed, mushrooms, nothing, special, calamari, fried, chewy, service, average, best]

[looooooooove, itwut, great, cell, phone, storeespecially, live, santa, barbraand, place, nice, amount, space, happy, get, helped]	[mom, bought, groupon, archery, lessons, lane, rental, facility, nice, however, dont, bother, lessons, involved, less, five, minutes, instruction, would, visit, facility, wouldnt, bother, lessons]	[brookes, music, needs, go, pop, edm]	[disappointed, fries, overlooked, burger, undercooked, ordered, egg, burger, runny, made, bottom, bun, useless]
[evening, started, great, min, seated, ordering, served, cocktails, eletricity, offfor, entire, time, till, leftdrinks, house, since, tolerated, uncomfortable, situation, guedt, town]	[stayed, one, night, dec, getting, snowed, nashvilleyep, snowed, fun, place, staykinda, vegas, style, accommodations, someone, gets, vegas, times, year, say, enjoyed, staying, downtown, hermitageand, choice, next, time, head, nashville]	[would, say, rockledge, nail, salon, hit, miss, manicures, pedicures, generally, goodbut, waxing, problematic, get, one, inexperienced, girls]	[mac, cheese, good, toppings, kind, unspectacular, corn, bread, crumbs, little, tasteless, also, overpriced, mac, cheese]
[met, fearless, new, cm, brett, lunch, break, right, hummus, fo, sho, hofbrauhaus, draft, munich, righteous, would, call, hole, wall, cept, made, gas, station, potential, depending, party, bring, yelpers, made, enlightening, bs, session, id, alos, saw, potential, menu, healthy, sammies, way, break, day, brett, lets, soon]	[update, barbers, club, moved, locations, north, side, route, extremely, difficult, hassle, get, also, trudy, awesome, barber, longer, works, done, barbers, club, cherry, hill, bummer, lot, potential]	[gas, station, gas, station, nothing, exciting, particular, one, sits, gibsonton, drive, hwy, onramp, tampa, convenient, plenty, pumps, usual, sundries, store, restroom, needed, guess, need, gas, station, stop]	[drink, special, really, good, balsamic, vinegar, basil, alcohol, nice, touch, fried, chicken, ok, small, bit, greasy, side, sides, mac, cheese, underwhelming, cole, slaw, good, could, vinegar, crunch, brussel, sprouts, favorite

Conclusion

Overall, I believe that my models created aren't as strong as I liked due to some preprocessing errors and the type of dataset I chosen. The yelp dataset contains texts that are contain slang which makes it hard to update the stop words to filter out unnecessary text. In addition, the reviews do not contain that many text in nature which makes it hard for certain models to analyze. However, the strongest models for the dataset will be the NMF, LDA where both models have coherence scores around 0.44. I do believe that BTM has a strong model too with my interpretation of the topic models but unsure of the method of calculating coherence since the bitermplus documentation doesn't state it.