Capstone project task 1

Abstract

In order to finish this task, I used python and jupyter notebook to develop and run my code.
I used graphlab to run text alalytic algorithms and matplot to visualize.

Implementation

At first I imported all dataset. Because of the task requirement, I think that only the text reviews themselves are important, so I remove columns like: 'votes', 'data' and id columns.

```
reviews = gl.SFrame.read_json('yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_
review.json',orient='lines')
reviews.remove_columns(['business_id','date','review_id','user_id','votes','type'])
```

After that, I created a dictionary map each word appeared in the review with its tf_idf score. I also removed stop words from the dictionary as they are mostly invaluable.

```
reviews['word_count'] = gl.text_analytics.count_words(reviews['text'],delimiters = delimiters
dict_trim_by_keys(stopwords, exclude=True)
reviews['tf_idf'] = gl.text_analytics.tf_idf(reviews['word_count'])
```

Because the data is quite big, I only use 10% of all reviews (about 110,000 reviews) to analyse. The data after preprocessing look like this:

```
sample = reviews.sample(.1,seed = 317)
sample.head()
```

stars	text	word_count	tf_idf
5	Dr. Eric Goldberg is a fantastic doctor who has ...	{'accessible': 1L, 'fantastic': 1L, ...	{'accessible': 6.440940320962146, ...
3	Ate a Saturday morning breakfast at the Pine ...	{'saturday': 1L, 'madison': 1L, 'qualms': ...	{'saturday': 3.7162420980736974, ...
5	i rarely give five star reviews but for what ...	{'breakfasts': 1L, 'give': 1L, 'taste': 1L, ...	{'breakfasts': 6.55469249425622, 'gi ...
3	This is definitely not your usual truck stop. ...	{'good': 1L, 'rude': 1L, 'restaurant': 1L, ...	{'good': 0.9400482737462729, ...
4	Unlimited hot coffee. I don't have any ...	{'generous': 1L, 'good': 1L, 'restaurant': 1L, ...	{'generous': 4.523116918480465, ...
5	Californians are all about the In-N-Out, w ...	{'love': 1L, 'family': 1L, 'feel': 1L, 'almo ...	{'love': 1.8910237607652902, ...
4	My meaty goodness...why isn't Culvers a ...	{'cheese': 1L, 'picnic': 1L, 'chain': 1L, 'bac ...	{'cheese': 2.5850249953356883, ...
5	Love it!!!!! Love it!!!!!! love it!!!!!!! ...	{'love': 4L, "culver's": 1L} ...	{'love': 7.564095043061161, ...
3	If you live in Madison, you too probably have a ...	{'charter': 1L, 'love': 1L, 'inadequate': 1L, ...	{'charter': 9.46779250322994, 'lo ...
2	My arch-enemy.\n\nI've never encountered a ...	{'contacted': 1L, 'extra': 1L, 'lack': 1L, ...	{'contacted': 6.237942631329764, ...

[10 rows x 4 columns]

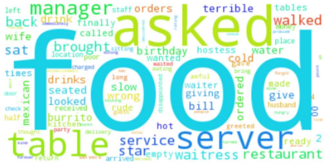
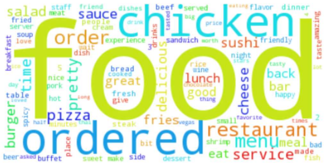
Task 1.1:

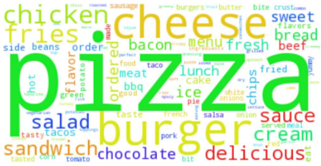
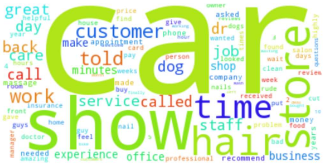
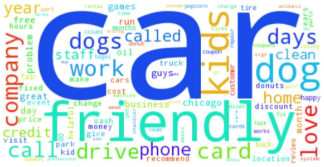
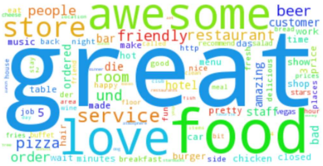
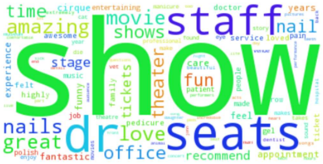
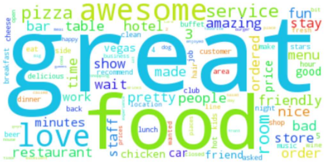
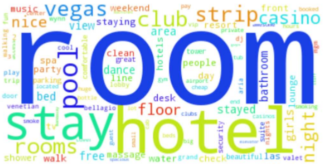
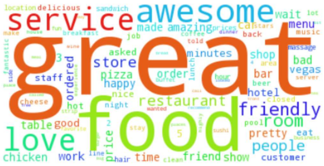
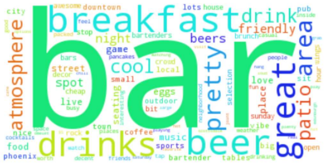
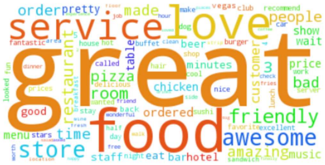
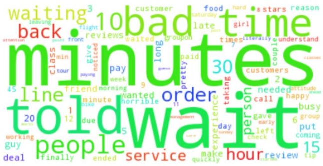
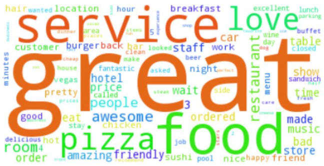
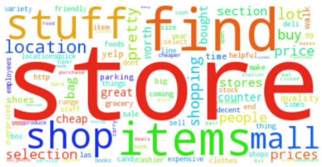
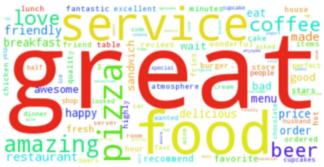
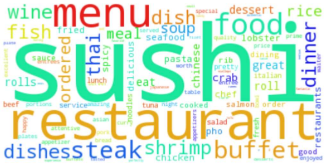
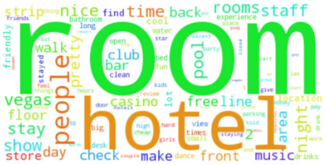
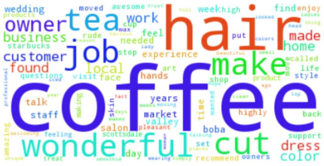
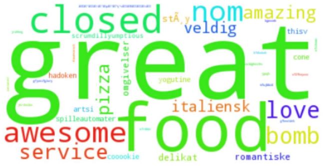
In this task, I will use 2 different methods of lda topic modeling named Collapsed Gibbs sampling (cgs) and AliasLDA method (alias) to create 10 topics each model. After that, I will creat visualization base on 2 model topics and compare them.

```
def model_cgs(data):
    return gl.topic_model.create(data,num_topics=topic_size, num_iterations=200,print_interval=50,method='cgs')
def model_alias(data):
    return gl.topic_model.create(data, num_topics=topic_size, num_iterations=50,print_interval=50,method='alias')

topic_model_cgs = model_cgs(sample['tf_idf'])

topic_model_alias = model_alias(sample['tf_idf'])
```

<table>
<tr><th>Topic</th><th>Sample data using alias method</th><th>Sample data using cgs method</th></tr>
<tr><td>0</td><td>
</td><td>
</td></tr>
</table>

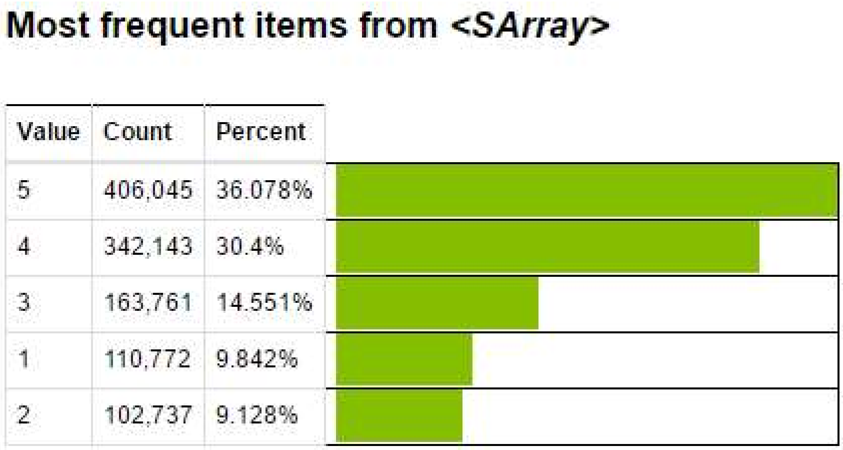
Topic	Sample data using alias method	Sample data using cgs method
1		
2		
3		
4		
5		
6		
7		
8		
9		

It is clear that alias method did better job than cgs method.
Model using alias method can differentiate topics very good and topics look very distinct: restaurants, coffee shops, stores, movie theatres, ...
By contrast, cgs method did not do well its job because more than half the topics scored the word *'great'* as the most important words. Therefore, in task 1.2 I will only use alias method.

Task 1.2:

In this task, I will get about 100,000 positive reviews (5 stars) and about 100,000 negative reviews (1 or 2 stars) as sample and run topic models on them.

```
gl.canvas.set_target('ipynb')
reviews['stars'].show('Categorical')
```



Most frequent items from <SArray>

Value	Count	Percent
5	406,045	36.078%
4	342,143	30.4%
3	163,761	14.551%
1	110,772	9.842%
2	102,737	9.128%

It is clear that reviews with 4 or 5 stars account for 66% of original data and reviews with 3 or less stars share the remaining 33%. I will use 25% of positive reviews (5 stars) and 30% of negative reviews(less than 3 stars) as sample and run topic model on these samples.

```
positive_reviews = reviews[reviews['stars'] == 5]
sample_positive = positive_reviews.sample(.25 ,seed = 317)
sample_positive.shape
```

```
(101630, 4)
```

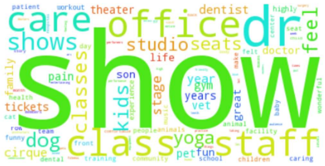
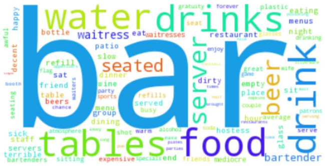
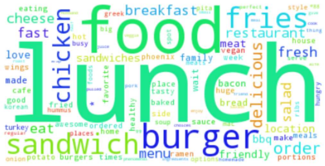
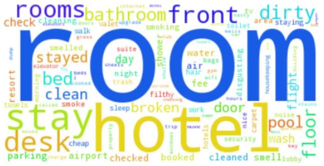
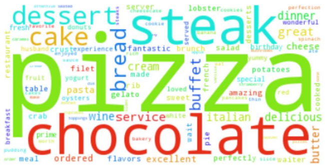
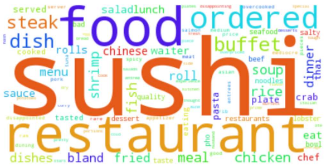
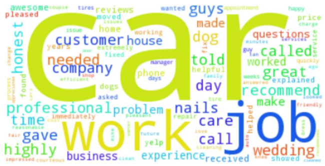
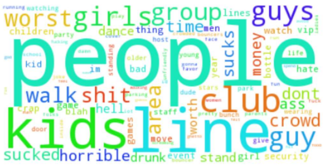
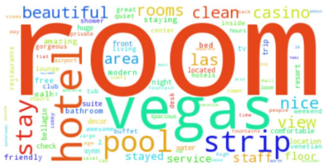
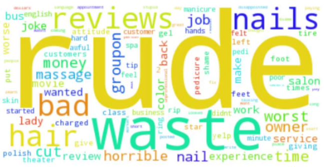
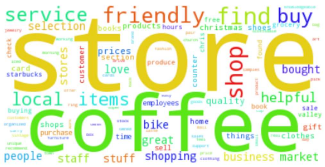
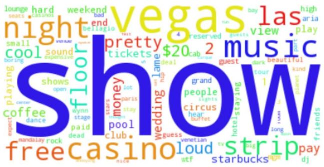
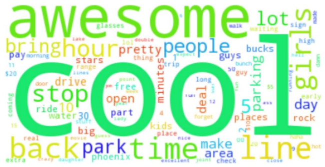
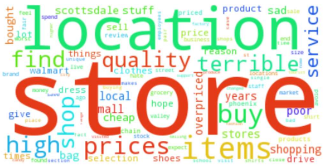
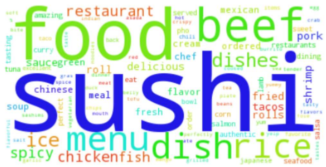
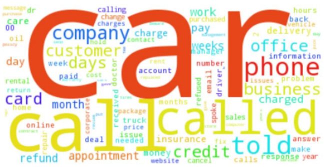
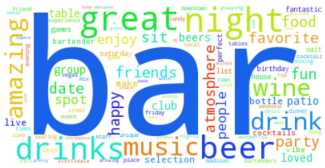
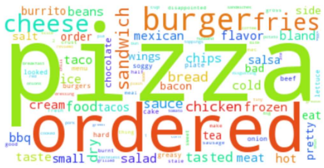
```
negative_reviews = reviews[reviews['stars'] < 3]
sample_negative = negative_reviews.sample(.5,seed = 317)
sample_negative.shape
```

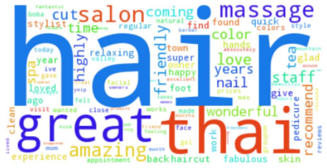
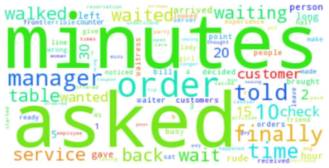
```
(106853, 4)
```

```
positive_model_alias = model_alias(sample_positive['tf_idf'])
```

```
negative_model_alias = model_alias(sample_negative['tf_idf'])
```

Topic	Positive	Negative

0	show class staff office dr care shows classes kids dog yoga studio seats feel life son year years gym pet fun theater dentist doctor tickets pain cirque	bar water drinks drink server tables food beer seated waitress bartender table beers sat friend served hostess dirty place terrible sick
1	lunch food fries chicken sandwich burger breakfast cheese fast delicious salad fresh restaurant meat vegan bacon bread sandwiches love eat menu order bbq location friendly portions	room hotel stay desk rooms bathroom front dirty stayed bed clean pool floor night door broken water air fee check parking airport checked booked cleaned smell wash
2	pizza chocolate steak dessert cake bread buffet cream italian delicious service wine pasta table filet fantastic dinner great amazing excellent butter meal ordered flavors perfectly	sushi restaurant food ordered steak dish buffet menu rolls roll shrimp chinese soup rice sauce noodles dinner plate crab dishes bland fried meal chicken chef
3	car work job customer house company honest needed professional time highly gave problem nails care call told called great recommend dog guys made day wedding business experience	people kids line club girls group guys worst time walk shit area sucks money crowd guy sucked horrible drunk hell bad dance stand
4	room vegas hotel stay pool strip beautiful rooms clean casino area las nice view stayed service staff floor suite bathroom friendly	rude waste reviews nails hair bad groupon money massage job worst owner service horrible nail experience time cut review lady salon
5	store coffee service friendly find buy shop local items helpful bought prices love selection stores great bike shopping people staff stuff business market starbucks	show vegas night las music cool floor free casino strip loud pretty tickets $20 people pool money circus coffee starbucks view
6	cool awesome lot girls bring hour people pretty stop open parking day back park time line make area places deal free	location store find quality terrible service shop buy prices items high local mall cheap overpriced poor years shopping scottsdale walmart stores
7	sushi food beef restaurant dishes menu dish rice ice spicy chicken fish roll delicious sauce green fresh tacos rolls shrimp pork fried chinese soup	car call called company customer phone office told credit card days business charge month refund appointment work pay charged insurance
8	bar great night drinks music beer amazing wine drink friends atmosphere happy date spot food favorite fun party enjoy sit beers patio group club loved	pizza ordered burger fries cheese sandwich chips salsa bread bacon wings taco mexican flavor food tacos sauce chicken frozen cream bbq taste small salad tasted meat hot burrito beans

9		

Even though the most scored words in models are words that related to business categories of the places being reviewed. We can notice that there are words like *love*, *awsome* or *amazing* in positive feedbacks while in negative reviews there are many negative words like *rude*, *waste*, or *horrible*.