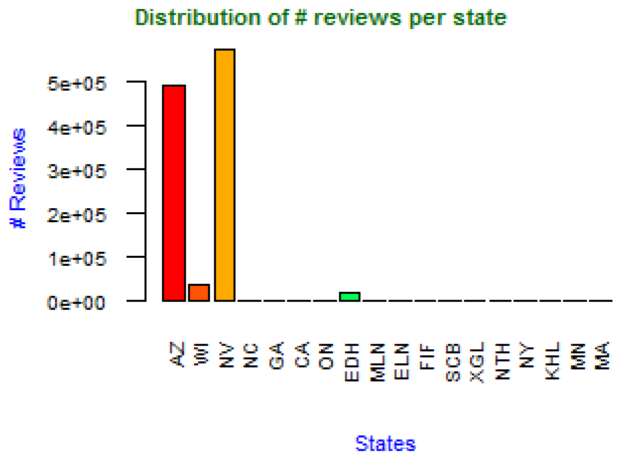
Task 1 - Data Mining Capstone Project

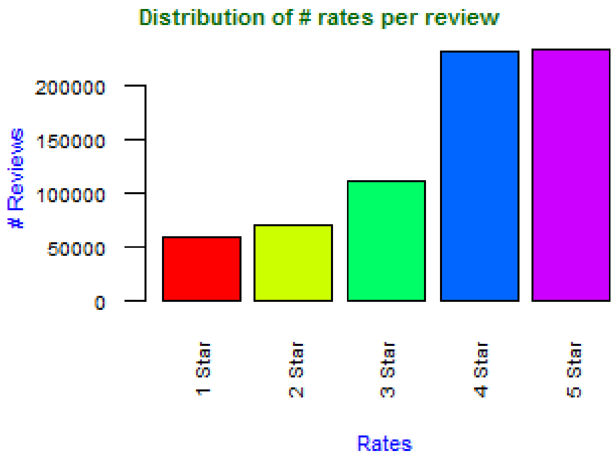
- *R language is used to prepare data and extract topics from the reviews*
- *D3 and R is used to visualize data*

Load datasets

The datasets **yelp_academic_dataset_review.json** and **yelp_academic_dataset_business.json** are loaded in R objects

The distribution of rates of restaurants and of reviews over different states



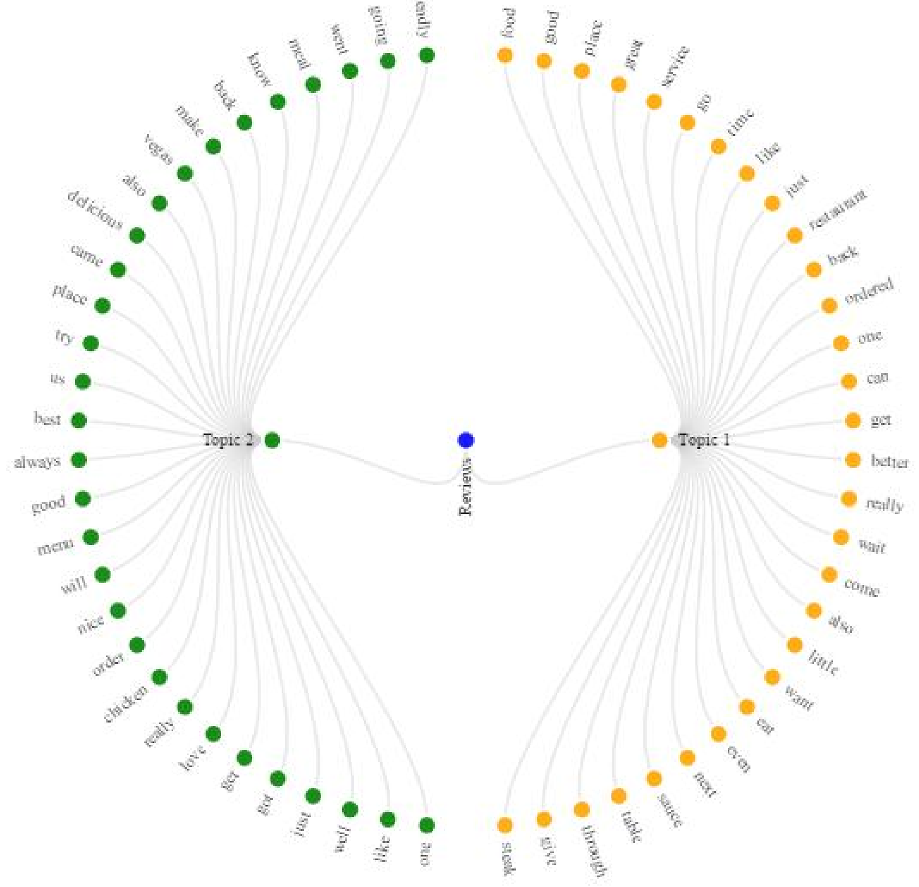


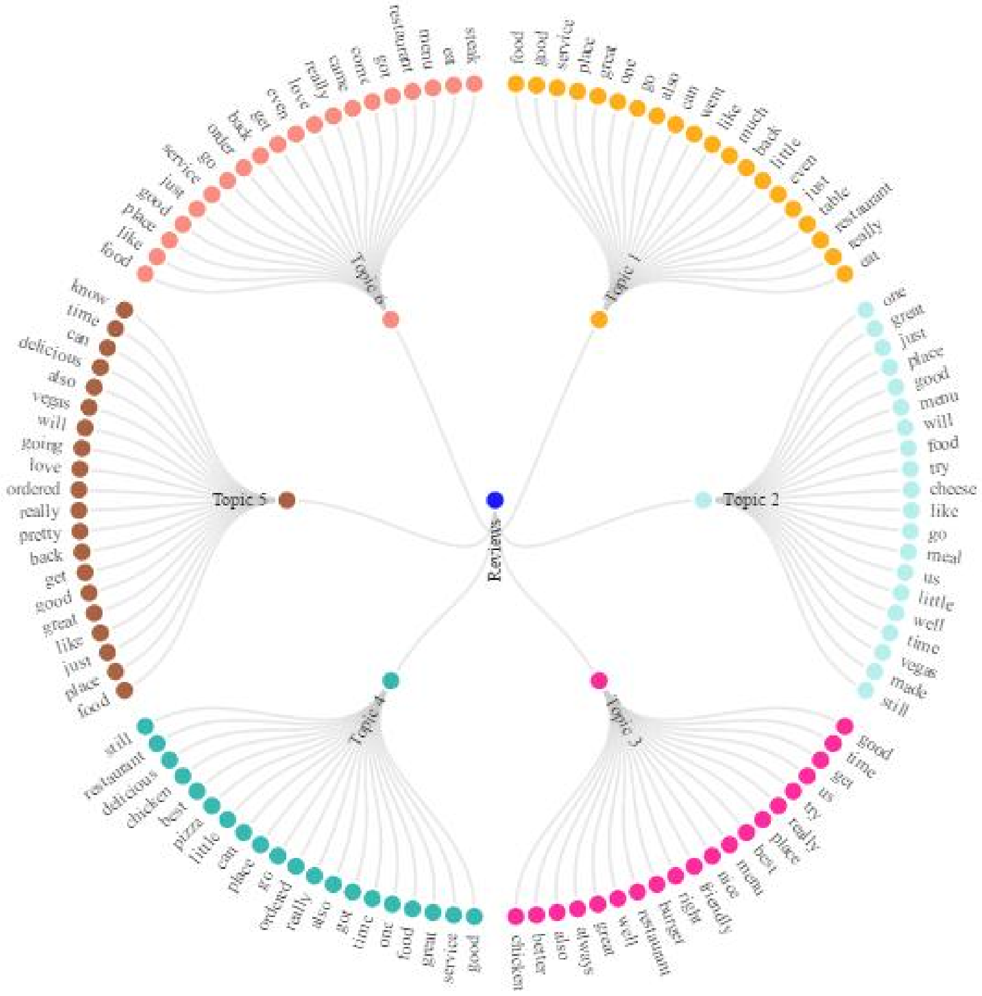
Prepare data of restaurant reviews for mining

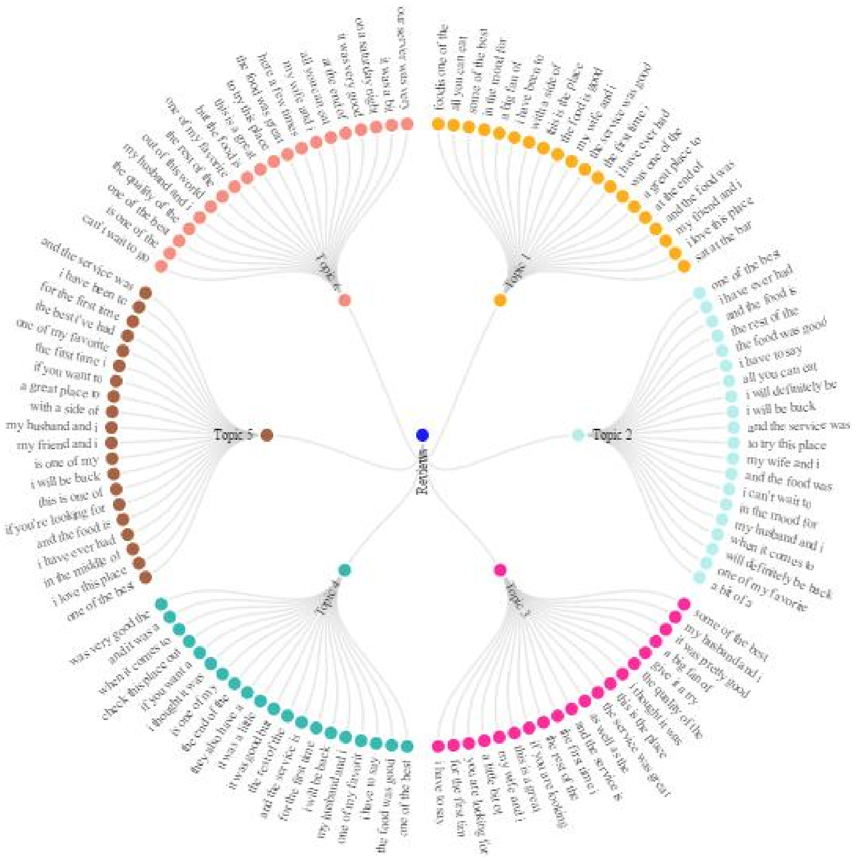
- Restaurant reviews are extracted and only a sample ~ **100000** reviews is used for mining
- **Low frequency** Term and Terms appear in less than **5 documents** are filtered from sample
- The sample is then randomly splitted into 70% for train set and 30% for test set
- Negative and Positive reviews of Vietnamese cuisine are extracted from the sample

Topic Model based on N-Grams

- The Model **Latent Dirichlet Allocation** (LDA in R) is used. 19 models based on **1-Gram** (word) with # topic from 02 to 20 are trained and **Perplexities** are calculated based on **test sets.** The best model suggested by **min value of perplexities** is 02 topics. In fact, a manually judgment finds a high correlation between topics in all models, thus 02 topics models display enough information about topic
- However, in order to justify and make a comparison above conclusion, **2-topics** model is presented along with **6-topics** model with 20 top terms for each topic .
- Beside, I also apply **LDA** based on **4-Grams** and the result is the same of **1-Gram.** However, terms **4-Grams** makes meaning of topics clearer

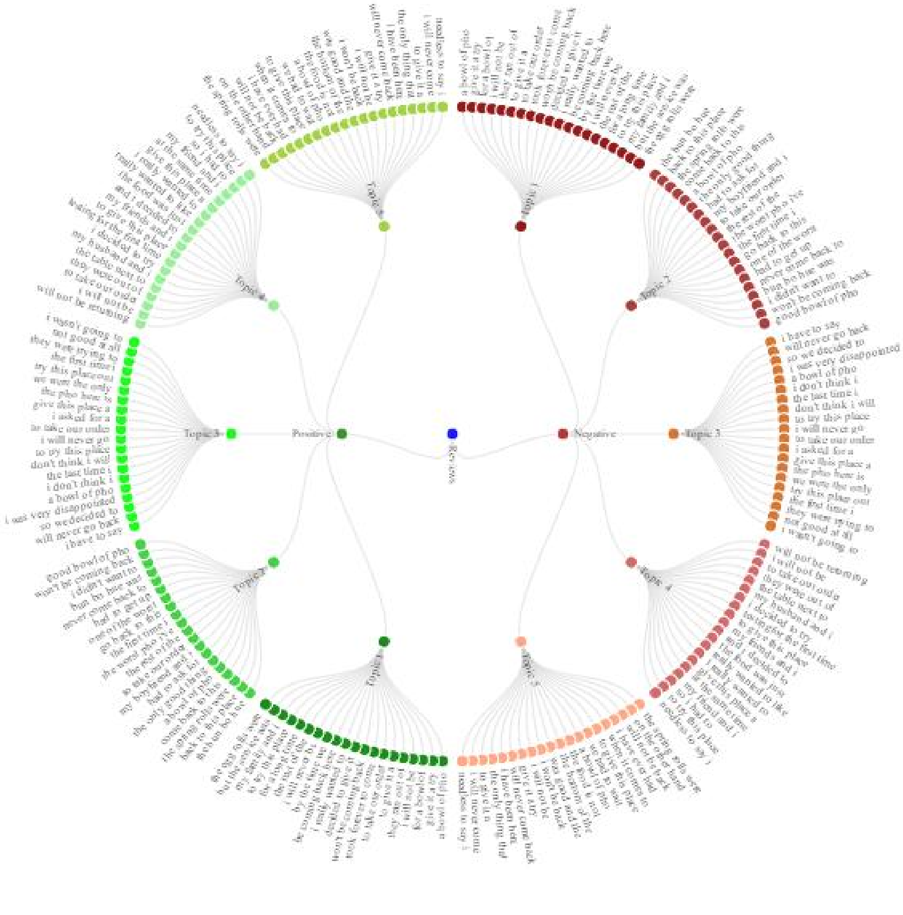

Reviews
Topic 1
food
good
place
great
service
go
time
like
just
restaurant
back
ordered
one
can
get
better
really
wait
come
also
little
want
eat
even
next
sauce
table
through
give
steak
Topic 2
friendly
going
went
meal
know
back
make
vegas
also
delicious
came
place
try
us
best
always
good
menu
will
nice
order
chicken
really
love
get
got
just
well
like
one


Reviews
Topic 1
food
fresh
service
place
great
one
go
also
can
went
like
much
back
little
even
just
table
restaurant
really
eat
Topic 2
one
great
just
place
good
menu
will
food
try
cheese
like
go
meal
us
little
well
time
vegas
made
still
Topic 3
good
time
get
us
try
really
place
best
menu
nice
friendly
right
burger
restaurant
well
great
always
also
better
chicken
Topic 4
still
restaurant
delicious
chicken
best
pizza
little
can
place
go
ordered
really
also
got
time
one
food
meat
service
good
Topic 5
know
time
can
delicious
also
vegas
will
going
love
ordered
really
pretty
back
get
good
great
like
just
place
food
Topic 6
food
like
place
good
just
service
go
order
back
get
even
love
really
came
come
got
restaurant
menu
eat
steak



Topic Model based on 4-Grams of Negative and Positive Reviews

- There are no differences between topics of **Negative** and **Positive** reviews if models are fitted based on **1-Gram, 2-Grams and 3-Grams.** Nevertheless, there is a clear difference if **4-Grams** is used for fiiting LDA model. The 6-topics model is visualized in this report


Reviews
Positive
Negative
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6