Yelper

A Collaborative Filtering Based Recommendation System

Team DeepBlue
Chuan Sun
Capstone Project @ NYC Data Science Academy
9/20/2016

Agenda

Why?

Why we need recommendation?

What?

What are the components for Yelper?

How?

How to build Yelper?

Demo

User-business network; Yelper main page; Simulation of users' recommendation requests handling

Summary & Future works

What we learn and what's next?

Information overload is a real phenomenon which prevents us from taking decisions or actions

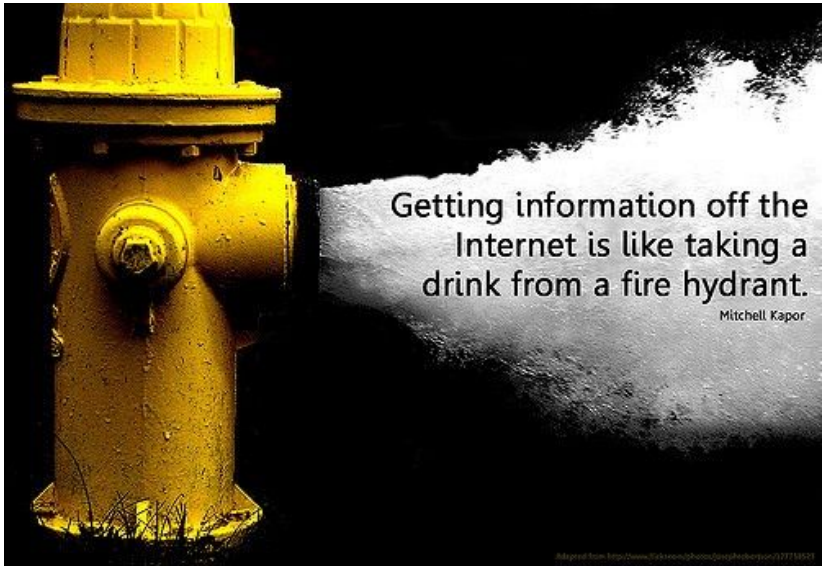
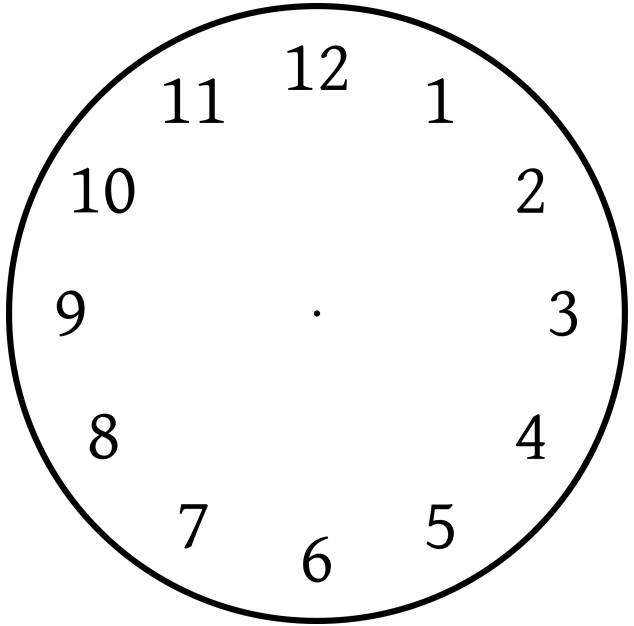


Image link

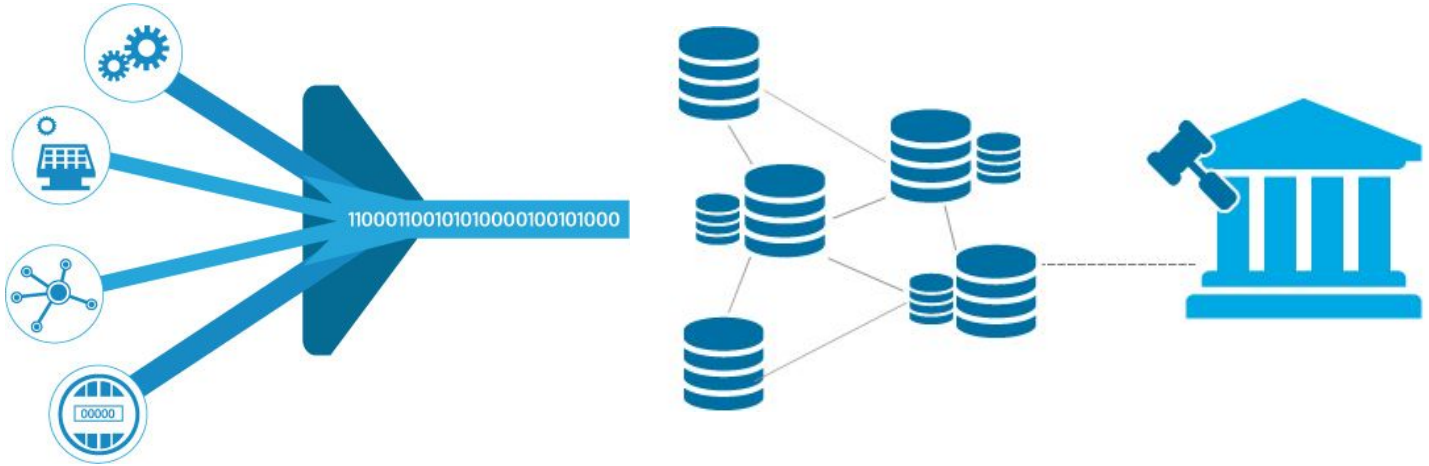
But life is short. We only have 24 hours per day



Recommendation is becoming extremely common in recent years

Real-time recommendation system may become a new normal in data era

- Gain real-time insight
- Enable rapid development
- Perform real-time analytics



Agenda

What?

What are the components for Yelper?

Demo

User-business network; Yelper main page; Simulation of users' recommendation requests handling

Why?

Why we need recommendation?

How?

How to build Yelper?

Summary & Future works

What we learn and what's next?

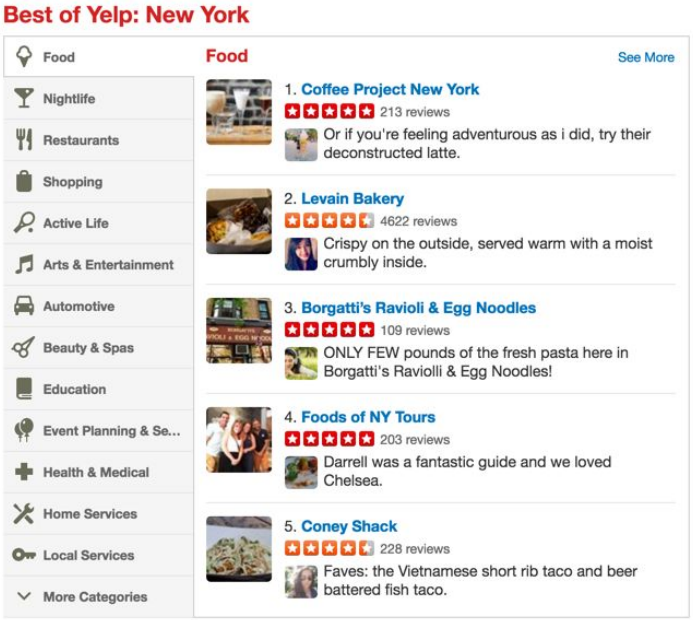
Yelp Challenge 2016 Dataset is a good sandbox to build recommendation engine

The Challenge Dataset:

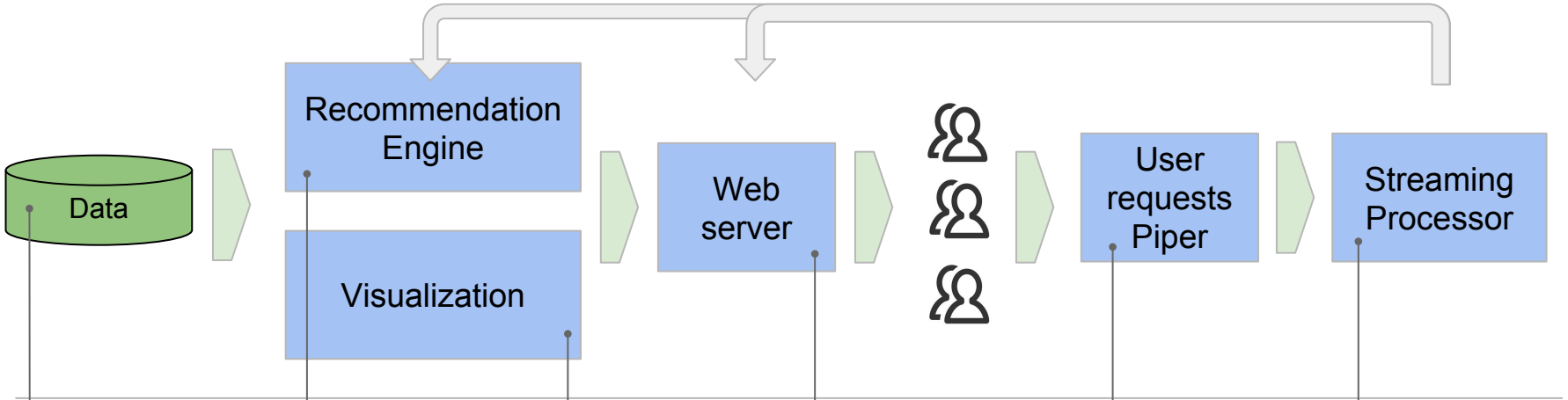
- **2.7M** reviews and **649K** tips by **687K** users for **86K** businesses
- **566K** business attributes, e.g., hours, parking availability, ambience.
- Social network of **687K** users for a total of **4.2M** social edges.
- Aggregated check-ins over time for each of the **86K** businesses
- **200,000** pictures from the included businesses

Cities:

- U.K.: Edinburgh
- Germany: Karlsruhe
- Canada: Montreal and Waterloo
- U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison



Yelper consists of 5 major components



Data source
Provide raw or processed data as the fuel for recommendation engine

Predict ratings
Adopt collaborative filtering or content-based filtering

Visual insight
Distill connected components or in/out degree for users and businesses

User frontend
Allow customer to get customized recommendations

Pipe user requests into streaming
Handle real-time data feeds as message broker

Process user requests
Digest continuous data from Kafka in a scalable and fault-tolerable way

Agenda

Why?
Why we need recommendation?

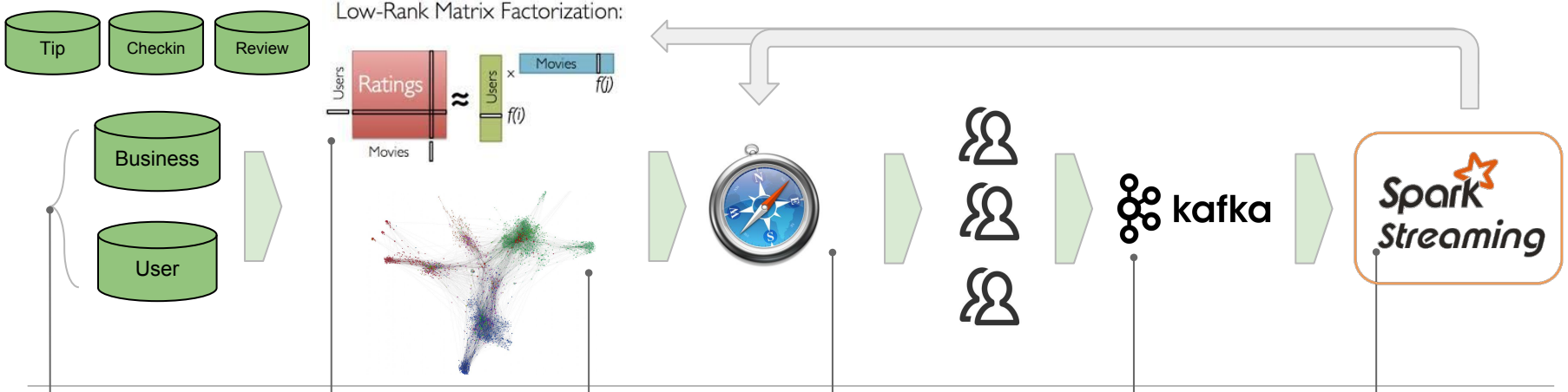
What?
What are the components for Yelper?

How?
How to build Yelper?

Demo
User-business network; Yelper main page; Simulation of users' recommendation requests handling

Summary & Future works
What we learn and what's next?

Yelper is built on top of Spark platform



Data source

(1) Use business and user data in Yelp dataset
(2) Map raw IDs to integers to make later stage a lot easier
(3) Splitting business data by city allows fine-tuned models

Predict ratings

(1) Use Spark MLlib to train Collaborative Filtering models per city
(2) Tune the best rank
(3) Filter user-input keywords to prune user's preferences

Visual insight

(1) Use Spark GraphX to extract connected components
(2) Use graph-tool library to draw large graphs
(3) Use D3.js to build dynamic graphs

Client frontend

(1) Use Cherrypy to allow Flask talk to Spark context
(2) Use Google Map API to show recommended results

Pipe user requests into streaming

(1) Pipe user requests into streaming
(2) Create json-based topics to allow consumer-provider communication

Process user requests in stream

(1) Use Spark SQL to query data frames
(2) Support parallelism
(3) Fault-tolerant

Agenda

What?
What are the components for Yelper?

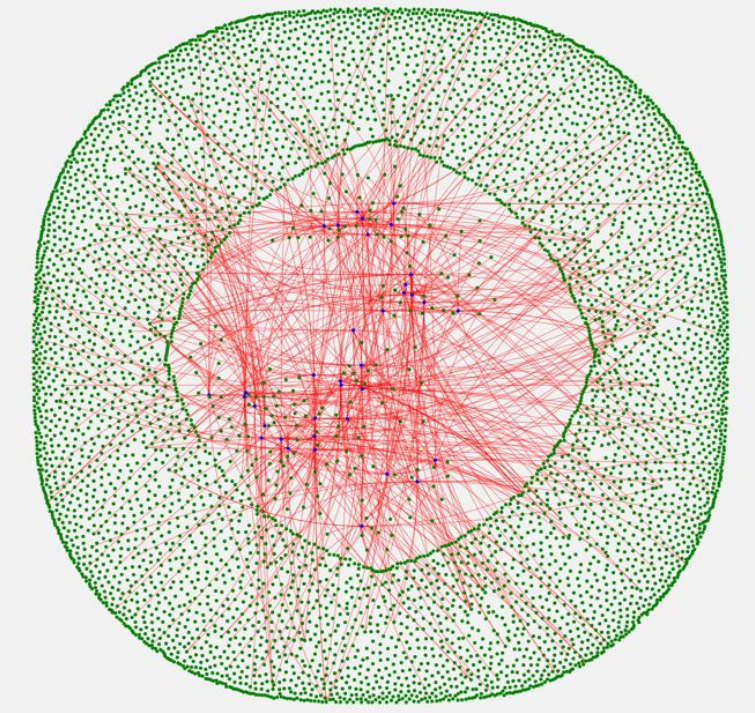
Demo
User-business network; Yelper main page; Simulation of users' recommendation requests handling

Why?
Why we need recommendation?

How?
How to build Yelper?

Summary & Future works
What we learn and what's next?

The city-wise user-business network in Yelp tells a lot about a city



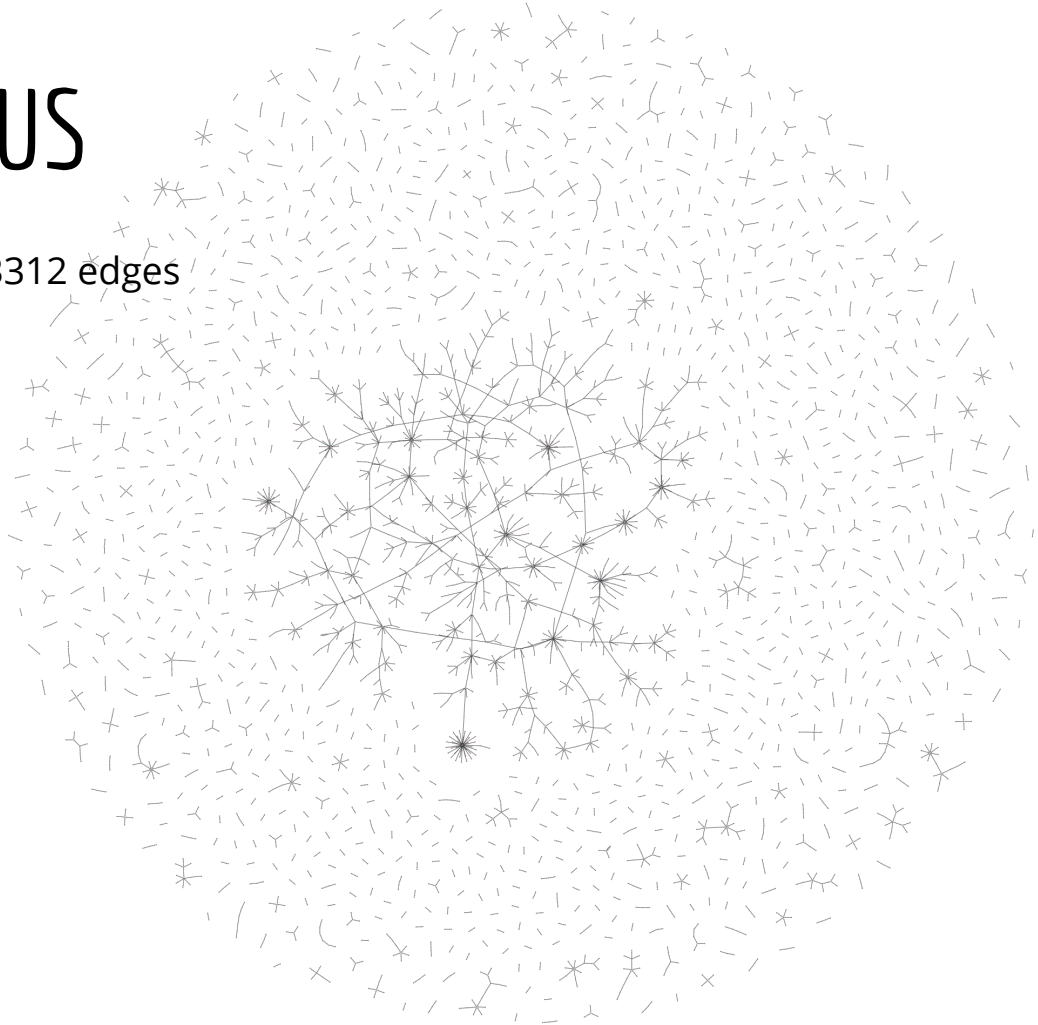
- Motivation
  - User-business interaction may reflect the economy trend in a city
- Each city is distinct, so does its user-business network
- Nodes: u1, u2, ... b1, b2, ...
- Edges: u1 → b1, u1 → b2, ...
- The network is a bipartite
- If Yelp provides the timestamp of rating data, we may see the evolution of economy development

Demo 1: User-business network visualization in the city of Madison, US

Charlotte, US

Randomly selected 3312 edges

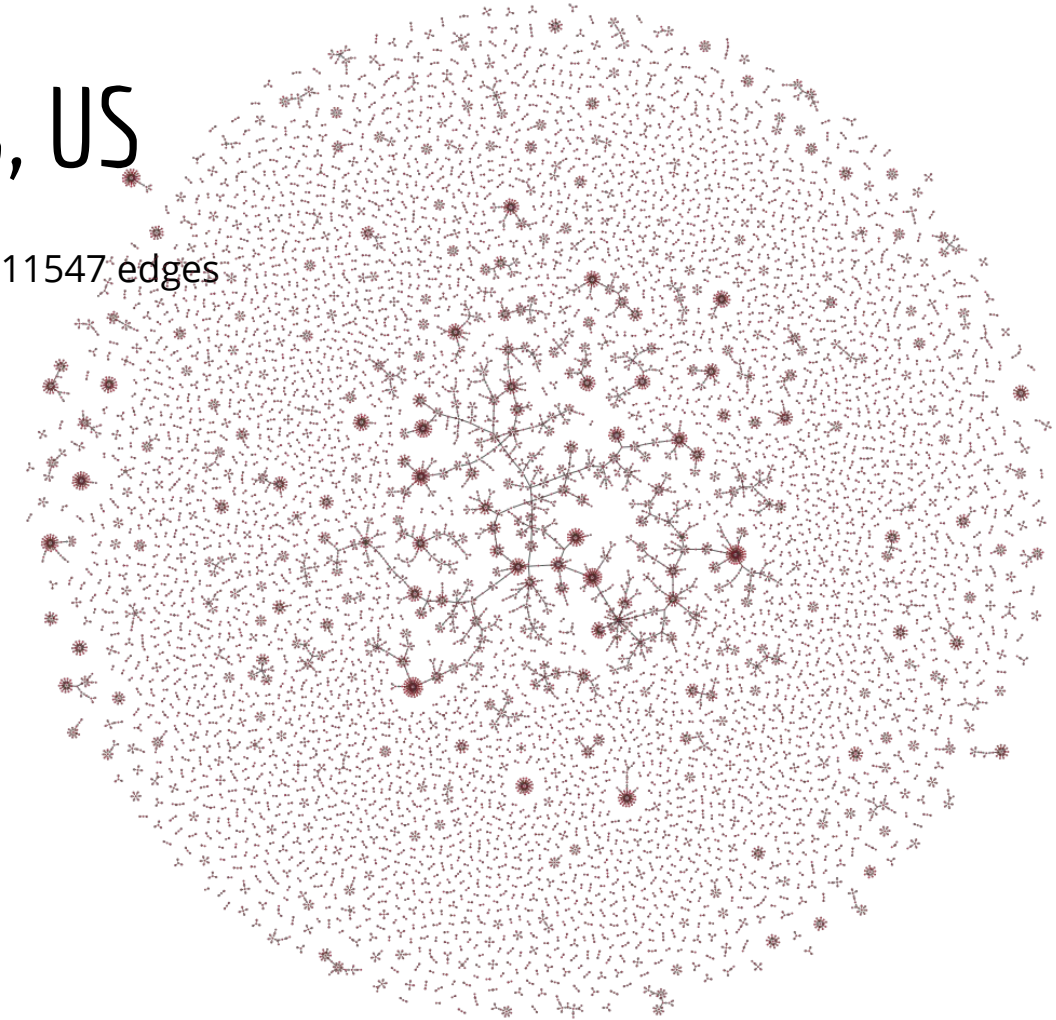
(1% of total ratings)



Las Vegas, US

Randomly selected 11547 edges

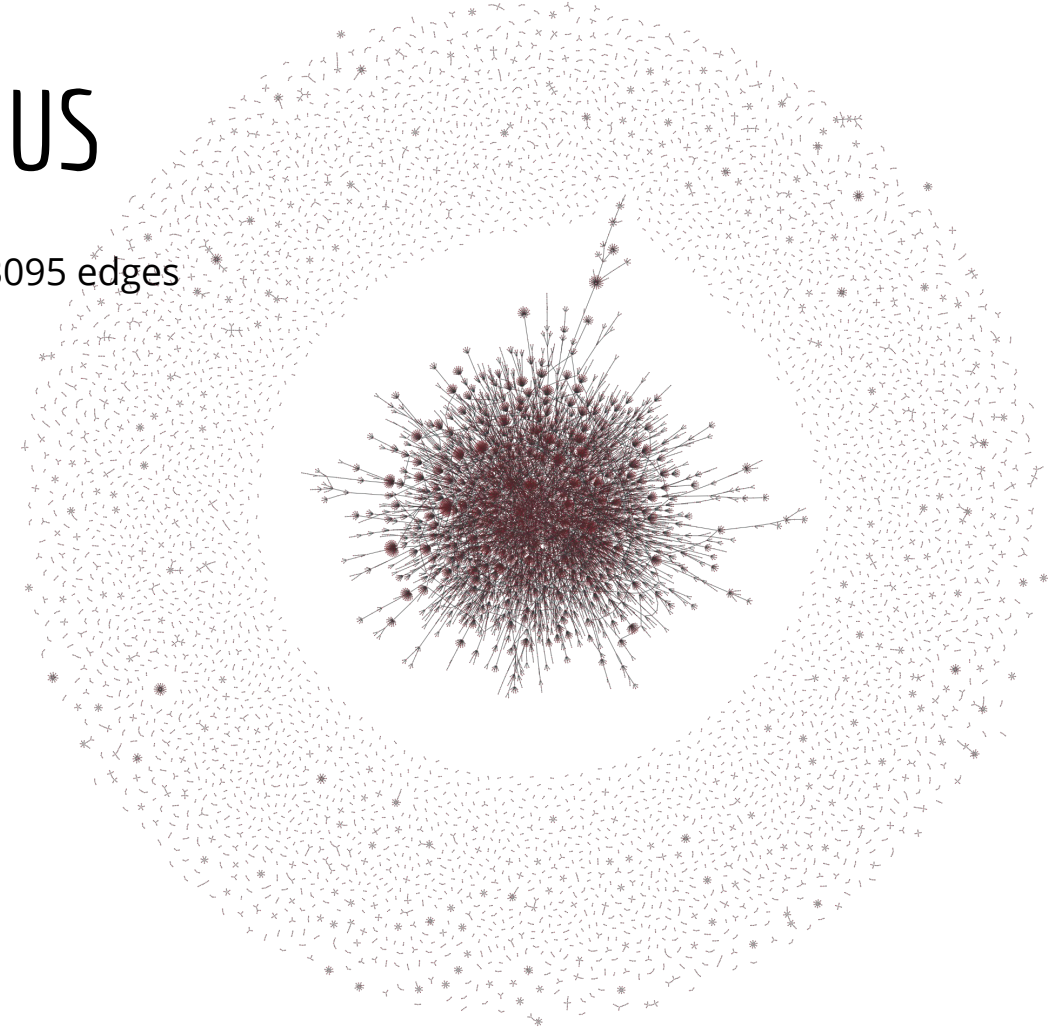
(1% of total ratings)



Las Vegas, US

Randomly selected 23095 edges

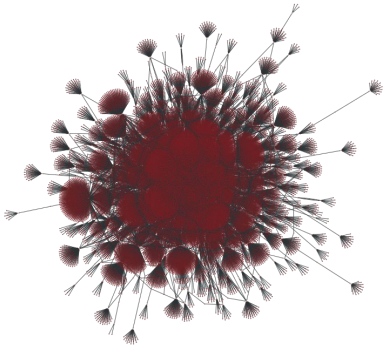
(2% of total ratings)

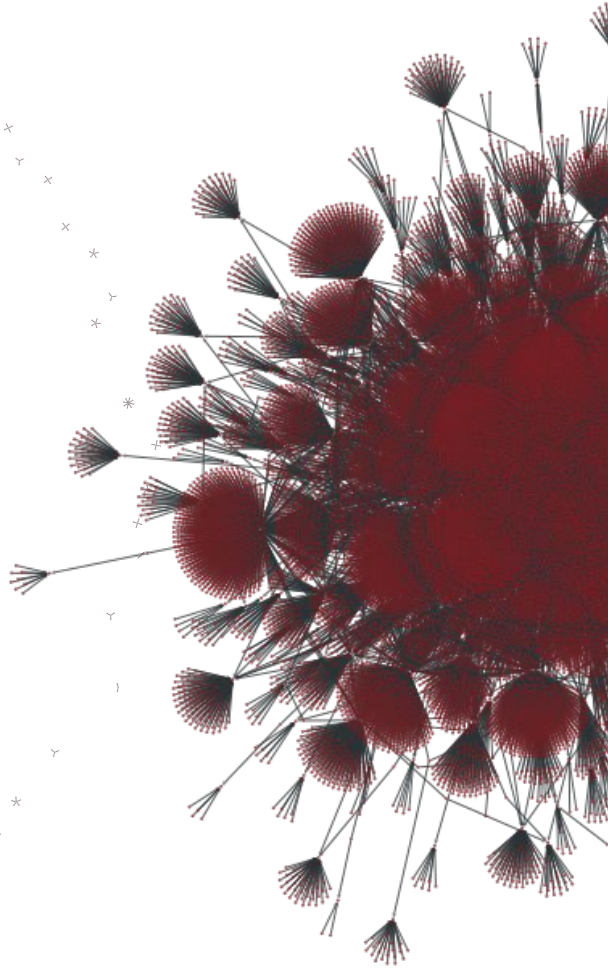


Las Vegas, US

Sequentially selected 23095 edges

(2% of total ratings)

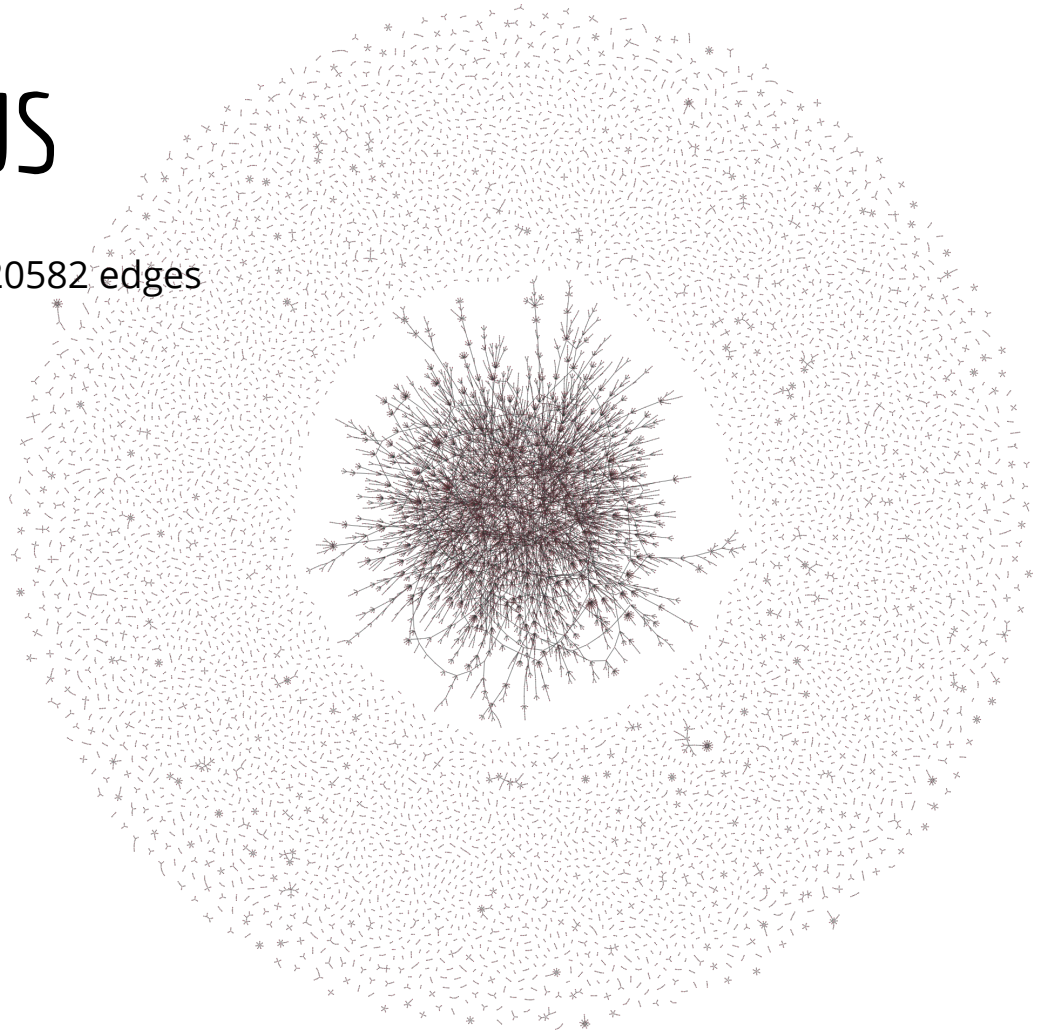




Phoenix, US

Randomly selected 20582 edges

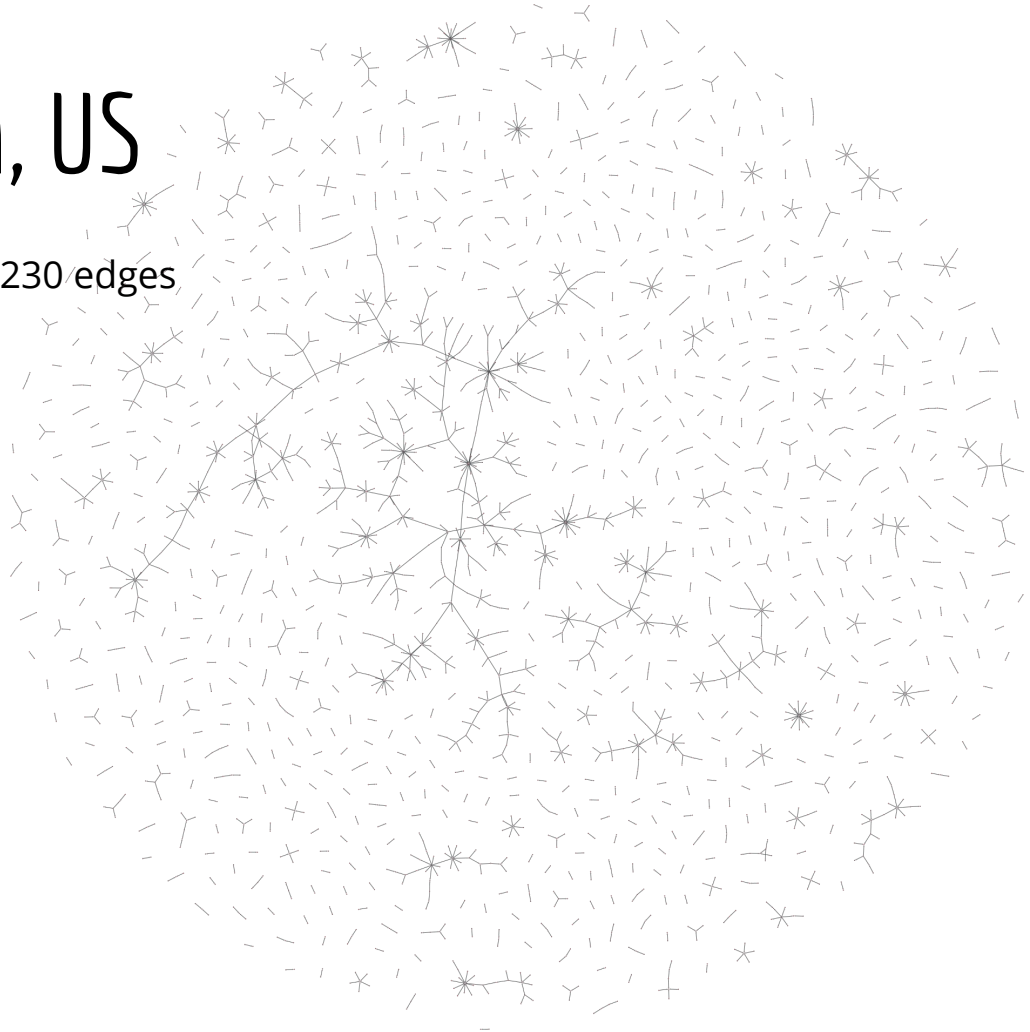
(2% of total ratings)



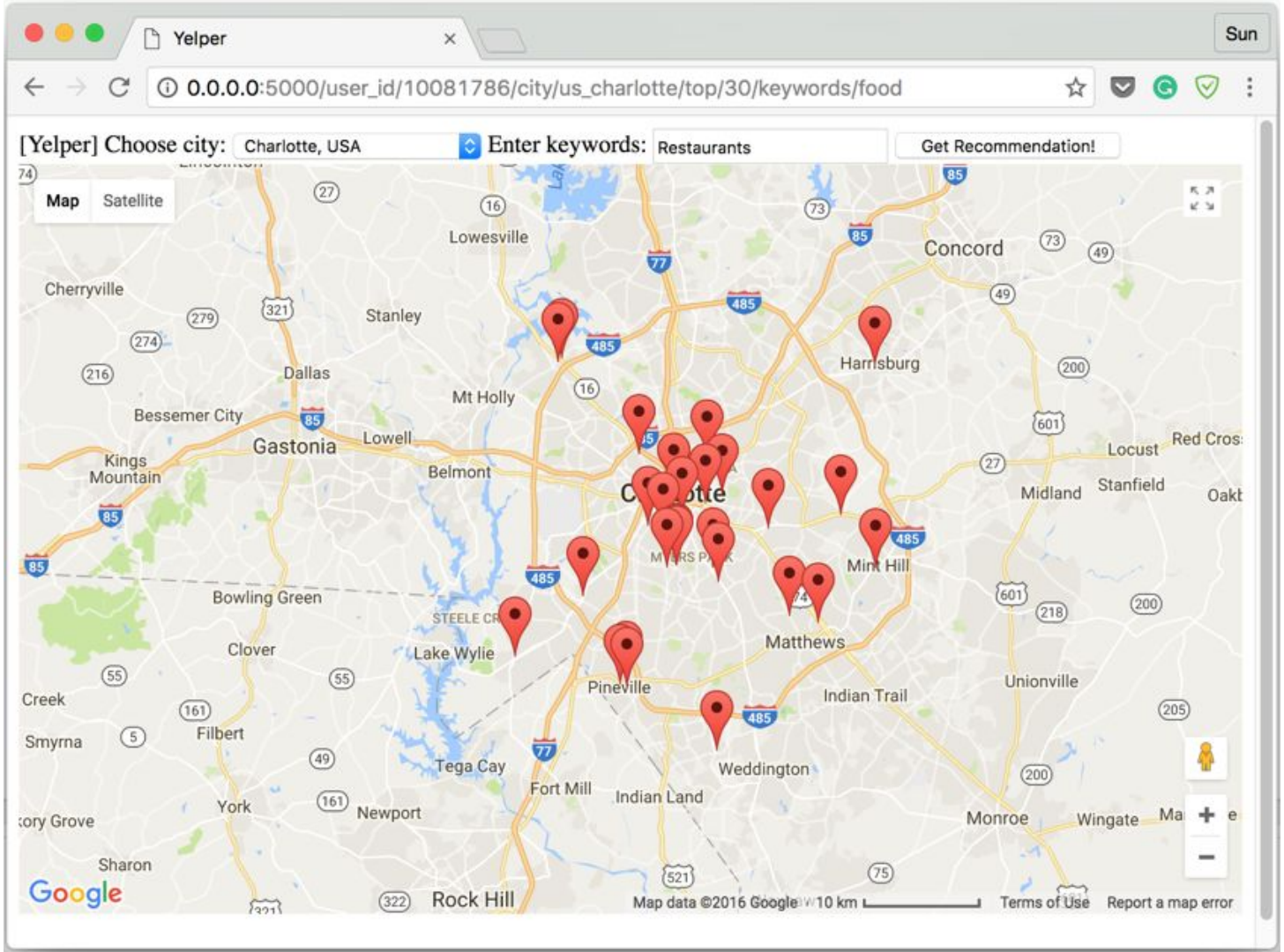
Pittsburgh, US

Randomly selected 2230 edges

(2% of total ratings)



Demo 2: Yelper recommendation page


Yelper
Sun
0.0.0.0:5000/user_id/10081786/city/us_charlotte/top/30/keywords/food
[Yelper] Choose city:
Charlotte, USA
Enter keywords:
Restaurants
Get Recommendation!
Map
Satellite
Lowesville
Concord
Cherryville
Stanley
Harrisburg
Dallas
Mt Holly
Bessemer City
Gastonia
Lowell
Belmont
Kings Mountain
Midland
Stanfield
Locust
Mint Hill
Bowling Green
Clover
Lake Wylie
Matthews
Pineville
Indian Trail
Unionville
Creek
Smyrna
Filbert
Tega Cay
Fort Mill
Indian Land
Weddington
Monroe
Wingate
York
Newport
Sharon
Rock Hill
Google
Map data ©2016 Google
10 km
Terms of Use
Report a map error

Demo 3: Simulation of user requests handling for recommendation using Spark Streaming and Kafka

bin/zookeeper-se... ⌘1 | bin/kafka-server-sta... ⌘2 | ..eek9/capstone (zsh) ⌘3 | ~ (zsh) ⌘4

..eek9/capstone (zsh)

```
16/09/20 08:03:36 INFO MemoryStore: Block broadcast_20 stored as values in memory (estimated size 8.
2 KB, free 365.5 MB)
16/09/20 08:03:36 INFO MemoryStore: Block broadcast_20_piece0 stored as bytes in memory (estimated s
ize 4.8 KB, free 365.5 MB)
16/09/20 08:03:36 INFO BlockManagerInfo: Added broadcast_20_piece0 in memory on 25.233.49.216:52863
(size: 4.8 KB, free: 366.2 MB)
16/09/20 08:03:36 INFO SparkContext: Created broadcast 20 from broadcast at DAGScheduler.scala:1012
16/09/20 08:03:36 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 16 (PythonRDD[46] a
t RDD at PythonRDD.scala:48)
16/09/20 08:03:36 INFO TaskSchedulerImpl: Adding task set 16.0 with 1 tasks
16/09/20 08:03:36 INFO TaskSetManager: Starting task 0.0 in stage 16.0 (TID 16, localhost, partition
 0, ANY, 7792 bytes)
16/09/20 08:03:36 INFO Executor: Running task 0.0 in stage 16.0 (TID 16)
16/09/20 08:03:36 INFO BlockManager: Found block input-0-1474373015200 locally
16/09/20 08:03:36 INFO PythonRunner: Times: total = 2, boot = -1496, init = 1498, finish = 0
16/09/20 08:03:36 INFO PythonRunner: Times: total = 5, boot = -1499, init = 1504, finish = 0
16/09/20 08:03:36 INFO Executor: Finished task 0.0 in stage 16.0 (TID 16). 1858 bytes result sent to
 driver
16/09/20 08:03:36 INFO TaskSetManager: Finished task 0.0 in stage 16.0 (TID 16) in 18 ms on localhos
t (1/1)
16/09/20 08:03:36 INFO TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed, from
 pool
16/09/20 08:03:36 INFO DAGScheduler: ResultStage 16 (runJob at PythonRDD.scala:441) finished in 0.02
1 s
16/09/20 08:03:36 INFO DAGScheduler: Job 16 finished: runJob at PythonRDD.scala:441, took 0.028218 s
-------------------------------------------
Time: 2016-09-20 08:03:36
-------------------------------------------
```

```
======= Yelper Recommended Results ========
user id: 36434133
city: us_phoenix
requested keywords: museum
Recommended business ids: [111, 329, 792, 15, 475, 435, 606, 288, 823, 678, 586, 909, 95, 115, 411,
347, 354, 600, 455, 566, 959, 304, 635, 162, 197, 366, 899, 769, 677, 140, 581, 520, 3, 302]



16/09/20 08:03:36 INFO JobScheduler: Finished job streaming job 1474373016000 ms.0 from job set of t
ime 1474373016000 ms
16/09/20 08:03:36 INFO JobScheduler: Total delay: 0.069 s for time 1474373016000 ms (execution: 0.05
3 s)
16/09/20 08:03:36 INFO PythonRDD: Removing RDD 42 from persistence list
16/09/20 08:03:36 INFO BlockManager: Removing RDD 42
16/09/20 08:03:36 INFO BlockRDD: Removing RDD 41 from persistence list
16/09/20 08:03:36 INFO BlockManager: Removing RDD 41
16/09/20 08:03:36 INFO KafkaInputDStream: Removing blocks of RDD BlockRDD[41] at createStream at Nat
iveMethodAccessorImpl.java:-2 of time 1474373016000 ms
16/09/20 08:03:36 INFO ReceivedBlockTracker: Deleting batches: 1474373013000 ms
16/09/20 08:03:36 INFO InputInfoTracker: remove old batch metadata: 1474373013000 ms
16/09/20 08:03:36 INFO BlockManagerInfo: Removed input-0-1474373013600 on 25.233.49.216:52863 in mem
ory (size: 174.0 B, free: 366.2 MB)
16/09/20 08:03:36 INFO MemoryStore: Block input-0-1474373016400 stored as bytes in memory (estimated
 size 166.0 B, free 365.5 MB)
16/09/20 08:03:36 INFO BlockManagerInfo: Added input-0-1474373016400 in memory on 25.233.49.216:5286
3 (size: 166.0 B, free: 366.2 MB)
16/09/20 08:03:36 WARN BlockManager: Block input-0-1474373016400 replicated to only 0 peer(s) instea
d of 1 peers
16/09/20 08:03:36 INFO BlockGenerator: Pushed block input-0-1474373016400
^CTraceback (most recent call last):
  File "/Users/sundeepblue/Bootcamp/allweek/week9/capstone/./handle_user_requests_streaming.py", lin
e 89, in <module>
    ssc.awaitTermination()
  File "/usr/local/Cellar/apache-spark/2.0.0/libexec/python/lib/pyspark.zip/pyspark/streaming/contex
```

..eek9/capstone (zsh)

```
User 74995577 wanted recommendation for walmart in city us_pittsburgh
User 9436765 wanted recommendation for hotel in city us_pittsburgh
User 79374837 wanted recommendation for exhibition in city us_phoenix
User 24739764 wanted recommendation for chinese foodbest ice cream in city us_phoe
nix
User 36844896 wanted recommendation for sight seeing in city us_lasvegas
User 91645901 wanted recommendation for chinese foodbest ice cream in city germany
_karlsruhe
User 69497998 wanted recommendation for macys in city uk_edinburgh
User 77154767 wanted recommendation for best indian food in city us_pittsburgh
User 94675065 wanted recommendation for shopping in city us_urbana_champaign
User 37416352 wanted recommendation for movie in city uk_edinburgh
User 32401061 wanted recommendation for best indian food in city us_lasvegas
User 75837638 wanted recommendation for movie in city us_urbana_champaign
User 54045496 wanted recommendation for chinese foodbest ice cream in city us_lasv
egas
User 6309513 wanted recommendation for museum in city us_madison
User 82906313 wanted recommendation for club in city us_urbana_champaign
User 55588605 wanted recommendation for best indian food in city us_madison
User 95375423 wanted recommendation for library in city us_lasvegas
User 12831786 wanted recommendation for restaurants in city us_lasvegas
User 61342611 wanted recommendation for movie in city us_phoenix
User 23323476 wanted recommendation for museum in city us_pittsburgh
User 2393371 wanted recommendation for eat in city us_madison
User 6886187 wanted recommendation for bookstore in city germany_karlsruhe
User 63471081 wanted recommendation for shopping in city us_charlotte
User 85315426 wanted recommendation for walmart in city us_madison
User 84501112 wanted recommendation for cinerma in city us_lasvegas
User 20928159 wanted recommendation for macys in city canada_montreal
User 33434204 wanted recommendation for exhibition in city us_charlotte
User 66505930 wanted recommendation for shopping in city us_madison
User 28593932 wanted recommendation for best indian food in city us_pittsburgh
User 91433509 wanted recommendation for chinese foodbest ice cream in city canada_
montreal
User 6384487 wanted recommendation for museum in city us_urbana_champaign
User 19767379 wanted recommendation for shopping in city us_charlotte
User 58086703 wanted recommendation for chinese foodbest ice cream in city us_urba
na_champaign
User 30561976 wanted recommendation for exhibition in city uk_edinburgh
User 5245242 wanted recommendation for sight seeing in city canada_montreal
User 14326179 wanted recommendation for club in city germany_karlsruhe
User 89331194 wanted recommendation for club in city us_pittsburgh
User 29274283 wanted recommendation for shopping center in city us_phoenix
User 83618317 wanted recommendation for restaurants in city us_charlotte
User 6056581 wanted recommendation for club in city us_pittsburgh
User 72328113 wanted recommendation for hotel in city us_pittsburgh
User 41214122 wanted recommendation for eat in city us_urbana_champaign
User 85472372 wanted recommendation for best indian food in city germany_karlsruhe
User 67629068 wanted recommendation for sight seeing in city us_lasvegas
User 93248529 wanted recommendation for sight seeing in city canada_montreal
User 57174247 wanted recommendation for club in city us_lasvegas
user-request-topic
0
1607
➜  capstone clear

➜  capstone python ./user_requests_producer.py
➜  capstone python ./user_requests_producer.py
➜  capstone python ./user_requests_producer.py
User 43079451 wanted recommendation for chinese foodbest ice cream in city us_char
```

Agenda

What?
What are the components for Yelper?

Demo
User-business network; Yelper main page; Simulation of users' recommendation requests handling

Why?
Why we need recommendation?

How?
How to build Yelper?

Summary & Future works
What we learn and what's next?

Summary & future works

- Summary
  - Preprocessing by dividing business data by cities to allow fine tuned and customized recommendations
  - Collaborative filtering based recommendation using Spark MLlib
  - User-business graph visualization using D3 and graph-tool library
  - User-business graph analysis using Spark GraphX in Scala
  - Real-time user request handling simulation using Spark Streaming and Apache Kafka
  - Google Map view to recommend high rated businesses for users
- What we learn
  - Streaming data analysis and mining could be the new norm in the future and we as Data Scientist should be prepared for it
- Future works
  - More graph analysis
    - Graph pagerank analysis using GraphX
    - Community discovery (similar to Facebook social network)
  - Improve recommendation
    - Content-based recommendation
    - Clustering all businesses
    - Extract object from business photos using Convolutional Neural Network

Thanks!