Analysing Foursquare Tips

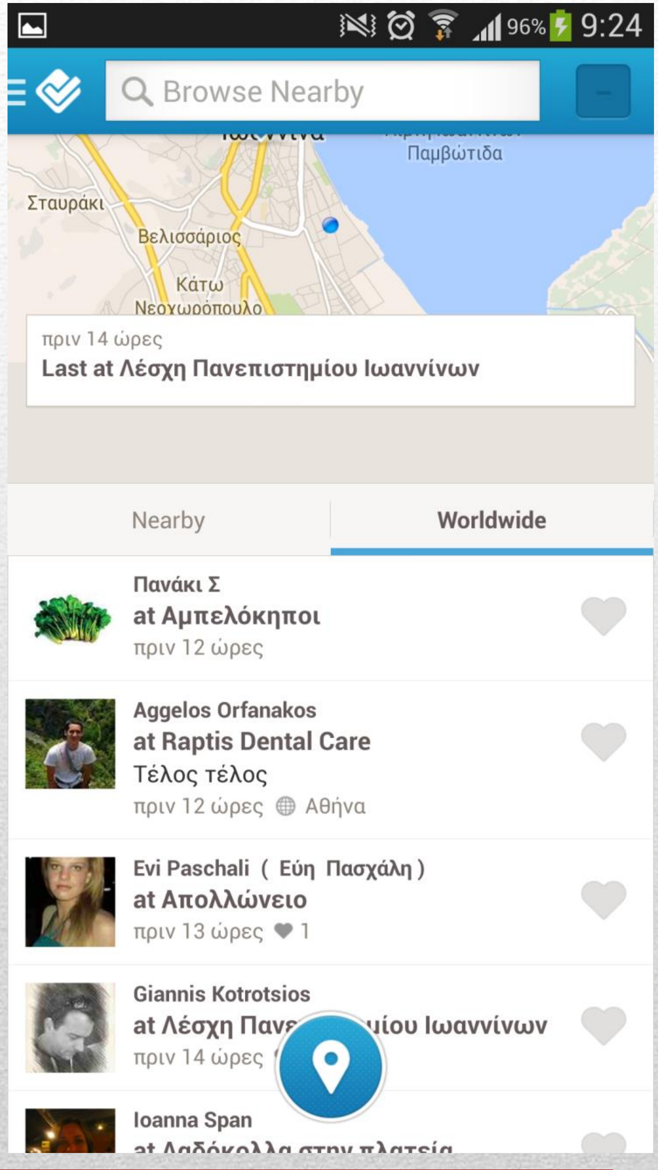
Kotrotsios Ioannis
MSc Student, University of Ioannina
ikotrots@cs.uoi.gr

➢ Location-based Online Social Networks

- Applications which allow users to interact, share their locations, meet up, recommend places based on their physical location etc.
- LBSN’s bridge the gap between online social networking and the physical world
- A smartphone or tablet with GPS is usually needed.

➢ About Foursquare

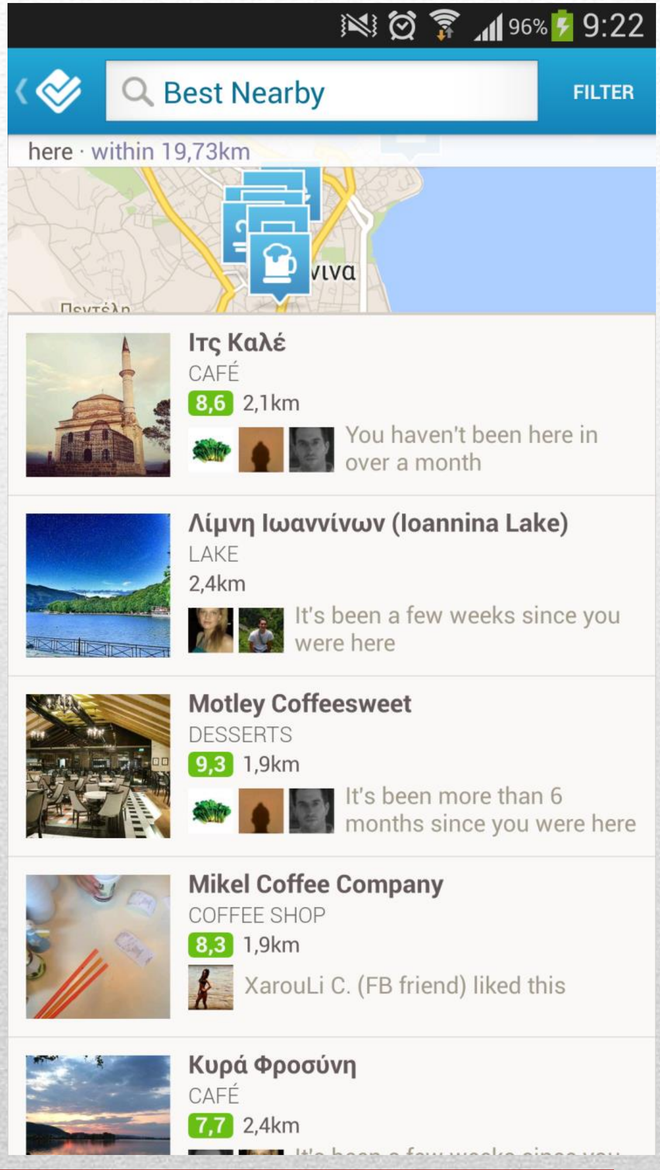
- Free Mobile and web application
- Connect with friends
- Check in to places you visit
- Leave tips and read other users’ tips
- Explore new places
- Get deals
- Earn points, budges and mayorships
- Get personalized recommendations
- And many more…



➢ **Foursquare in numbers**

- Launched in 2009
- Over 40 million people worldwide
- Over 4.5 billion check-ins, with millions more every day
- Businesses: Over 1.5 million using the Merchant Platform
- Employees: Over 160

- Foursquare can constitute a great marketing opportunity for local businesses.



- **Foursquare Merchant Platform**

- Free tool to help businesses attract new customers or reward current ones
- Access to venue statistics
- Foursquare specials, discounts, prizes
- Offered when users check-in at or near the venue

➢ Foursquare API

- Third-party applications can connect with Foursquare and extend the official application functionality
- Developers can get access and collect valuable data for analysis
- Some rate and content limits exist (no access to users' check-ins)

- **Some of the things are we trying to do**

- Identify interesting facts about users' habits in a location
- Provide the user more personalized recommendations
- Predict the tip that a user may leave to a venue

- All of the above using mainly the tips that have been already left by the users

➢ Talk Outline

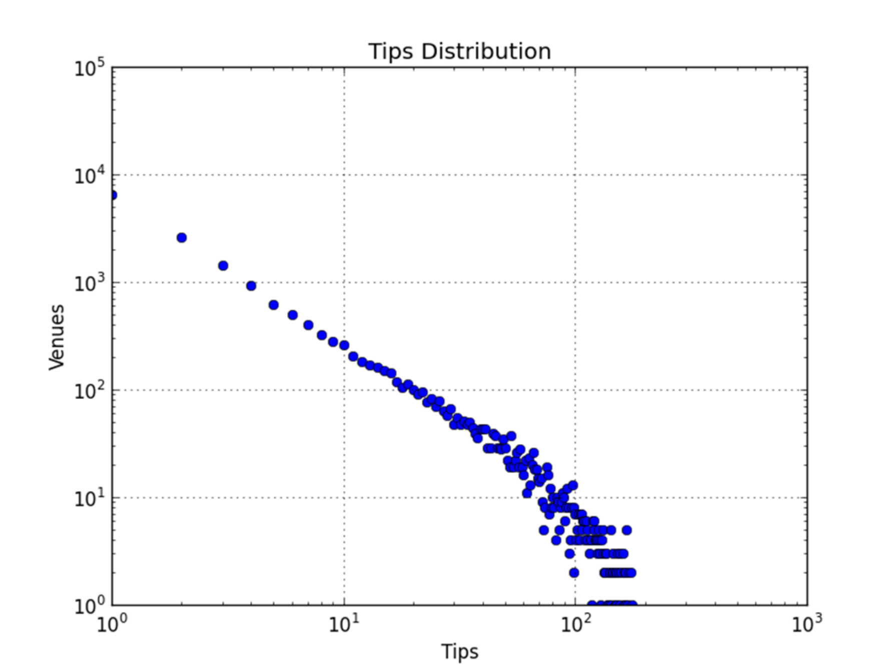
- Information about the collected data
- General Statistics
- Working with tips
- Tagging users and venues
- Use of topic modeling on tips
- Conclusions

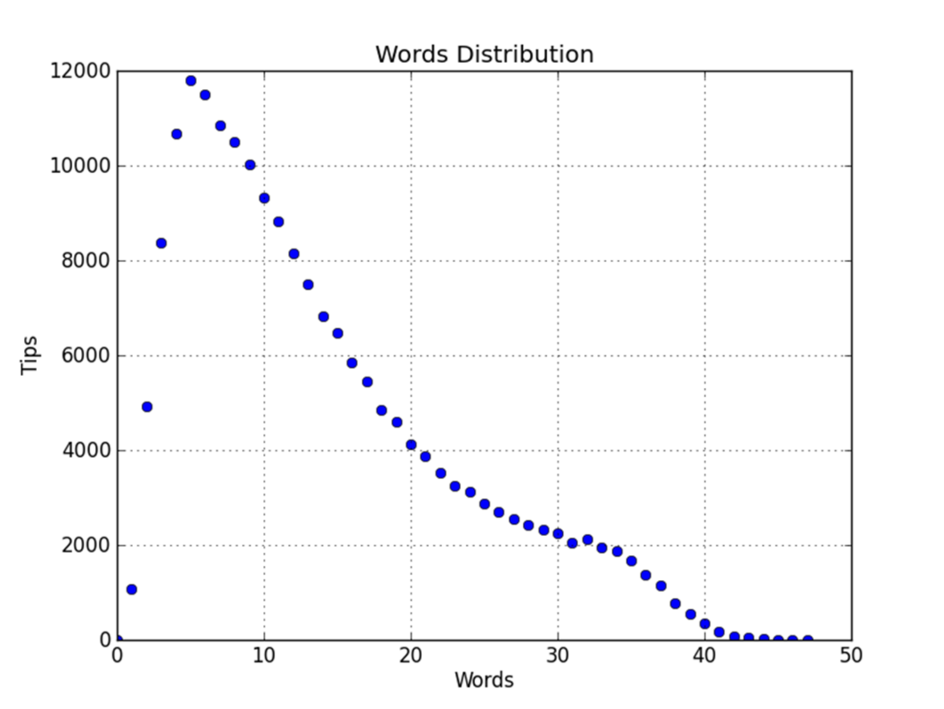
➢ Collected Data

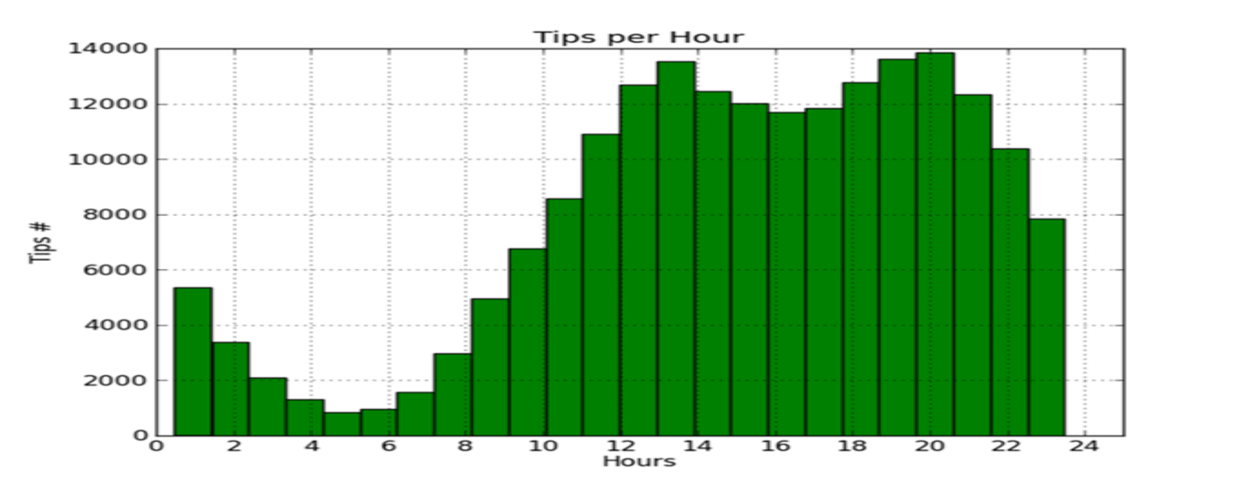
- Using Foursquare API we collected venue, mayor and tip data from a region of New York City (Manhattan)

- About 41.200 venues with 195.000 tips (2,7 million words)
- About 18.200 mayors with 170.600 tips (2,1 million words)

➢ General Data Statistics



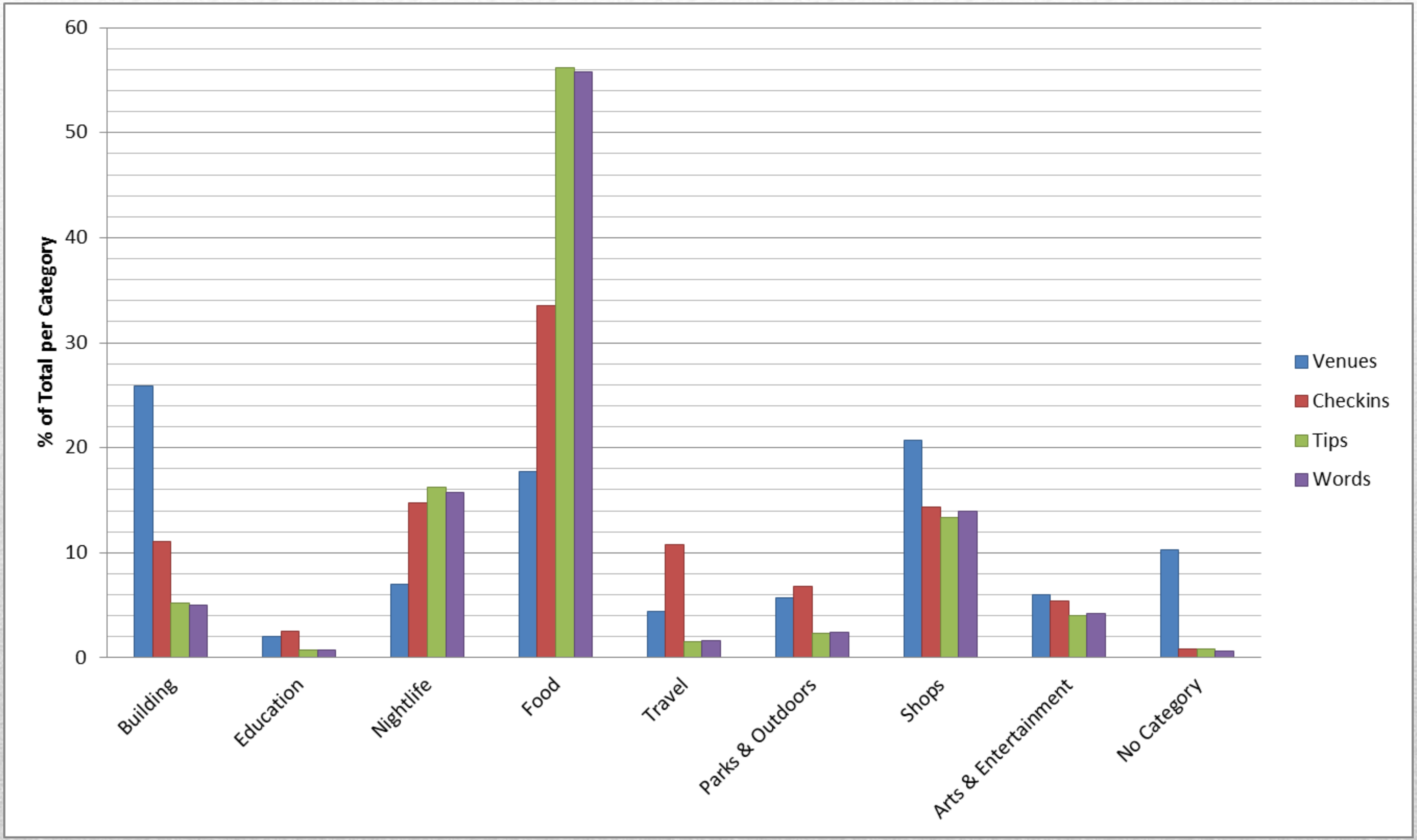




➢ Statistics Per Venue Category

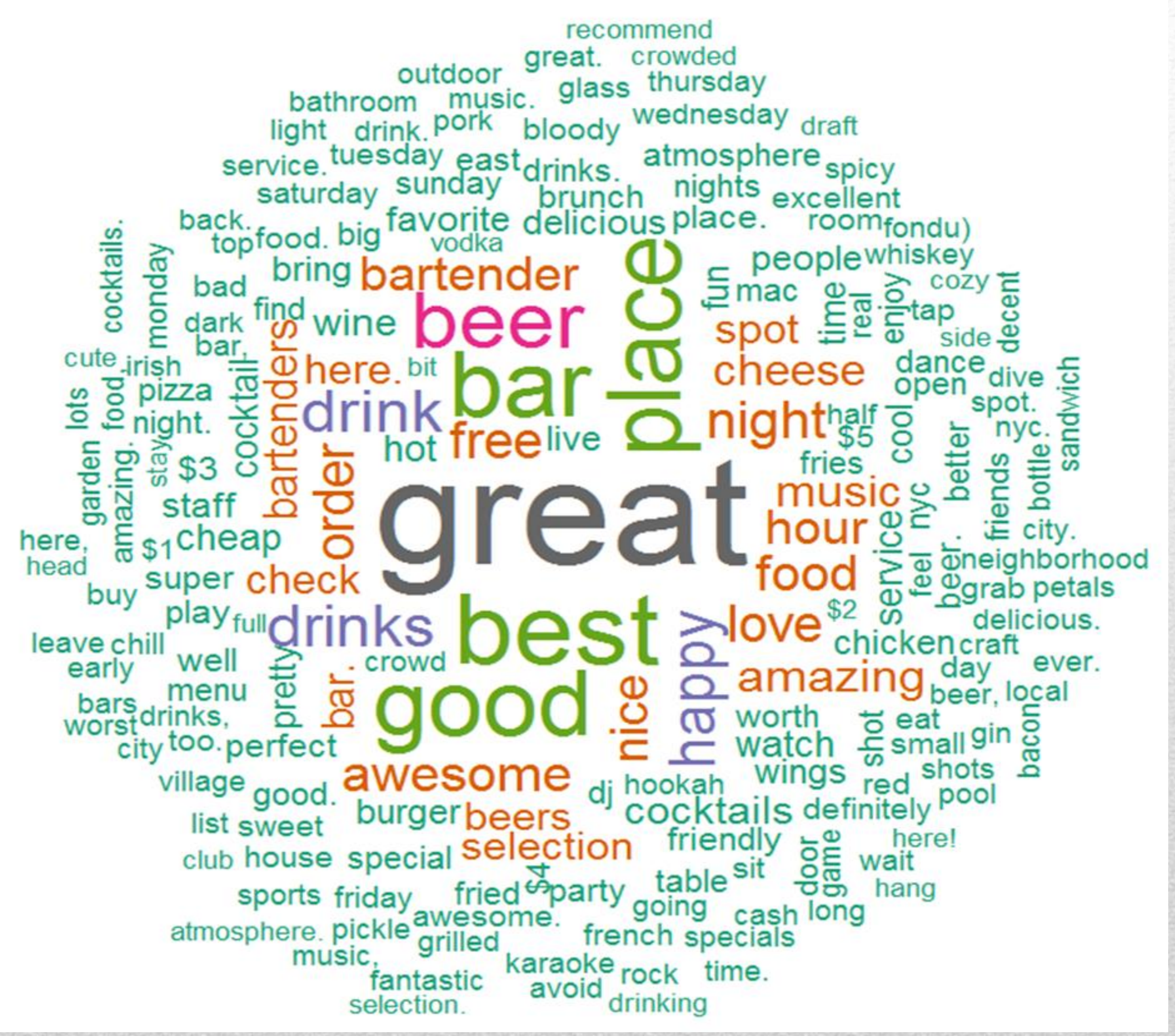
Categories	Venues	%	Checkins	%	Tips	%	Words	%
Building	10.649	25,9	2.889.958	11,1	10.074	5,2	132.853	5,0
Education	849	2,0	638.596	2,5	1.327	0,7	18.463	0,7
Nightlife	2.897	7,0	3.843.434	14,8	31.371	16,2	421.266	15,7
Food	7.293	17,7	8.672.397	33,5	108.743	56,2	1.493.946	55,8
Travel	1.827	4,4	2.787.081	10,8	2.843	1,5	42.898	1,6
Parks & Outdoors	2.364	5,7	1.753.723	6,8	4.491	2,3	63.328	2,4
Shops	8.525	20,7	3.744.627	14,4	25.870	13,4	373.894	14,0
Arts & Entertainment	2.504	6,0	1.393.044	5,4	7.657	4,0	112.369	4,2
No Category	4.253	10,3	202.624	0,8	1.478	0,8	16.210	0,6
Total	**41.161**		**25.925.484**		**193.854**		**2.675.227**	

➢ Statistics Per Venue Category



- **Most Common Words (Food Category)**

- **Most Common Words (Nightlife Category)**



- **Now we are going to work with tips…**

Dataset Selection

- As we needed sufficient tip data we selected to work with a small dataset of users and venues that had more than 50 tips.
- About 950 venues with 52.600 tips and 560 users with 87.800 tips
- Intersection between venue and mayor tips: 6.023 tips

Tip Words Clean up

- Conversion to lower
- Punctuation and unnecessary marks removal
- Urls and numbers removal
- Noun conversion to singular
- Stopwords removal

- We needed to define similarity/build model for users and venues

- How?
- Using characteristic words in tips to tag each user and venue
- Constructing topics from words in tips and use them as intermediate

Creation of venue and User word tags

- ***tf–idf***, *term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection or corpus*

- Used tf-idf to score words in every venue and every user

- <u>tf:</u>
  - *Word appearances in the venue or user*
- <u>idf:</u>
  - *Number of venues that their tips include the word*

- Used words with top-N scores as tags for each venue and user

- **We attempted to match every user to a "similar" venue based on the previous word tags**
- **We considered the matching successful if the user has indeed left a tip at the matched venue and then calculated an overall success percentage**

- If the matching is done randomly the success percentage is about 1,1%
- If we match every user with the venue that has the larger tag intersection the percentage is about 31,4%
- If we take the complete tag word lists with the *tf-idf* scores, compute the cosine similarity and match every user with the top venue (larger cos similarity) the percentage is about 28,5%

- Then we removed all the user tips (about 6.023) from the initial venue tip dataset, recomputed the venue tags and repeated the last two matchings. The success percentages were 8,7% and 15,5% respectively

Topic Modeling

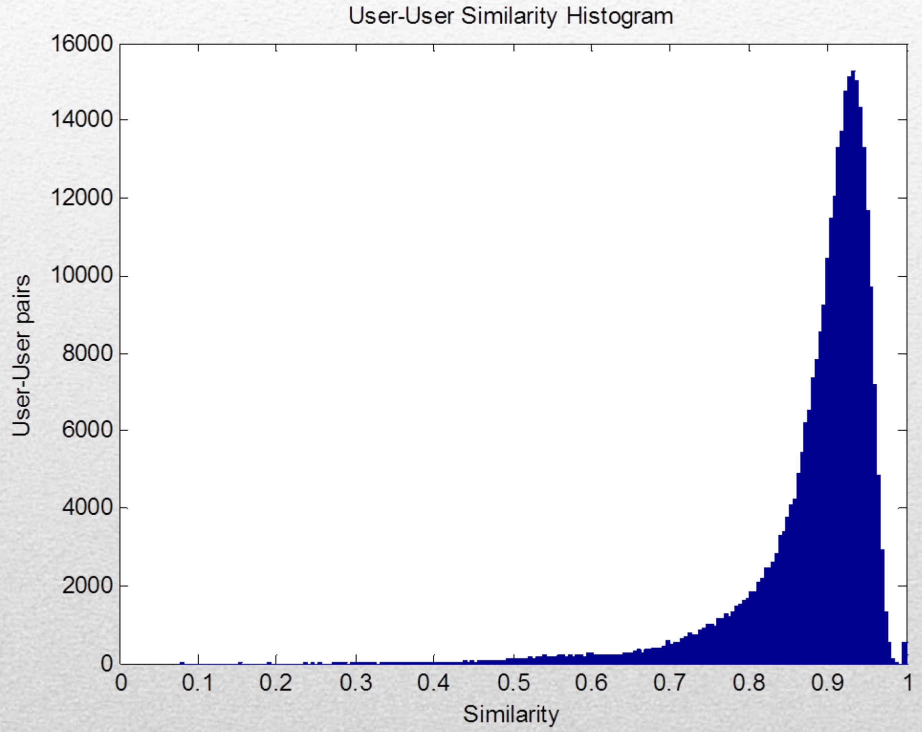
- *Topic models provide a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that frequently occur together.*

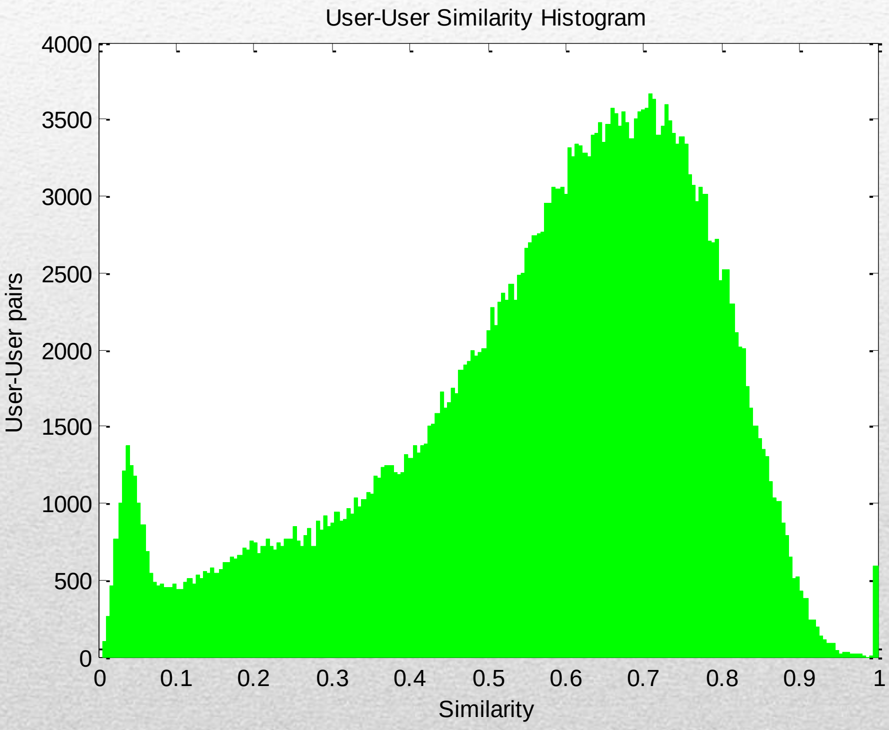
- We used the MALLET topic modeling tool in order to create topics and use them to identify similarities between venues and users.

- We created one document for every venue with all their tips together and used them to build a MALLET topic model
- Then we did the same with the user tips and performed topic inference using the above model
- Additionally, we built another topic model this time with the documents created from user tip data.

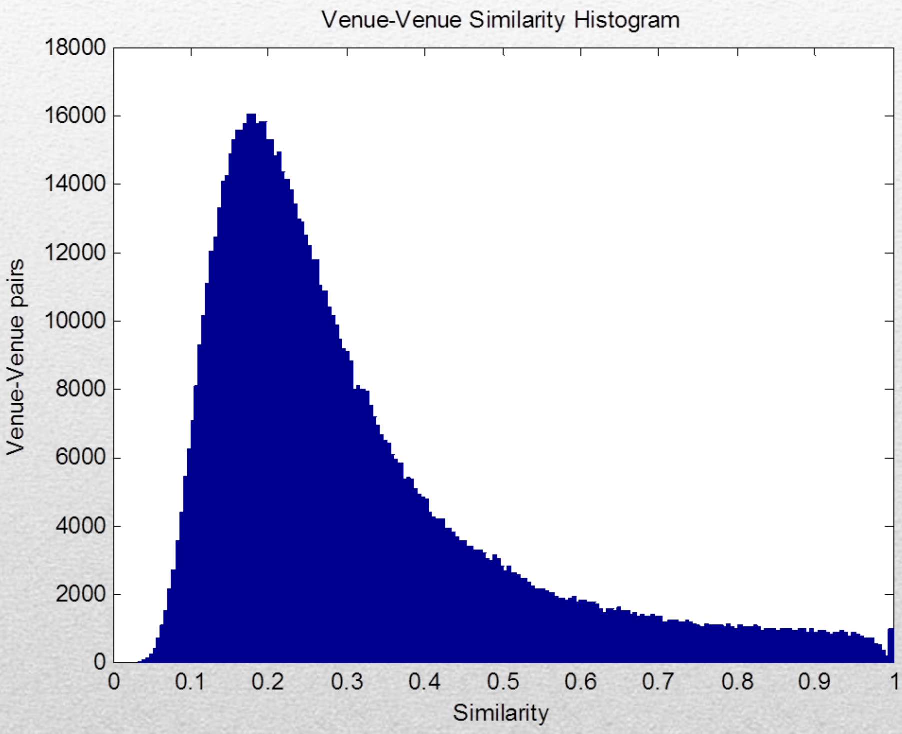
- **Topic Modeling**
- The Mallet topic modeling tool outputs one vector for each document which represents the topic composition of the document
- We used these vectors to compute various similarities between users and venues in order to verify if we can use topic models for deciding if a venue or user is similar to an other venue or user
- All of the similarities in the next slides were computed using the very common cosine similarity

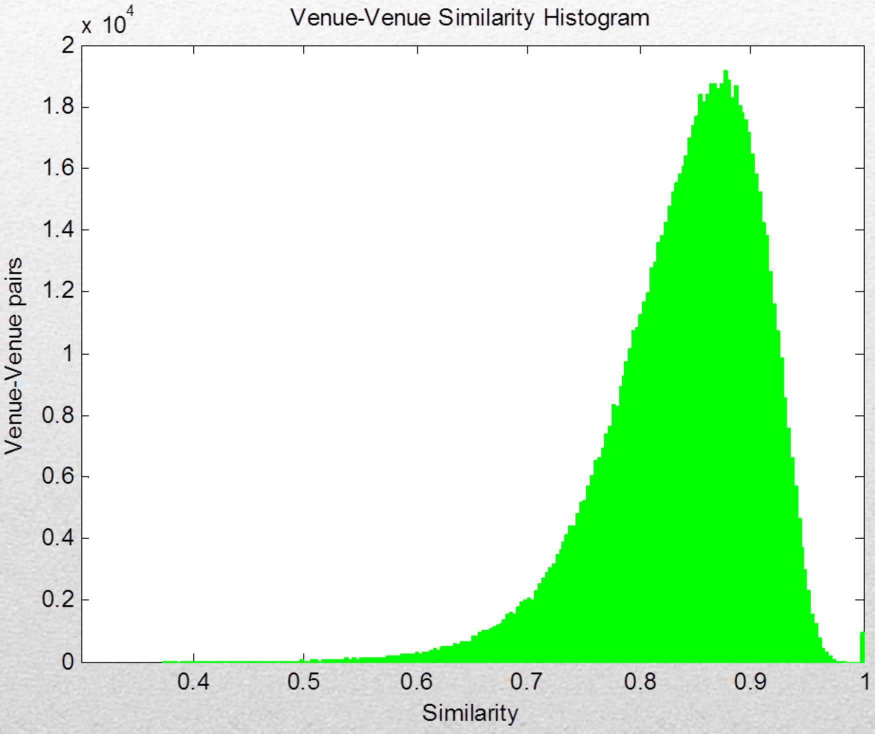
- User - User Similarity Histograms
- Blue: venue data model, Green: user data model)



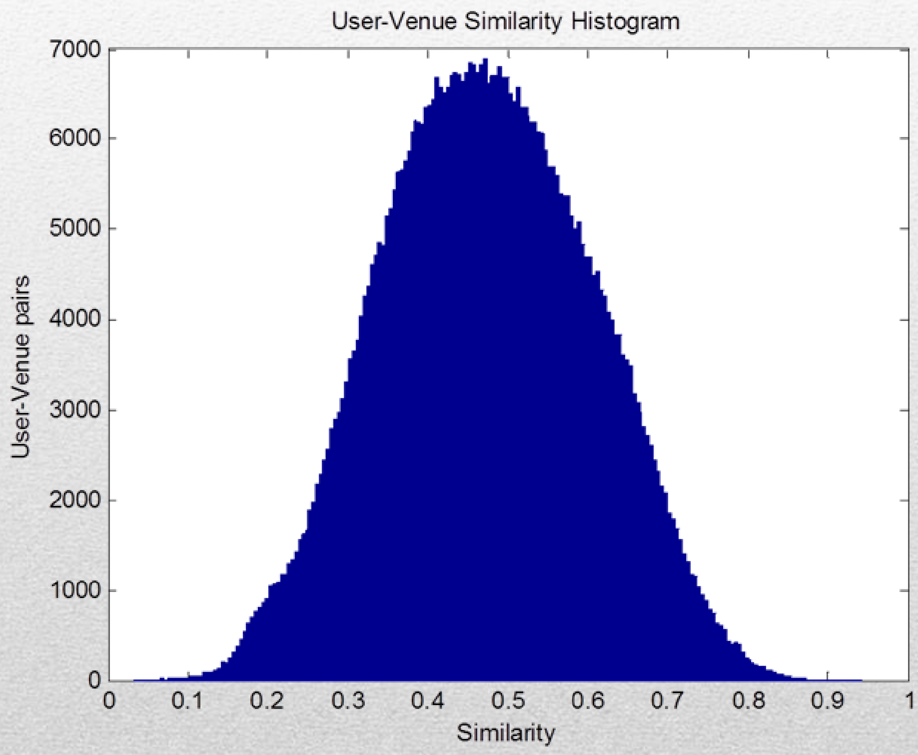


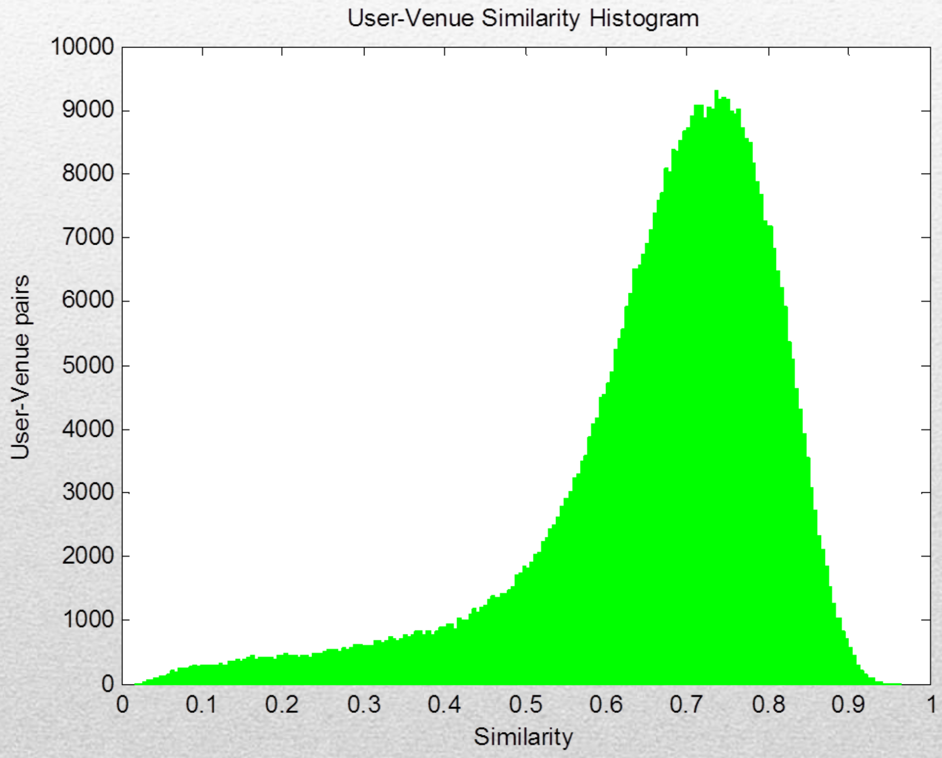
- Venue - Venue Similarity Histograms
- Blue: venue data model, Green: user data model)



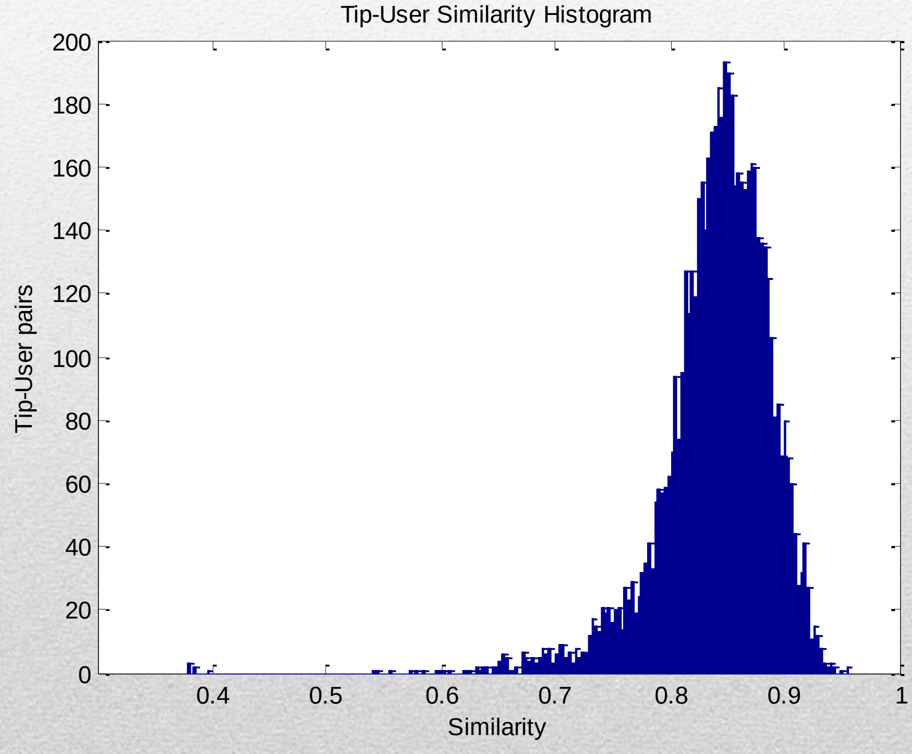


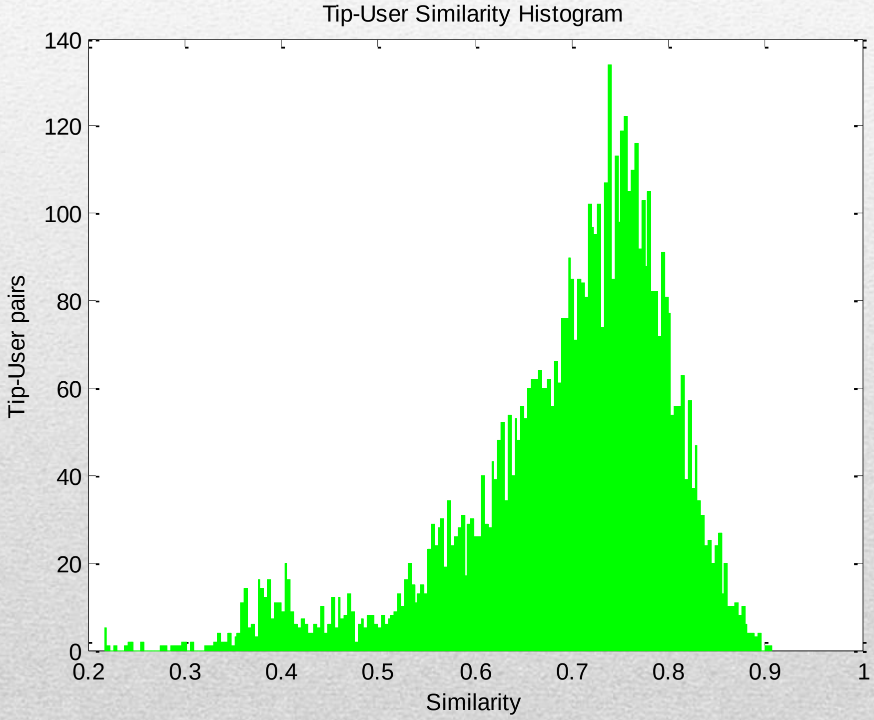
- User - Venue Similarity Histograms
- Blue: venue data model, Green: user data model)



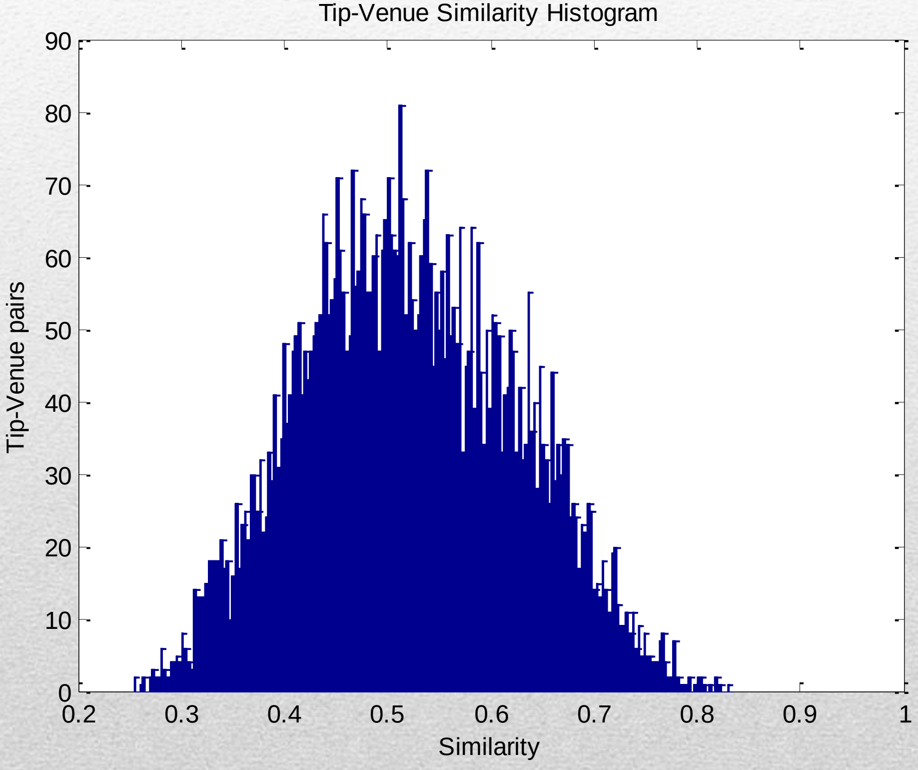


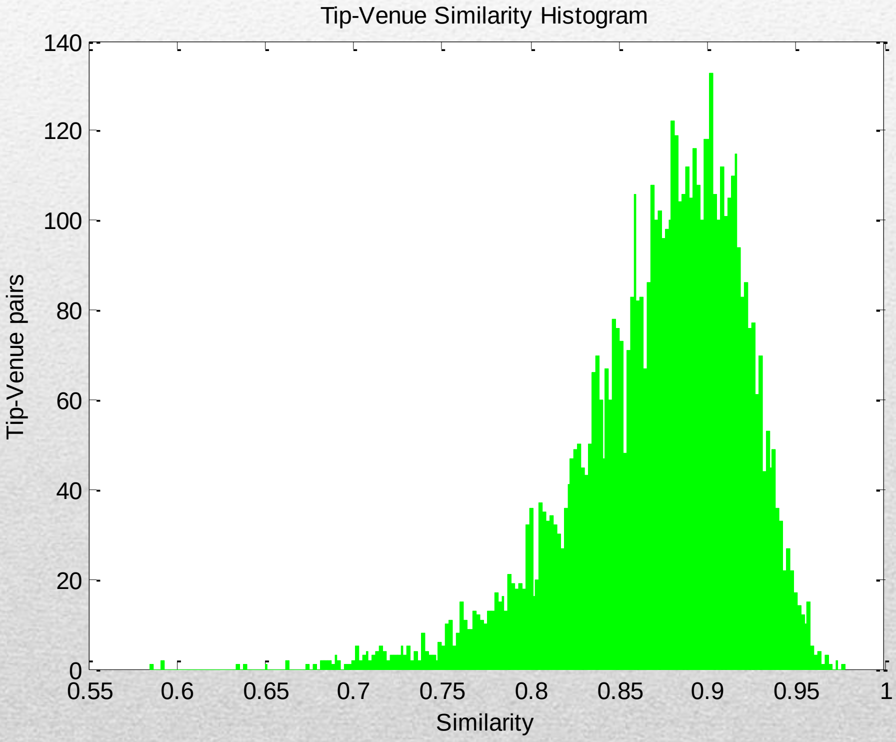
- For each tip that we knew its corresponding user and venue we computed the similarities with them





- Tip – Venues similarity Histograms





- **Conclusions about topic models**

- Seems that they can be used to distinguish between users and venues
- They do not work out of the box
- Further work should be done to improve the quality of the topics

- **Next steps into improving topics**

- Work with input data before constructing topics
- Use only nouns, give different weights to words
- Combine different methods/models

- Thank you