Prediction of User Appreciation of Yelp's Businesses Based on Text Reviews Hidden Features

Author: Umberto Di Fabrizio
UIN: 651053197
Author: Vittorio Selo
UIN: 657698682

CS 521 -Statistical Natural Language Processing

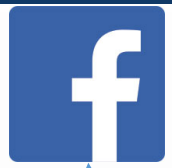
Outline

- Introduction
- Data Analyses
- Approach
- Experiment Results
- Conclusion

Outline

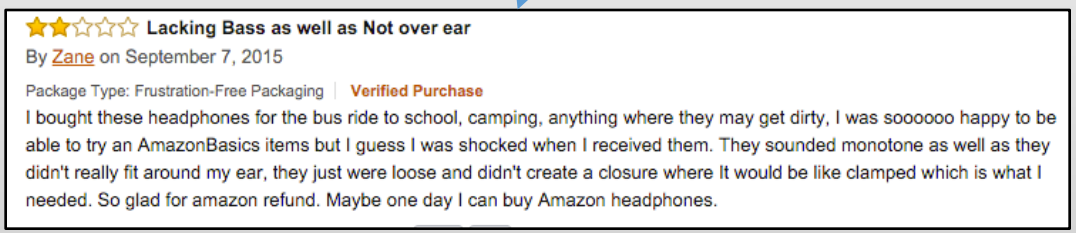
- Introduction
- Data Analyses
- Approach
- Experiment Results
- Conclusion

Introduction(1/3)



NETFLIX

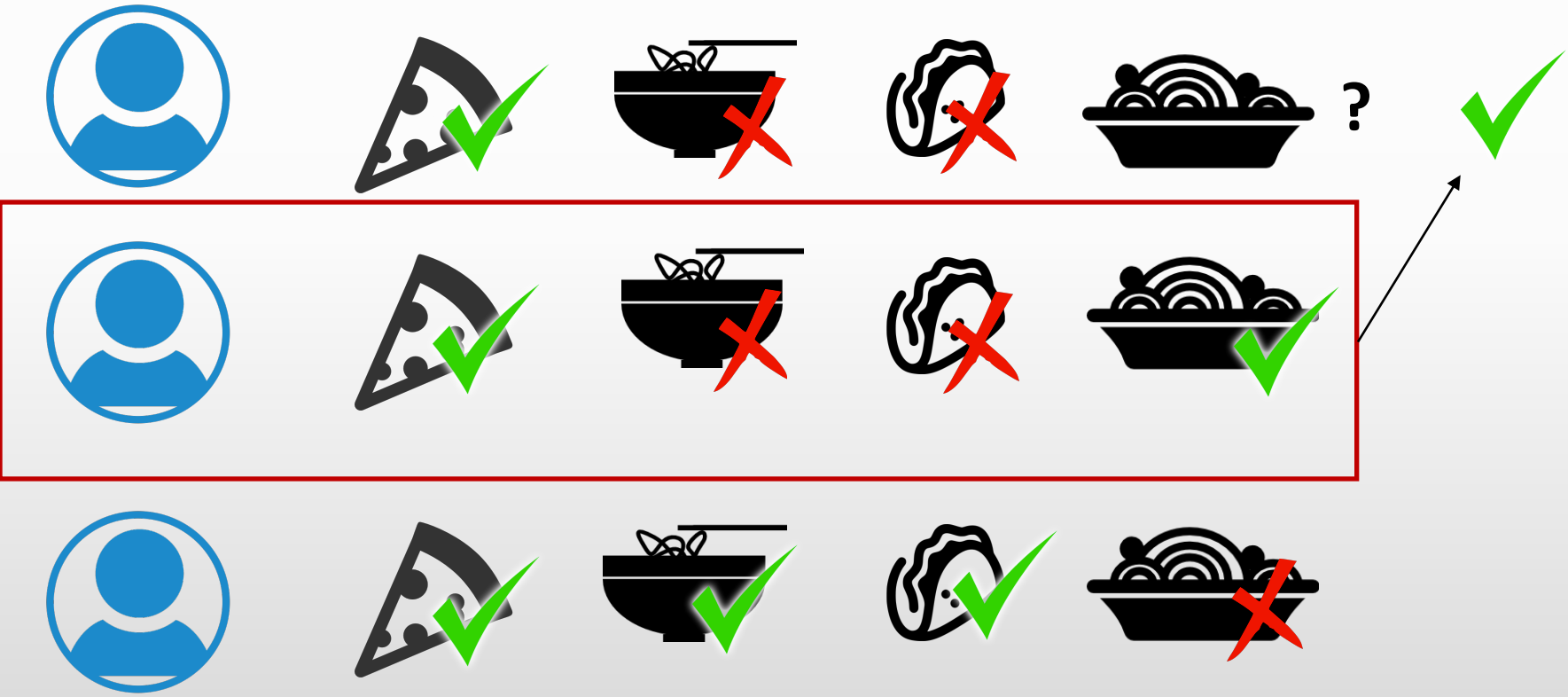
subscribe



Introduction(2/3)

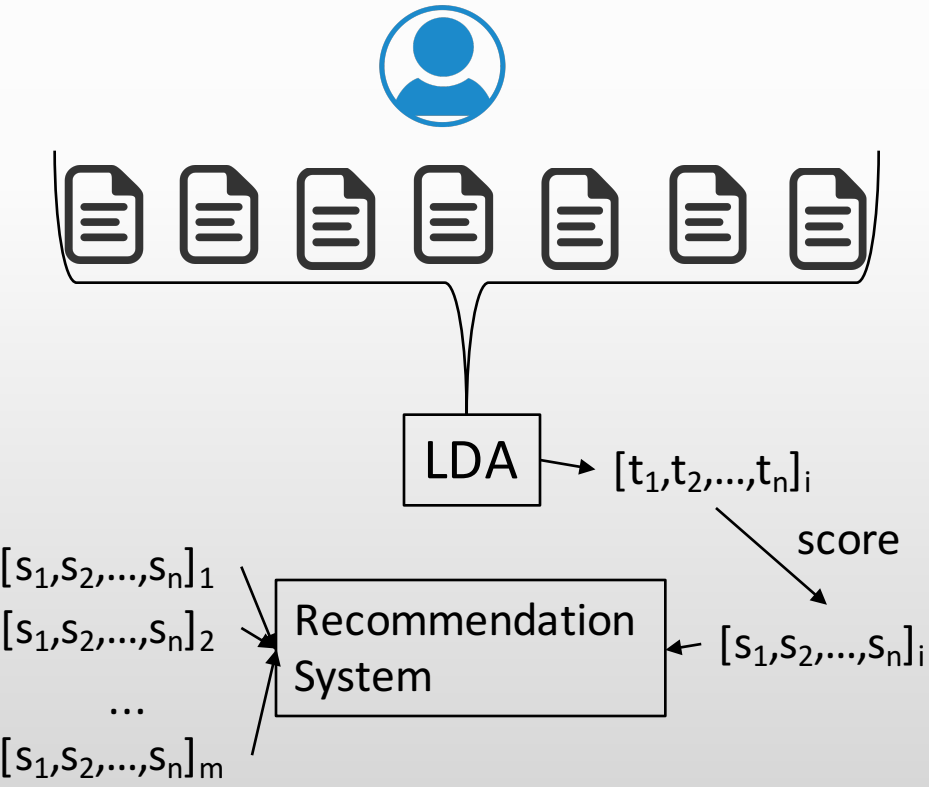
A key tool for the *Social Network Services* is a good Recommendation System

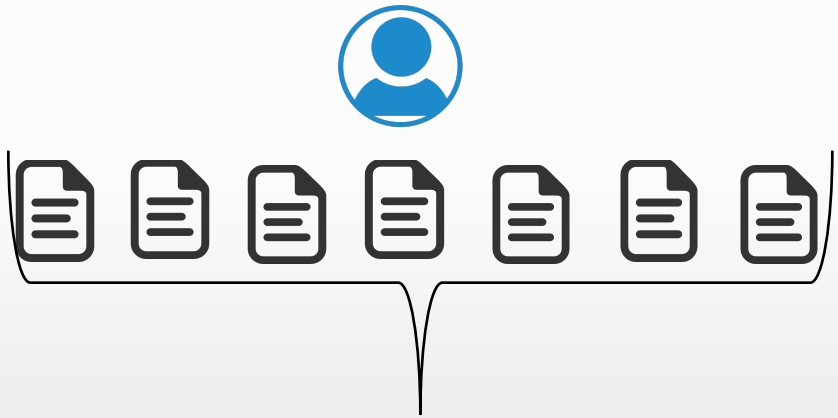
But… what is a Recommendation System?



Introduction(3/3)

Most of the websites are based on *user-generated content*
It means we have a lot of text information to exploit
The goal is to extract tastes of the users in order to enhance the performance of recommendation systems





Hidden-features

Personal tastes of an user which can be even unconscious for the user himself

Predict what the user likes and what not

Outline

- Introduction
- Data Analyses
- Approach
- Experiment Results
- Conclusion

Data Analyses(1/1)

We have used the Yelp Dataset Challenge[1] that is public available

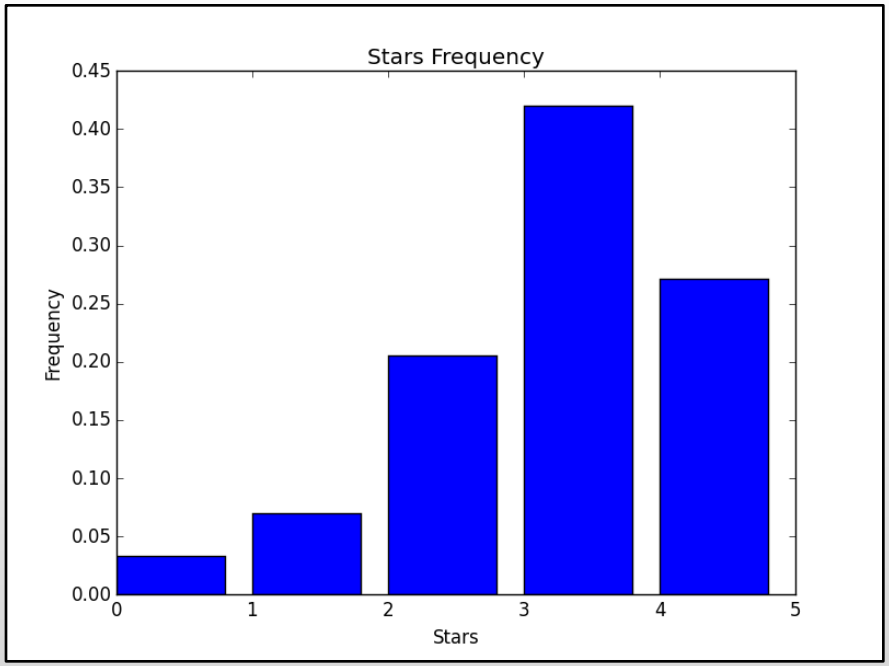
- 1.6M reviews
- 500 Tips
- 366K Users
- 61K Business

about

Several cities across the world (Edinburgh, Karlsruhe, Montreal, Waterloo, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison)

Edinburgh:

- 23780 reviews
- 3150 users
- 4576 businesses



However...
In this way we can only access to reviews of businesses in Edinburgh

[1]:http://www.yelp.com/dataset_challenge

Outline

- Introduction
- Data Analyses
- Approach
  - Data Preprocessing
  - Algorithms
- Experiment Results
- Conclusion

Approach(1/2)

Data Preprocessing – Hidden Features

I had a pretty good experience at the Doric. Yes, the restaurant is upstairs and you have to do a little exploring to find the staircase. The food was decently priced and the beer was pretty good. Service was fast and helpful.

POS(Penn Treebank)
What is important?

I/PRP had/VBD a/DT **pretty/RB good/JJ experience/NN** at/IN the/DT Doric/NNP.
Yes/UH, the/DT **restaurant/NN** is/VBZ **upstairs/JJ** and/CC you/PRP have/VBP to/TO do/VB a/DT **little/JJ exploring/NN** to/TO find/VB the/DT **staircase/NN**.
The/DT **food/NN** was/VBD **decently/RB** priced/VBN and/CC the/DT **beer/NN** was/VBD **pretty/RB good/JJ**.
Service/NNP was/VBD **fast/RB** and/CC **helpful/JJ**.

Approach(2/2)

Data Preprocessing – Hidden Features

Some parameters definition:

- r_m set of reviews of user *m*
- X_m set of relevant words of the user *m*
- n_{xmz} the numer of times the word *x* appear in the review *z*
- s_{mz} the rank that user *m* gives to review *z*

For each extracted words we have defined the following features:

Frequency:

$$f(x,m) = \frac{\sum_{i \in r_m} n_{xmi}}{\sum_{i \in r_m} \sum_{k \in X_m} n_{kmi}}$$

Regularity:

$$r(x,m) = \frac{\sum_{i \in r_m} n_{xmi} \frac{1}{\sum_{k \in X_m} n_{kmi}}}{|r_m|}$$

Relevance:

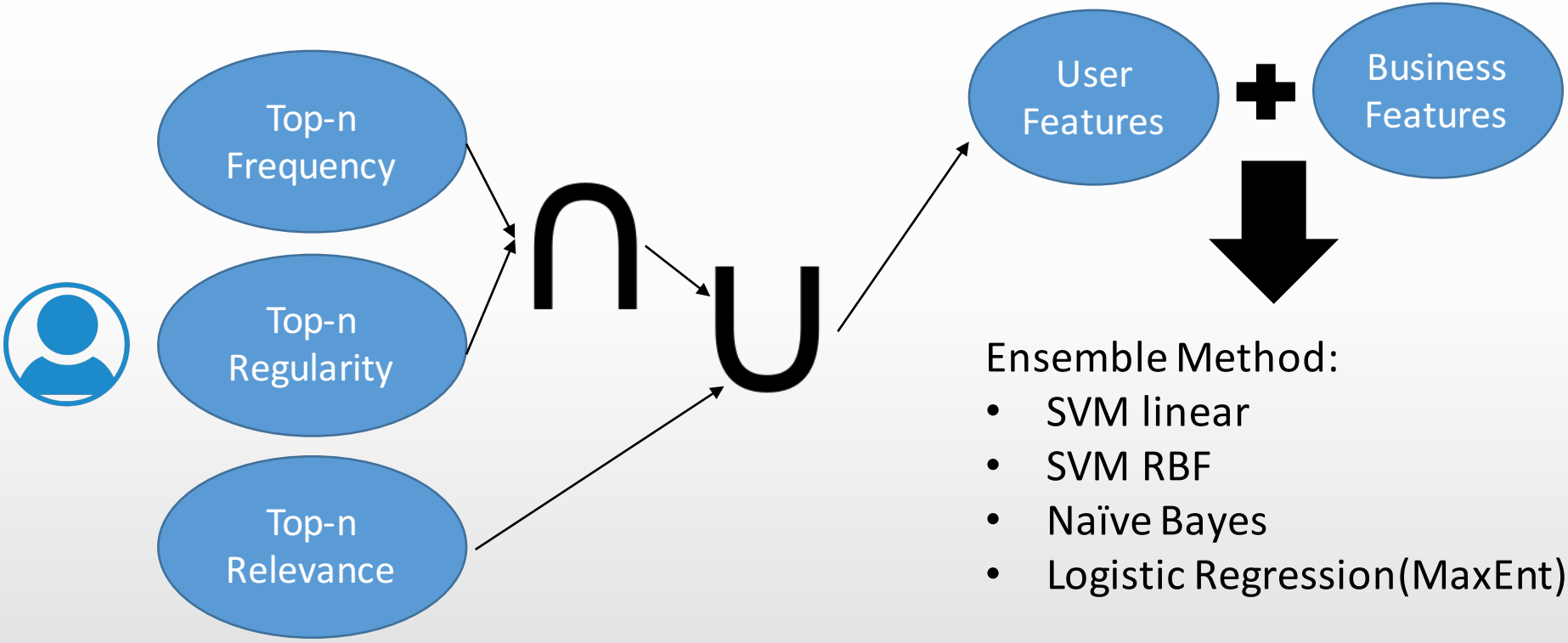
$$i(x,m) = \frac{\sum_{i \in r_m} n_{xmi} \frac{s_{mi}}{\sum_{k \in X_m} n_{kmi}}}{|r_m|}$$

Outline

- Introduction
- Data Analyses
- Approach
  - Data Preprocessing
  - Algorithms
- Experiment Results
- Conclusion

Approach(1/2)

Algorithms - Method 1°



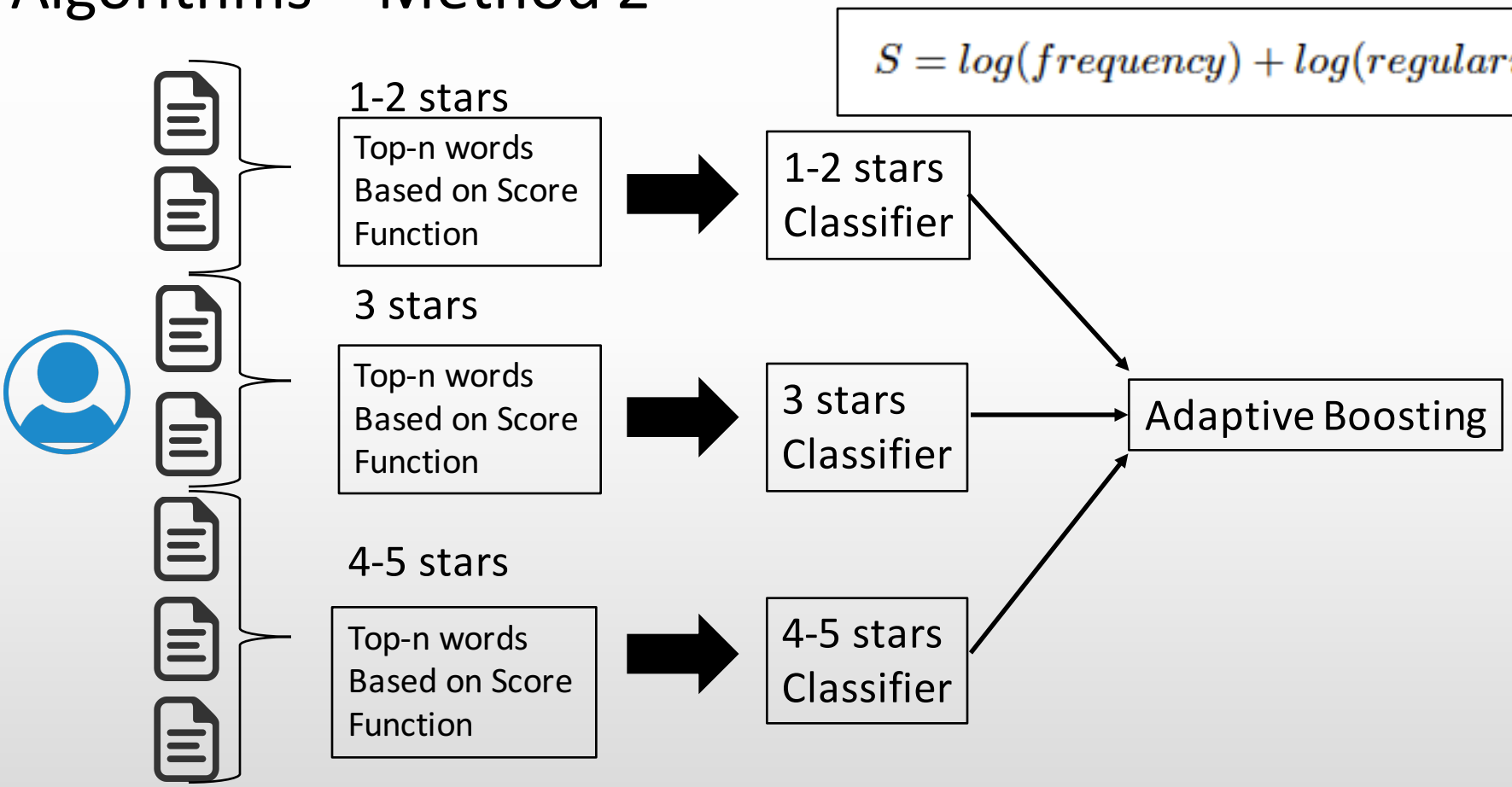
Ensemble Method:

- SVM linear
- SVM RBF
- Naïve Bayes
- Logistic Regression(MaxEnt)

Approach(2/2)

Algorithms – Method 2°

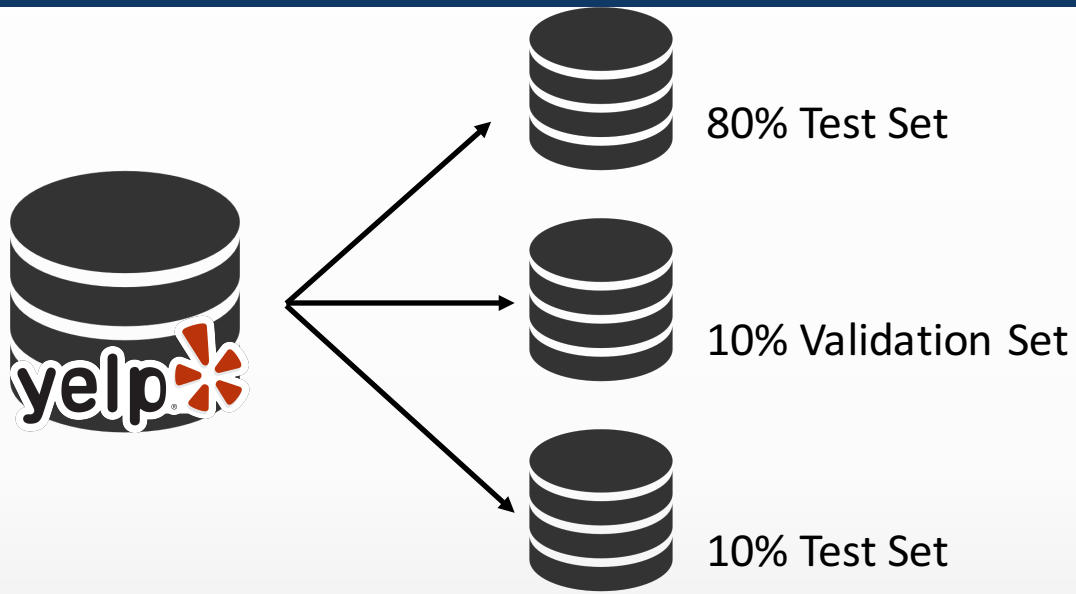
$$S = log(frequency) + log(regularity)$$



Outline

- Introduction
- Data Analyses
- Approach
- Experiment Results
- Conclusion

Experiment Results(1/3)



Baseline:

- Majority Class – We have 9 classes, accuracy ≈11%
- Majority Class – Actually only 5 classes appear, accuracy ≈20%
- Majority Class per user – For each user we predict its majority class ≈ 44.9%

Experiment Results(2/3)

A big issue in Recommendation System is: *cold start problem*
Since, we do not want to deal with it, we decide to consider user with at least a minimum number of reviews.

Min Reviews	20	25	50	75
Baseline	0.441	0.442	0.449	0.449
Method1	0.445	0.446	0.455	0.445
Method2	//	//	0.539*	//

Second Method Accuracy per class:

Class	[1-2]	[3]	[4-5]
Baseline	0.917	0.768	0.463
Method2	0.917	0.795*	0.587*

*p-value ≈ 0

Experiment Results(3/3)

Error Analysis

True \Predicted	1	2	3	4	5
1	0	5	8	5	4
2	1	17	20	43	11
3	0	17	144	113	27
4	0	22	67	422	66
5	0	14	18	153	140

Method	Average Error
Baseline	-0.15
Method1	-0.10
Method2	-0.08

Outline

- Introduction
- Data Analyses
- Approach
- Experiment Results
- Conclusion

Conclusion

- It is possible to extract user related information only from text
- Flexible Approach
- Computational Attractive (100 ≈ 4 minutes)

Future Works:

- Add sentiment analysis
- Extend the considered words with synonyms and/or hyperym
- Integrate the proposed method with existing recommendation techniques
- Try the method on other social media and on stratified datasets

Questions?

Thank you!

Detailed Analysis - Input

$[u_1, u_2, u_3, \dots, u_k]$

- Restaurant{0.012, 0.033, 0.11}
- Beer {0.023, 0.001,0.12}
- ...
- Bar{0.017, 0.023,0.09}

$[b_1, b_2, b_3, \dots, b_h]$

- Restaurant{0.003, 0.031, 0.01}
- Beer {0.012, 0.01,0.09}
- ...
- Service{0.043, 0.053,0.09}

Details!

- Restaurant{0.003, 0.031, 0.01}
- Beer {0.012, 0.01,0.09}
- ...
- Bar{0, 0,0}

$[u_1, u_2, u_3, \dots, u_k]$ $[b_1, b_2, b_3, \dots, b_k]$

Input

MODEL

Top n:
Model1 -> n = 30
Model2 -> n = 100

Model1:
$30 \le k \le 60$ -> $90 \le \#features \le 180$
Model2:
$0 \le k \le 100$ -> $0 \le \#features \le 200$