GOURMETNET

Metehan YILDIRIM & Mete Han KAHRAMAN & Ilayda CAVUSOGLU
Department of Computer Engineering
Hacettepe University
Ankara, Turkey

ABSTRACT

In our report we explain the process of building a recommendation system for Yelp. Then focus the work we have done so far which is categorizing restaurants. We did this so later we can build a recommendation system on top of it. We first tried to use the restaurants of tags and using Word2Vec[1]. Then getting word vectors we tried to apply K-Means Cluster[2] method to get some categories. This approach has failed and later we tried using manual categories.

1 A FOOD RECOMMENDATION SYSTEM : GOURMETNET

Our goal is to develop a food recommendation system for Yelp. The program will learn a person's taste according to the person's previous ranks that are given to restaurants and recommend a restaurant.

The research field for recommendation systems are not very active. Despite the interest of the companies on this topic. We found this topic to be very practical for real life purposes.

2 RELATED WORK

This is not a hot area on Machine Learning. But there is a lot of research caused by the industry on recommendation systems.[3] This topic comes up usually as Collaborative Filtering. We specifically looked at recommendation systems for Yelp.[4,5,6]

All approaches first worked on categorizing the foods because the data was sparse. Then using K-Nearest Neighbour and Collaborative Filtering on this categories. And most approaches also used a Graph we have yet to use it and we will see if it will be needed.

3 METHODOLOGY

There are many approaches to this but we preferred the baseline method the collaborative filtering. Collaborative filtering can be applied in two ways, a narrow one and a more general one.

Narrow one is a method that makes automatic predictions. It collects preferences or taste information from many users. Then predicts about a user's interests. In our case narrow one will be used.

First problem is the sparseness of our data. Applying K-Nearest Neighbours[7] on this data is very impractical. So first we need to group our data. To do that Google's pretrained Word2Vec[8] corpus was going to be used. We thought using about them on the restaurant tags. After grouping our data with Word2Vec vectors K-Means Clustering was going to be applied. We later saw that this approach didn't work. For more information please seek section 4.

K-means clustering partitions N observations into k clusters. Each observation belongs to the cluster with the nearest mean. A observation is a prototype of it's own cluster. Objective function :

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\|x_i^{(j)} - c_j\|}_{distance}$$

c_j : is the centre of the cluster
$x_i^{(j)}$: is the data point

4 EXPERIMENTAL EVALUATION

We are going to use Yelp dataset. Yelp data set includes many attributes. We analyse them and choose the most appropriate ones.

The required business attributes are stars , review_count, name, city, categories and business_id. We shall recommend a restaurant which belongs to the city that the user is currently located so city attribute is needed. Restaurants that do not have "categories" attribute is going to be deleted.

The required user attributes are average stars and user_id.

For both users and cities average rate will be used.

We stripped the data only using these features. On some restaurants some fields were empty so we disregarded those. On closer observation we saw that majority of the shops weren't even restaurants. So we disregarded every shop without the tag named "Restaurant". And then we also disregarded restaurants with reviews less than 10. So after this we had around 25000 rows.

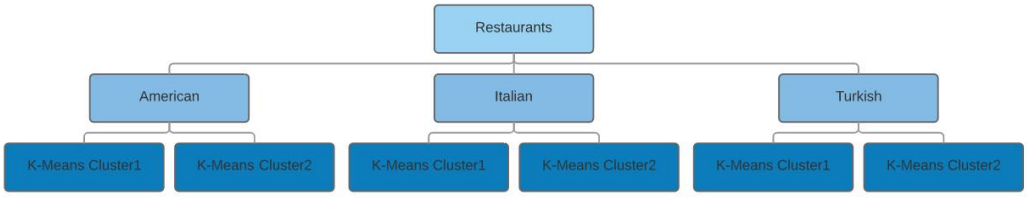
4.1 WORD2VEC ON K-MEANS APPROACH

So to categorize that we first thought of using Word2Vec for semantic similarities on tags and then using K-Means Clustering to group similar businesses together. This approach failed. Below you can see some examples of the clustering. We picked K as 100 and this is the result. As you can see "chinese" and "american" has the same label. This is not desired for us. So we needed a different approach.

Category	Label
Shopping	1
Bar	5
Nightclub	5
halal	11
japanese	24
american	34
chinese	34
italian	36
french	41
fries	72
sushi	72
club	76
turkish	76
coke	76
hamburger	78
hotdog	78
cheeseburger	78
steak	78
pencil	94
potato	98

4.2 MANUAL CATEGORIZATION

So we analysed the data even further and observed that most restaurant categories had their cuisine names. So we clustered our data on cuisines. We are thinking about clustering each cuisine clusters based on their other tags.



5 REFERENCES

[1]https://papers.nips.cc/paper/5021-distributed-representations-of-words\
-and-phrases-and-their-compositionality.pdf

[2]https://msu.edu/~ashton/classes/866/papers/2010_jain_kmeans_
50yrs__clustering_review.pdf

[3]https://www.cs.elte.hu/blobs/diplomamunkak/msc_alkmat/2016/
kelen_domokos_miklos.pdf

[4]http://snap.stanford.edu/class/cs224w-2012/projects/
cs224w-054-final.pdf

[5]https://pdfs.semanticscholar.org/8b2b/ada221819161961l6f1711d456ea212f2b3b.
pdf

[6]https://pdfs.semanticscholar.org/a7d2/5c03ec2a7dfe54c7b2e39729e906283e8e07.
pdf

[7]https://msu.edu/~ashton/classes/866/papers/2010_jain_kmeans_
50yrs__clustering_review.pdf

[8]http://www.gelbukh.com/ijcla/2014-1/IJCLA-2014-1-Complete.pdf#
page=27