UNIVERSITY OF ILLINOIS AT CHICAGO
CS521 - STATISTICAL NATURAL LANGUAGE PROCESSING

Project Proposal

MS.Umberto Di Fabrizio, MS.Vittorio Selo
October 17, 2015

I. PROBLEM

Nowadays with the growth of websites that offer user-generated content (e.g. Yelp, IMDb, TripAdvisor) there is an increasing need, for companies, to extract the tastes of the users in order to make unique the user experience.
Generally to achieve this outcome the system is trained using high-level features, for instance in a movie platform those features would be: the rating of the film, the genre, the actors and so on.
Our idea is to extract *hidden-features* of an user in order to understand their personal tastes. For example suppose a user writes reviews of mexican restaurants, why is one restaurant better than another although the food is good in both? Usually, people pay attention to details such as lights, atmosphere, the type of customers and so on, and sometimes without even being aware of their pet peeves.
The issue is how to find out, in an automatic way, those tastes that rule people feelings about a place (or an item) and which source to use.

II. SOLUTION

We plan to develop a framework to detect and extract people personal and intimate tastes.
To extract the *hidden-features* we will mine the reviews of an user in order to collect the most common topics (furniture, lights, etc.) and understand what he really observes to judge a business (e.g. restaurant, pub, hotel). The hypothesis is that if an user always talks about certain topics then he cares a lot about that topic and this can guide the recommendation system. Suppose:

- If the user says that he did NOT like pizza as many times as he DID like it, then pizza is something very important for the user, so the recommendation system will suggest places were pizza as positive reviews.
- If the user always says that it did not like the sofas maybe the user really does not like to seat on sofa when eating (maybe he prefers chairs). So the recommendation system will not suggest any place that as a review talking (positively or negatively) about the sofas.

In order to suggest which place to recommend, our system will compute the *k-top* topics for a business based on all the reviews for that place (this will be done offline). Once we know both the topics of an user and the topics of a business we will use a similarity function to understand if those topics have the same polarity. We understand that there is plenty of work to be done, so we divided our pipeline as follows:

1) Take reviews and mine topics. This will require some time because we need to understand how many reviews we need about an user to significantly detect his core topics. Then we will analyze those result to understand if the topic detection is enough to extract the user preferences or if we need to add any extra feature (words that appear few times may be very important and yet not detected by the topic modeling algorithm). By the end of this phase we will already have significant results about the feasibility of the recommendation system and it is considered by itself a possible project ending.
2) Assign polarity to each of the topic detected for an user accordingly to how the user talked about that topic. This will require to use some tool for sentiment analysis of sentences.
3) Create a similarity function between clusters of topics to understand if two topics are similar. There is some literature about this, although it is very recent[2].
4) Create the *k-top* topics for each business, com-

pare it with the user *m-top* topics and the system will return the *top-n* businesses that score higher with respect to the similarity function.

III. DATA

Yelp dataset challenge[1].

IV. ADDITIONAL INFORMATION

We are both two Master Degree students without a thesis. Umberto Di Fabrizio is taking the Neural Networks class with Prof.Graupe but has not a project yet.

REFERENCES

[1] www.yelp.com/dataset_challenge
[2] Measuring the Similarity between Automatically Generated Topics,2014,Aletras at al.