Natural Language Processing and the Web (Presentation)

------- Tianyang Zhou (2589435), Yantao Shi (2673707), Lin Li (2378303)

Topic

Basic Idea:
We will create a system based on the TripAdvisor dataset to find similar hotel(s) according to the description article.

Data Preprocessing:
Firstly, we will only keep the comments from users in all the files (one for each hotel as before), and delete all the other information except location and overall rating. Each comment will be seen as a separate paragraph.

Algorithmus:
We will use the LibSvm from the DkPro Project.

Input is the all those preproccessed documents of hotels, we will use the words from the given feature words file together with their td-idf weights to form vectors for each document and then pass it to SVM.
As label for output is the overall rating. For the first phase we will train the machine to get a better prediction of this overall rating according to the comments. We will try to minimize the MSE of our output in comparison with the true value.

After we finished training our machine, then we will get a descriptive article, and take it as input, then we will get a rating as output. Based on this rating, we will find an array of possible hotels, which have a similar overall rating. And then we will filter the hotels due to the location and some other parameters.

Evaluation:

We will compare the output with the real overall rating to see if the prediction is good.

For the train and test dataset we would like to use the Cross-Validation over the original dataset.

Possible Improvements:

1. according to the train results, we will delete some feature words from the original wordset (for example, if they never appear). Furthermore, we would also add some other feature words on depends (for the begining manuelly, maybe also use third-party resource during the developement).
2. filter some comments out to improve the prediction. For example if the comments are too short.

Problems:

1. How similar is the descriptive article to the user comments? For now we just suppose that they are the same, but actually the grammatic and structual differences between those texts could have an influence over the result. And that is also an issue what we should work out.