Natural Language Processing and the Web (Final-Presentation)

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

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

We want to look for potential competitors based on customer reviews of hotels/restaurants, that is, to find peers who are similar to each other. At the same time, this method can also be used for those tourists who do not want to stay in the same hotel/restaurant in the same city, we recommend similar hotels/restaurants for them to choose.

Guideline

TECHNISCHE UNIVERSITÄT DARMSTADT

The whole process will be executed in two steps:
First step: we calculate the tf-idf values of the predefined feature words for every hotel file (in this case, a document means every single paragraph in the file) with help of UIMA annotation. We transfer this as a vector for the file to libsvm and get predicted label for the test hotel file. Then we will select the hotels, which locate in the same city with the test hotel file, and then pick the hotels with the most similar labels (overall rating) with the predicted label of the test hotel file as candidates.
Second step: we let the user type in the feature words they want to use, and use these feature words together to form a new feature and calculate the tfidf value again among the candidate hotels. In the end we pick out one hotel with the highest tfidf value of this new feature.

Elements

TECHNISCHE UNIVERSITÄT DARMSTADT

Original data:

Data provided by TripAdvisor, Most of the files contain the overall rating, hotel name and location. Then the comments from users with <Content> tag.

<Overall Rating>4 rating
<Avg. Price>$302
<URL>http://www.tripadvisor.com/ShowUserReviews-g60878-d100504-r22932337-
Hotel_Monaco_Seattle_a_Kimpton_Hotel-Seattle_Washington.html

hotel name location

<Author>selizabethm
<Content>Wonderful time- even with the snow! What a great experience! From the goldfish in the room (which my daughter loved) to the fact that the valet parking staff who put on my chains on for me it was fabulous. The staff was attentive and went above and beyond to make our stay enjoyable. Oh, and about the parking: the charge is about what you would pay at any garage or lot- and I bet they wouldn't help you out in the snow!
<Date>Dec 23, 2008
<No. Reader>-1
<No. Helpful>-1
<Overall>5

comment

picture 1: original hotel file

hotelName Hotel_Monaco_Seattle_a_Kimpton_Hotel
location: Seattle_Washington
overallRating 4

<Content>Wonderful time- even with the snow! What a great experience! From the goldfish in the room (which my daughter loved) to the fact that the valet parking staff who put on my chains on for me it was fabulous. The staff was attentive and went above and beyond to make our stay enjoyable. Oh, and about the parking: the charge is about what you would pay at any garage or lot- and I bet they wouldn't help you out in the snow!

picture 2: modified hotel file

Train data:

We will only take the part in which is tagged in the left picture and it looks like in the right picture.

Elements

TECHNISCHE UNIVERSITÄT DARMSTADT

Test data:
The test hotel files look exactly the same with train data.
(* because some original files don't have a url and thus have no location and hotel name, we just discard them**)**

Feature words:
The original feature words provided by dataset contains more than thousand feature words. At the very begining we just picked out 400 words due to their indexes (picture 3).

0	staff
1	location
2	night
3	clean
4	nice
5	time
6	breakfast
7	walk
8	day
9	service

picture 3: original feature words

3	clean
4	nice
10	restaurant
11	bed
12	friendly
13	bathroom
14	helpful
15	recommend
16	food
17	view

picture 4: delete some feature words

```
terrible horror horrible scare scared bloody
bad worse worst ugly waste disgrace nightmare aweful
lie trick fake
dirt dust unclean
love enjoy cheer
interest fantastic wonderful
```

picture 5: collapse similar words

Elements

TECHNISCHE UNIVERSITÄT DARMSTADT

Feature words:

During the svm processing we noticed that many words are not very relevant to the label, which is the rating. So we picked about 300 out of the words, like in picture 4.

We also noticed that calculate the similar words as separate feature made them all not that relevant for label, better is, that we collapse them into one feature. After deleting many other irrelevant words we got the final feature words, like in picture 5. Every line is a set of words, which form together a feature.

0	staff
1	location
2	night
3	clean
4	nice
5	time
6	breakfast
7	walk
8	day
9	service

picture 3: original feature words

3	clean
4	nice
10	restaurant
11	bed
12	friendly
13	bathroom
14	helpful
15	recommend
16	food
17	view

picture 4: delete some feature words

```
terrible horror horrible scare scared bloody
bad worse worst ugly waste disgrace nightmare aweful
lie trick fake
dirt dust unclean
love enjoy cheer
interest fantastic wonderful
```

picture 5: collapse similar words

Additional feature:

Elements

TECHNISCHE UNIVERSITÄT DARMSTADT

Additional feature:

Intuitively we know that the rating is most relevant to the ratio between positive and negative comments. So why don't we just calculate **how "positive" the comments overall are**?

We calculate for every paragraph the ratio between positive features and all features r_p,

and then calculate the average ratio over all paragraphs: $r_{all} = \sum_{n_{para}} r_p$

The vector file as input for the svm looks like this:

index value

label

```
4.5  0  4.238383038018905  1  0.0  2  6.849576910286561  3
0.0  4  13.45441008856036  5  13.16077902783271  6
8.332150383620975
3.5  0  7.3786905709933714  1  12.96918750639406  2
9.511397556565733  3  3.586789755406759  4  12.81977003273233
5  10.022498110584017  6  4.624490986492345
```

picture 7: vector file

Other improvements:

1. detect the negative expression
2. normalize the feature due to the size of comments

Processing

TECHNISCHE UNIVERSITÄT DARMSTADT

Create train and test files:
we create train and test files out of modified hotel files. In our case, we put for every location the second hotel file into test, and use the other hotel files as train files.

Algorithmus:
libsvm: we use the libsvm to process the vector file. We choose the nu-svr mode from libsvm to use the whole 7 features to regression over the label.
After we got the predicted label for the test file, we use a list to store the train files, where the train hotel with most similar label with the test hotel come first.
Often, there are **more than one candidates**, thus we need the **second step:**

```
optimization finished, #iter = 176
epsilon = 3.207053034053331E-4
obj = -8.32379239282361, rho = -3.477110590353275
nSV = 41, nBSV = 7
Mean squared error = 0.47786456920716464 (regression)
Squared correlation coefficient = 1.0000000000312694 (regression)
Please enter a location (lowercase)
```

picture 8: example result for the first step (svm)

Processing

TECHNISCHE UNIVERSITÄT DARMSTADT

Second step:

because we have multiple test files (each for one location), we let user choose one location:

```
optimization finished, #iter = 176
epsilon = 3.207053034053331E-4
obj = -8.32379239282361, rho = -3.477110590353275
nSV = 41, nBSV = 7
Mean squared error = 0.47786456920716464 (regression)
Squared correlation coefficient = 1.0000000000312694 (regression)
Please enter a location (lowercase)
```

picture 8: example result for the first step (svm)

if the location is incorrect, user must type once again until location is correct:

```
Please enter a location (lowercase)
darmstadt
The location you entered is not included, please enter a new location (lowercase)
texas
3
```

picture 9: enter location again if location isn't included

Processing

TECHNISCHE UNIVERSITÄT DARMSTADT

we get the candidates and let user define their feature words, and calculate tfidf value for this feature among the candidates and pick up the one with highest value:

```
Please enter your feature word(s) separated with space
sea sunshine warm
??
4
Dallas_Texas_3_InterContinental_Dallas.dat  processing
Dallas_Texas_4_The_Magnolia_Hotel_Dallas.dat  processing
Dallas_Texas_7_Hotel_Lawrence.dat  processing
Dallas_Texas_8_La_Quinta_Inn_and_Suites_Dallas_Addison_Galleria.dat  processing
We choose file: Dallas_Texas_7_Hotel_Lawrence.dat, because it has the highest score for the given feature words
```

picture 10: enter feature words and get the hotel returned

Further work

TECHNISCHE UNIVERSITÄT DARMSTADT

1. the data could be cleaner after preprocessing: for example some comments in non-english languages could be detected and discarded:

```
<Author>lass=
<Content>ein super Hotel   showReview(15568780, 'full');
<Date>Apr 24, 2008
<No. Reader>-1
<No. Helpful>-1
<Overall>5
<Value>-1
<Rooms>5
<Location>-1
<Cleanliness>5
<Check in / front desk>-1
<Service>5
<Business service>-1

<Author>lass=
<Content>鍇曞嗞鍘滈姹顆＄┐閨?#x201d;   showReview(19164371, 'full');
```

2. we could use some more information besides the comments, for example the rating of room, value and even how many people find the comment is useful.
3. we could use test data from other datasets

Resources

TripAdvisor dataset source: *http://times.cs.uiuc.edu/~wang296/Data/*