APPLICATION OF DATA-MINING TO BUILD RECOMMENDATION SYSTEM FOR HOTEL REVIEW

PROJECT BY:
EKAMBER CHADDA 11-CSS-17
HATIM TAI 11-CSS-21

AIM OF THE PROJECT

To build a system capable of processing the customers' reviews for different 5 star hotels of Delhi and mining some useful information from it .

In particular the project focused on extracting information on six predefined features : "breakfast " , "staff" , "service" , "location" , "swimming pool" and "rooms"

TOOLS USED

PYTHON
Programming language used for building the system

- NLTK (NATURAL LANGUAGE TOOL KIT)
- A leading platform for building Python programs to work with human language data.

STANDFARD PARSER
Parser to analyze grammatical structure of sentence

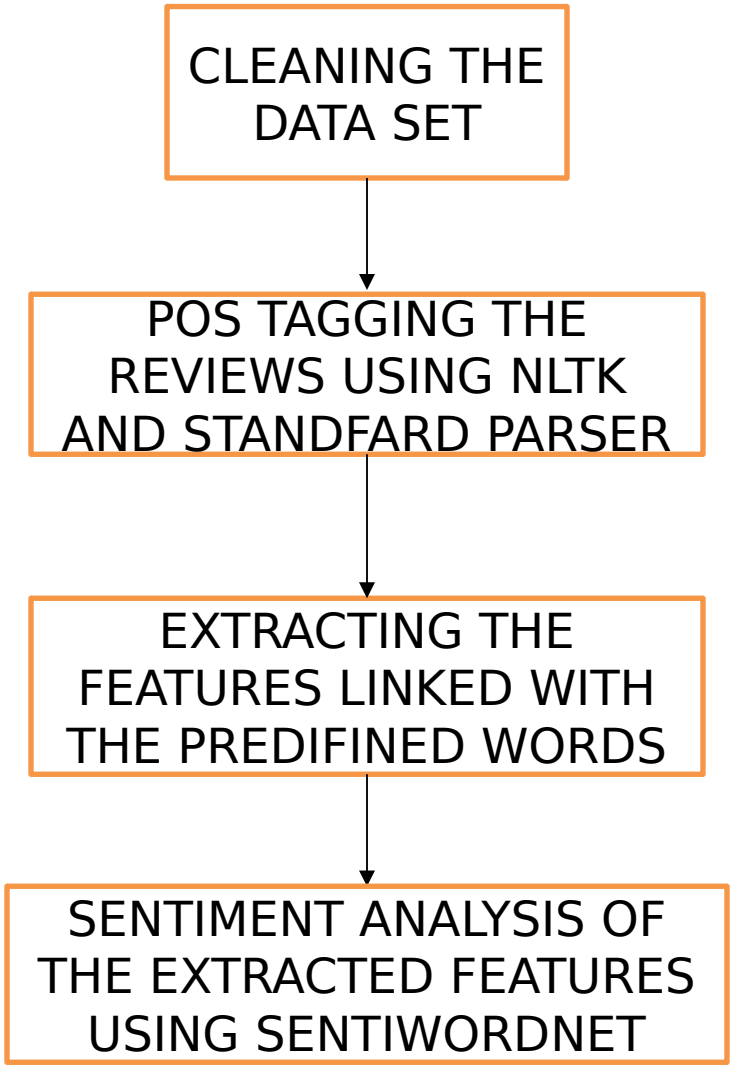
-
- SENTIWORDNET
- A lexical resource for opinion marking

DATA SET

Data set comprised of reviews collected from various travelling websites like Tripadvisor.com and booking.com of following hotels :

HYATT REGENCY	30 REVIEWS
THE LALIT	31 REVIEWS
TAJ PALACE	30 REVIEWS
THE OBEROI	30 REVIEWS
ITC MAURYA	27 REVIEWS

FLOW CHART OF THE PROJECT



POS TAGGING THE REVIEWS

“impeccable service yes it is pricey but the service makes it worth it the taj hotels are known for their warm indian hospitality certainly lived up to their reputation everywhere you went greeted with smiles i was not very well during my stay the reception staff could not do enough”

(after pos tagging)

[(u'impeccable', u'JJ'), (u'service', u'NN'), (u'yes', u'RB'), (u'it', u'PRP'), (u'is', u'VBZ'), (u'pricey', u'JJ'), (u'but', u'CC'), (u'the', u'DT'), (u'service', u'NN'), (u'makes', u'VBZ'), (u'it', u'PRP'), (u'worth', u'IN'), (u'it', u'PRP'), (u'the', u'DT'), (u'taj', u'NN'), (u'hotels', u'NNS'), (u'are', u'VBP'), (u'known', u'VBN'), (u'for', u'IN'), (u'their', u'PRP$'), (u'warm', u'JJ'), (u'indian', u'JJ'), (u'hospitality', u'NN'), (u'certainly', u'RB'), (u'lived', u'VBD'), (u'up', u'RP'), (u'to', u'TO'), (u'their', u'PRP$'), (u'reputation', u'NN'), (u'everywhere', u'RB'), (u'you', u'PRP'), (u'went', u'VBD'), (u'greeted', u'VBN'), (u'with', u'IN'), (u'smiles', u'NNS'), (u'i', u'FW'), (u'was', u'VBD'), (u'not', u'RB'), (u'very', u'RB'), (u'well', u'RB'), (u'during', u'IN'), (u'my', u'PRP$'), (u'stay', u'VB'), (u'the', u'DT'), (u'reception', u'NN'), (u'staff', u'NN'), (u'could', u'MD'), (u'not', u'RB')

EXTRACTING FEATURES

In order to extract the feathers linked with a predefined word (“breakfast” etc) we searched for the adjectives nearest to that word

“the buffet breakfast was incredible and delicious”

(u'the', u'DT'), (u'buffet', u'NN'), (u'breakfast', u'NN'), (u'was', u'VBD'), (u'incredible', u'JJ'), (u'and', u'CC'), (u'delicious', u'JJ')

SENTIMENT ANALYSIS OF REVIEWS

A sentiwordnet data file looks like:

```
24 # POS   offset  PosScore        NegScore        SynsetTerms
25 a       1000003 0.0     0.125   form-only#a#1
26 a       1000159 0.25    0.0     dress#a#1 full-dress#a#1
27 a       1000307 0.0     0.0     titular#a#5 nominal#a#6
28 a       1000440 0.0     0.0     prescribed#a#4 positive#a#5
29 a       1000554 0.0     0.25    perfunctory#a#2 pro_forma#a#1
30 a       1000681 0.0     0.5     semiformal#a#1 black-tie#a#1 semi-formal#a#1
31 a       10007   0.0     0.625   abstentious#a#1 abstinent#a#1
32 a       1000859 0.0     0.0     starchy#a#2 buckram#a#1 stiff#a#4
33 a       1001035 0.125   0.375   white-tie#a#1
34 a       1001157 0.0     0.0     informal#a#1
35 a       100126  0.5     0.0     viable#a#2
36 a       1001456 0.375   0.125   casual#a#3 everyday#a#2
37 a       1001581 0.0     0.0     free-and-easy#a#1 casual#a#8
38 a       1001755 0.0     0.375   folksy#a#2
39 a       1001882 0.0     0.625   unceremonious#a#1 unceremonial#a#1
40 a       1002013 0.0     0.25    formal#a#3
41 a       1002315 0.0     0.0     literary#a#3
42 a       1002508 0.0     0.0     informal#a#3
43 a       100261  0.0     0.125   vital#a#4
44 a       1002760 0.0     0.0     conversational#a#1 colloquial#a#1
45 a       1003005 0.0     0.0     vulgar#a#3 vernacular#a#1 common#a#5
46 a       1003296 0.0     0.0     epistolary#a#1 epistolatory#a#1
47 a       1003509 0.375   0.125   slangy#a#1
48 a       1003665 0.125   0.5     subliterary#a#1
49 a       1003815 0.25    0.375   unliterary#a#1 nonliterary#a#1
50 a       100393  0.0     0.75    dead#a#1
51 a       1003972 0.0     0.25    former#a#1
52 a       1004232 0.125   0.0     latter#a#1
53 a       1004423 0.125   0.25    last_mentioned#a#1
54 a       1004545 0.0     0.0     forsaken#a#1
55 a       1004767 0.0     0.0     deserted#a#1 abandoned#a#2
```

So the extracted features are searched in the sentiwordnet data file and a score is assigned to every feature according to the positive and negative scores of the respective word

For example
1073446 0.625 0.0 good