Sentiment Analysis and Rate Prediction of Hotel Reviews

Anupama Garani

Nasser Almohrij

AGENDA

- Introduction
  - Scope
  - Sentimental Analysis
  - Hotel Reviews
- Methods and Approach
  - Dataset Description
  - Classification Algorithm
- Results

INTRODUCTION

SCOPE

- Rating based on the classification of reviews
- Some of existing resources/websites
    - Oyster.com
    - Hotels.com
    - TripExpert
    - TripAdvisor
- Accommodation websites provide hotels' ratings which are not always reliable as some of these websites manipulates the ratings.
- Our method ensures fair rating based on the information gathered by classification of reviews and sentiment analysis.

SENTIMENT ANALYSIS

- It is a technique which refers to the use of text analysis and natural language processing to systematically identify and extract and quantify the subjective information.
- In this case we have two classes of sentiments
  - Positive
  - Negative
- Some instances from the data set
  - Great stay in an amazing City - Positive review
  - Poor value for money - Negative Review

HOTEL REVIEWS

- We have taken into consideration reviews of various hotels in each city as the dataset for sentiment analysis
- Reviews based on the experience by customers
- Data format is
    - <Date>< Review Title>< Review>

```
Nov 7 2009  A quality hotel with quality service    We stayed at the Ascott Beijing this summer for 1 week in a family of 5.
It is centrally located. When you are a european with kids staying for a few days it is ideal to have a fully equipped kitchen and a big Wal-Mart nearby.
We had a modern and spacious apartment with a big LCD flat television. The service was really good and important for us not speaking the language:
we never had to wait for a taxi and they always provided info for our directions.We would definitely return!
```

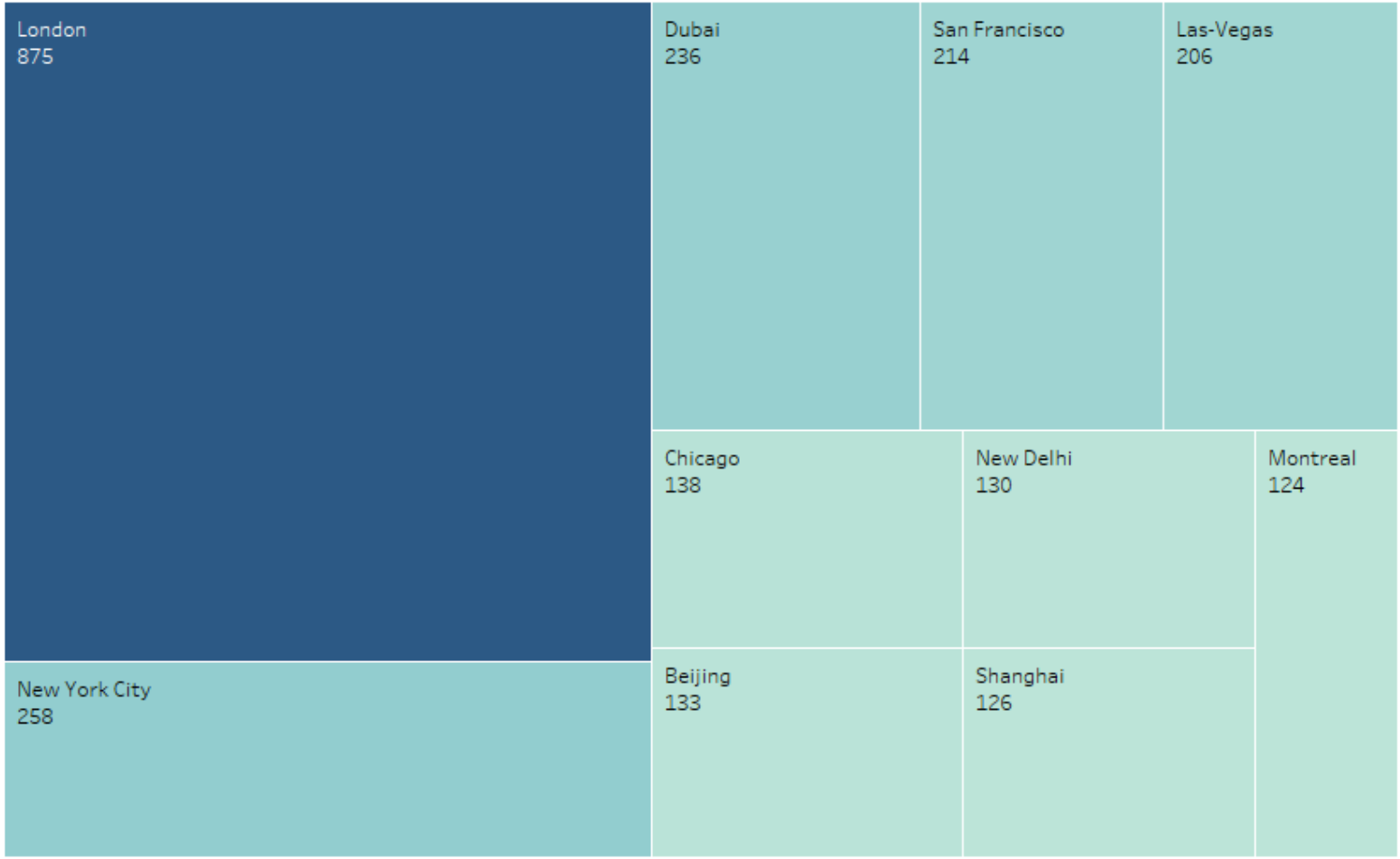
- On an average each review is about 2318 characters.

METHODS AND APPROACHES

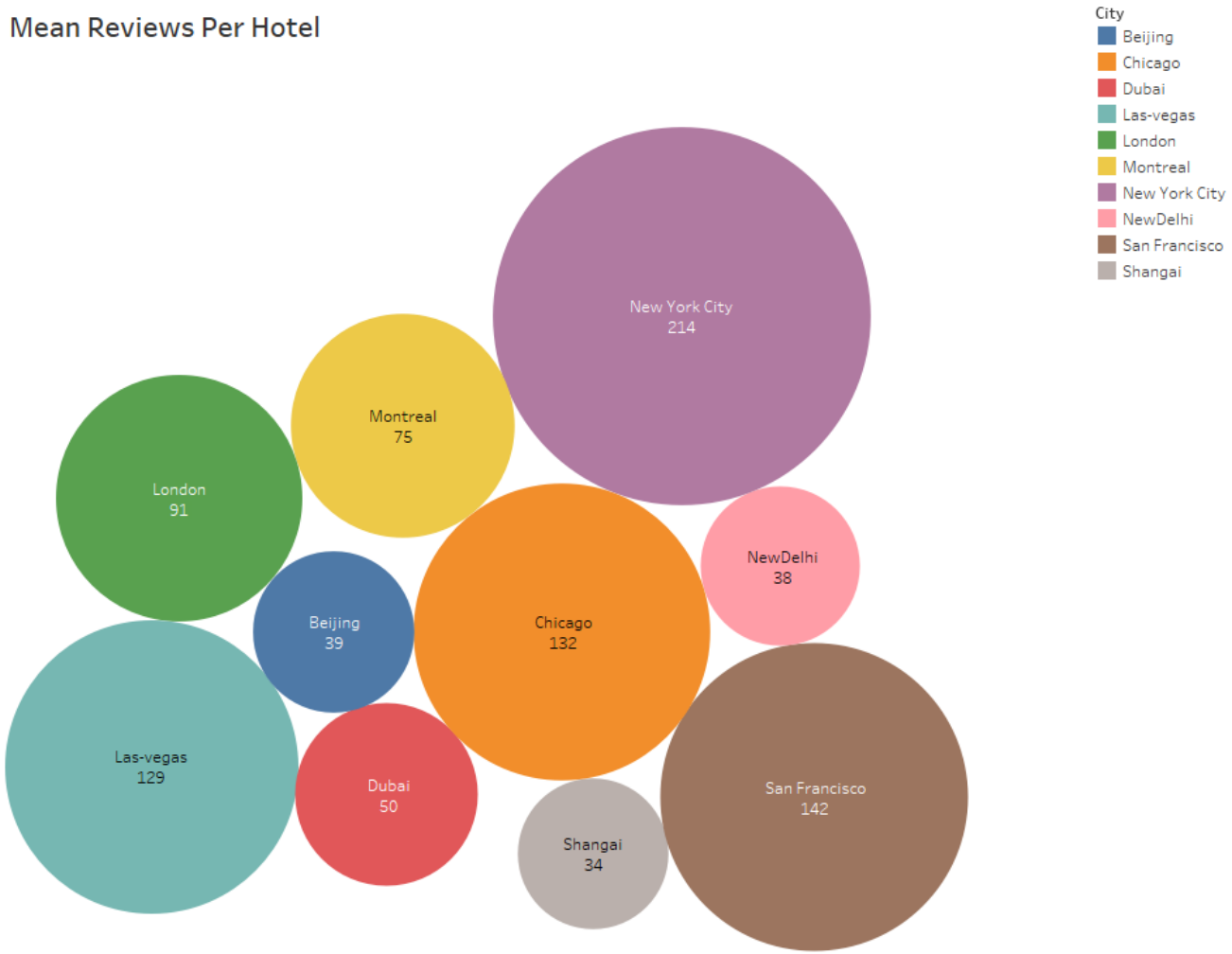
DATASET

- Statistics Collected Over 2002 - 2009
  - **Cities:** 10
    - Beijing, Chicago, Dubai, Las Vegas, London, Montreal, New Delhi, New York City, San Francisco, and Shanghai
  - **Hotels:** 2440
  - **Reviews:** 245,172

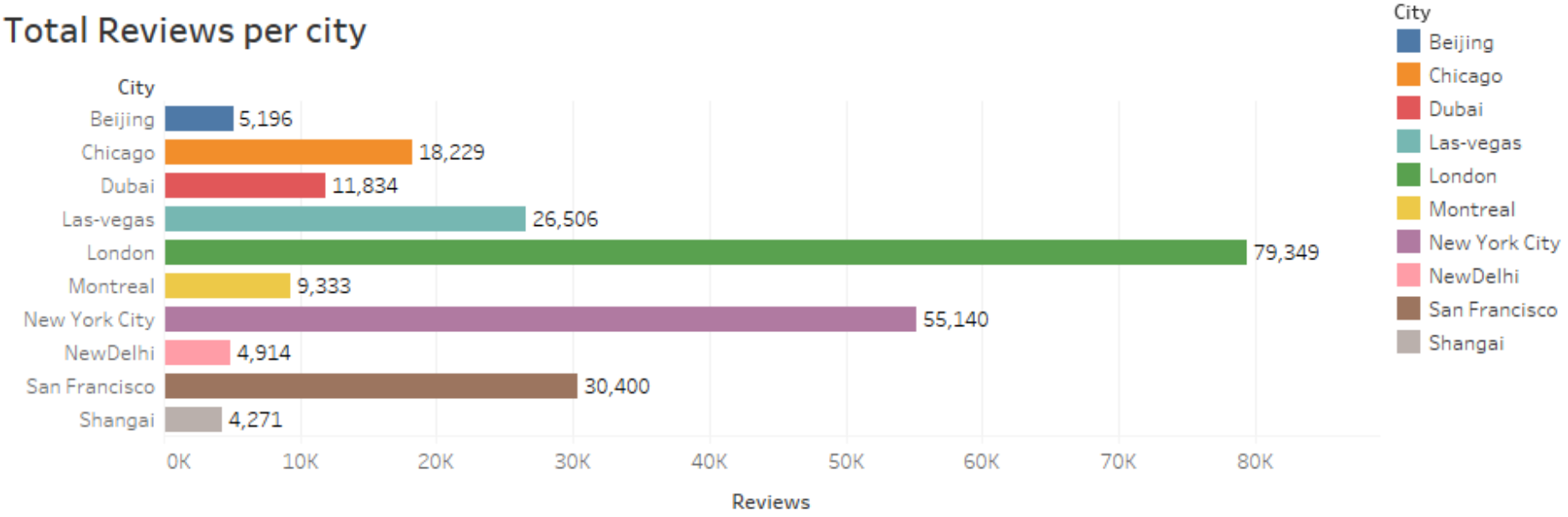
Hotels per city



Mean Reviews Per Hotel



Total Reviews per city



SAMPLES FROM DATASET

- Luvd it I was recommended this place by someone in Xl'an .I got there and hated the beds but loved everything else.
- Trop bruyant. Petit déjeuner à revoir. Bon emplacement.
- ??????10?
- Super Wahl für einen günstigen Peking-Trip!
- Great! (considered the price) Not expensive! Rooms are ok (bathroom can be a little more clean).

DATASET CLEANING

- The reviews of the dataset contains a lot of stray characters,date and non-English alphabets which slows down the process of classification.
  - ?,[,],/,\,|,*,&,;,:,!,$,%,(,),=,#,@,^,Oct,Jan,Mar,2009,2008..
  - Ü,é,à..
- Tools used for cleaning the stray characters: Regular expression in python
  - Before: Oct 7 2008 FABULOUS!!!!! Perfect Location.. close to everything.. The hotel is brand new:-D..
  - After: FABULOUS Perfect Location close to everything The hotel is brand new

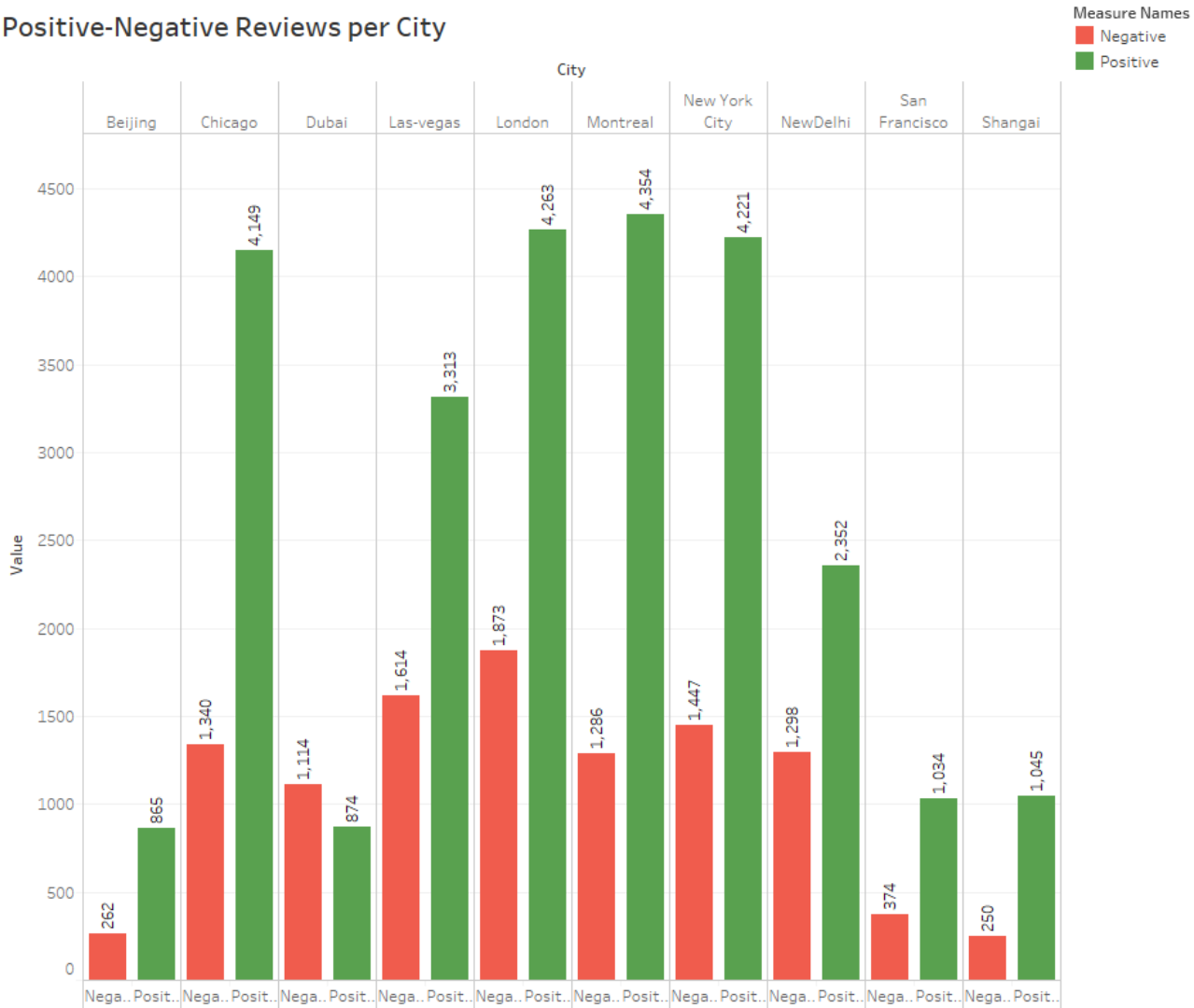
DATASET PREPROCESSING

- Data preprocessing does not include stemming since we would like to have high precision for the system.
- Casefolding normalization has been applied in the preprocessing of the data set.

CLASSIFICATION ALGORITHM

- Naive bayes classifier is used as the classification method to perform sentiment analysis on the hotels' reviews
- The algorithm is built using JAVA.
- Two classes
  - Positive
  - Negative

Positive-Negative Reviews per City



TRAINING SET

- 8000 reviews are used in the training set.
  - 4000 reviews each of positive and negative reviews.
- Python libraries used for sentiment analysis on the training data set
  - Scikit-learn
  - Numpy
  - NLTK

ldacluster - [C:\Users\Anupama\Documents\3SummerSemester\KPT\ldacluster] - ...\NBSVMClassifier.py - PyCharm Community Edition 2016.2.3

File Edit View Navigate Code Refactor Run Tools VCS Window Help

ldacluster > NBSVMClassifier.py

NBSVMClassifier

Project

- montrealOutput.txt
- msvcp71.dll
- msvcr71.dll
- naiveBayes.py
- NBSVMClassifier.py
- negativeReviews1.txt
- negativeReviews2.txt
- negativeReviews3.txt
- negativeReviewsChicago
- negativeReviewsdubai.tx
- negativeReviewsLasVega
- negativeReviewslondon.t
- negativeReviewsmontrea
- negativeReviewsnewBeiji
- negativeReviewsnewdelh
- negativeReviewsnewYork
- newDelhiOutput.txt

NBSVMClassifier.py | ClassPosNeg.py | pos.py | TextPreProcess.py | NumberReviews.py | recFile.py | FileClassification.py

```
26  tf_transformer = TfidfTransformer(use_idf=False).fit(X_train_counts)
27  X_train_tf = tf_transformer.transform(X_train_counts)
28  print(X_train_tf)
29  tfidf_transformer = TfidfTransformer()
30  X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
31  clf = MultinomialNB().fit(X_train_tfidf, dataset.target)
32
33  with open("ChicagoOutput.txt",'r') as f:
34      docs_new = f.read().split('\n')
35
36
37  #docs_new = ['worst hotel ever','very horrible hotel','do not recommend this to anyone','best
38  X_new_counts = count_vect.transform(docs_new)
39  X_new_tfidf = tfidf_transformer.transform(X_new_counts)
40  predicted = clf.predict(X_new_tfidf)
41  for doc, category in zip(docs_new, predicted):
42      with open("resclassif.txt",'w') as dest:
43          print('%r => %s' % (dataset.target_names[category],doc))
44          dest.flush()
```

NumberReviews.py | addinglinecount.py | lineOutput.txt

```
1  filename="C:/Users/Anupama/Documents/3SummerSemester/KPT
2  sum = 0;
3  with open(filename,'r') as f:
4      for line in f:
5          sum = sum + int(line)
6      print(sum)
```

Run NBSVMClassifier

! Too much output to process

```
'Positive' => EXcellent Location This hotel is perfect location  Only a block or so from Magnificent Mile  Michigan Ave  Train Subway is right in front of the hoteltook trains to Wri
'Positive' => Would definitely stay here again We Pricelined this hotel for a short getaway to meet friends in Chicago    Paid more than we wanted to but still a great deal Really gr
'negative' => Do not stay here under any circumstances Im writing this review to warn fellow travelers not to stay here  Because the wonderful Homewood Suites next door was booked up
'negative' => great location low on staff location was great resterant on sight was good indoor pool was clean warm enough to enjoy whirl pool great as well great location few steps
'Positive' => My first choice in Chicago This was our third time staying at the Hilton Garden Inn  We reserved the room through Priceline spending only  per night  TIP The quottwo Qu
'Positive' => Good Time had by Me I stayed for one week while attending a business conference  Within easy walking distance to Navy Pier Art Institute theater district etc  Grocery s
'negative' => Great location Pretty much had nothing much to complain about the hotel  Except when we checked in the staff who checked us in did not smile or greet us and continue ta
'Positive' => Everything You Could Want I recently spent a week at this location for business  As I traveled around Chicago during the week and visited friends at other hotels I conf
'negative' => The staff went above and beyond to make our experience wonderful My family and I were traveling to Frankfurt Germany and wanted a nonstop flight  Because we used to liv
'Positive' => Thank you Hilton magnificent mile We actually made a pitch to the Oprah show and met first in one of the convention rooms here  The wait staff did not know who we were
```

4: Run | 6: TODO | Python Console | Terminal | Event Log

TESTING SET

- 1000 reviews are used in the testing set.
  - 500 reviews are positive reviews
  - 500 reviews are negative reviews
- Testing set and training set are independent from each other.

USER INTERFACE

- JAVA SWING libraries is used to build the UI of the application.
- The UI provide search functionality.

RESULTS

CLASSIFIER EVALUATION

- Correctly Classified 851 out of 1000
- Accuracy: 85.1%
- Precision : 0.816
- Recall : 0.8774193548387097
- F-score: 0.8455958549222798