Opinion Based Entity Ranking (OpinRank) Dataset

With Relevance Judgments

Description of dataset used for the “Opinion Based Entity Ranking” project.

Kavita Ganesan
3/7/2012

OpinRank Review – Dataset (with Judgments)

Author: Kavita Ganesan (kganes2@illinois.edu)
HTML Version: http://www.kavita-ganesan.com/entity-ranking-data

DATASET OVERVIEW

This data set contains full reviews for cars and and hotels collected from **Tripadvisor** (~259,000 reviews) and **Edmunds** (~42,230 reviews).

CAR REVIEWS

DATASET DESCRIPTION

- Full reviews of cars for model-years **2007**, **2008**, and **2009** and corresponding aspect ratings used as relevance judgments
- There are about **140-250** cars for each model year
- Extracted fields include **dates**, **author names**, **favorites** and the **full textual review**
- Total number of reviews: **~42,230**
    - Year 2007 -18,903 reviews
    - Year 2008 -15,438 reviews
    - Year 2009 - 7,947 reviews

DATA FORMAT

In the data folder, there are three different subfolders (2007,2008,2009) representing the three model years. Each file (within these 3 folders) would contain all reviews for a particular car. You can ignore all the csv files in the data folder. The filename represents the name of the car. Within each car file, you would see a set of reviews in the following format:

```
<DOC>
<DATE>06/15/2009</DATE>
<AUTHOR>The author</AUTHOR>
<TEXT>The review goes here..</TEXT>
<FAVORITE>What are my favorites about this car</FAVORITE>
</DOC>
```

Note that each review is enclosed within a <DOC> element as shown above and all the extracted items are within this element.

RELEVANCE JUDGMENT FORMAT

In the judgment folder, you will see three different subfolders (2007,2008,2009) representing the three model years. Each "*.q" file (within these 3 folders) contain a set of queries and the corresponding relevance scores. The format is as follows:

#cat=<category or aspect>,

#query=<query1>;<query ID1>
#query=<query2>;<query ID2>
……..list of queries…….

#judge
relevant judgment file 1; relevance score 1
relevant judgment file 2; relevance score 2
relevant judgment file 3; relevance score 3

HOTEL REVIEWS

DATASET DESCRIPTION

- Full reviews of hotels in 10 different cities (Dubai, Beijing, London, New York City, New Delhi, San Francisco, Shanghai, Montreal, Las Vegas, Chicago) and corresponding aspect ratings used as relevance judgments
- There are about 80-700 hotels in each city
- Extracted fields include **date**, **review title** and the **full review**
- Total number of reviews: **~259,000**

DATA FORMAT

In the data folder, there should be 10 different sub-folders representing the 10 cities mentioned earlier. Each file (within these 10 folders) would contain all reviews related to a particular hotel. You can ignore all the csv files in the data folder. The filename represents the name of the hotel. Within each file, you would see a set of reviews in the following format:

```
Date1<tab>Review title1<tab>Full review 1
Date2<tab>Review title2<tab>Full review 2
.................
.................
```

Each line in the file represents a separate review entry. **Tabs** are used to separate the different fields.

RELEVANCE JUDGMENT FORMAT

In the judgment folder, you will see 10 different subfolders representing the 10 cities. Each "*.q" file (within these 10 folders) contain a set of queries and their corresponding relevance scores. The format is the same as used with Car Reviews.

CITATION REQUEST

If you use this dataset for your own research, please cite the following paper:

Kavita Ganesan and ChengXiang Zhai, "Opinion-Based Entity Ranking", Information Retrieval, 2011.

Bibtex:

```
@article {opinrank,

      title = {Opinion-Based Entity Ranking},
      journal = {Information Retrieval},
      year = {2011},
      keywords = {adhoc multifaceted search, entity oriented search,
      entity ranking, entity retrieval, product search},
      doi = {10.1007/s10791-011-9174-8},
      author = {Kavita Ganesan and ChengXiang Zhai}
}
```