OpinRank Review - Dataset

Author: Kavita Ganesan (kganes2@illinois.edu)

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**
- 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

Format

There are three different folders (2007,2008,2009) representing the three model years. Each file (within these 3 folders) would contain all reviews for a particular car. 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.

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)
- 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**

Format

There should be 10 different folders representing the 10 cities mentioned earlier. Each file (within these 10 folders) would contain all reviews related to a particular hotel. 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.

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}
}
```