1. [Deceptive Opinion Spam Corpus v1.4]

(http://myleott.com/op-spam.html)

**文件名：***op\_spam\_v1.4*

**描述**

Myle Ott等采集了TripAdvisor上最热门的20家旅馆的评论，从中挑选可信评论400条，与众包平台AMT构造的虚假评论400条结合起来构造的黄金数据集

**优点**：一个平衡的数据集

**缺点**：只有评论文本

1. [A reviews collection for varies purpose]

(http://odds.cs.stonybrook.edu/)

**Filename：*yelp collective***

**Description：**

YelpZip is the dataset currently used. It contains 608,598 reviews, and there exist 13.2% filtered reviews.

| **Dataset** | **Size** | **Description** |
| --- | --- | --- |
| [YelpCHI](http://odds.cs.stonybrook.edu/yelpchi-dataset/) | 67,395 hotel and restaurant reviews | Reviews from Yelp.com for Chicago Hotels and Restaurants. |
| [YelpNYC](http://odds.cs.stonybrook.edu/yelpnyc-dataset/) | 359,052 restaurant reviews | Reviews from Yelp.com for NYC restaurants |
| [YelpZip](http://odds.cs.stonybrook.edu/yelpzip-dataset/) | 608,598 restaurant reviews | Zip code wise reviews from Yelp.com for NY, NJ, VT, CT, and PA. |

**Sample review data**

|  |  |
| --- | --- |
| **Date** | 2015-01-02 |
| **label** | -1 |
| **prod\_id** | 4698 |
| **rating** | 1.0 |
| **user\_id** | 250899 |
| **reviewContent** | we went out to the city to visit Rockefeller center during the christmas holidays and I suggested hallal guys on 53rd and 6th. I work in the city so I know my places where you get good cart food. we ordered chicken over rice and lamb over rice. it came out so bland. we were so disappointed. I'm not sure whether its the holiday season rush that they made the once delicious food so bad. |

**Top 7 large product**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Product ID** | **Number of reviews** | **Fake reviews rate** | **begin time** | **end time** |
| ID3745 | 7378 | 0.095418813 | 2008-02-19 | 2015-01-06 |
| ID4698 | 6632 | 0.12017491 | 2006-06-28 | 2015-01-06 |
| ID3237 | 4716 | 0.148642918 | 2005-08-13 | 2015-01-06 |
| ID3136 | 3938 | 0.138141188 | 2005-06-28 | 2015-01-06 |
| ID1881 | 3143 | 0.071905822 | 2004-10-20 | 2015-01-06 |
| ID3875 | 3122 | 0.057975657 | 2005-04-21 | 2015-01-06 |
| ID2605 | 2999 | 0.064688229 | 2010-06-02 | 2015-01-06 |

**Merit：it’s enough construct a time series**

[**Disadvantage**](file:///D:\Program%20Files\Dict\7.5.2.0\resultui\dict\?keyword=disadvantage)**：it’s imbalanced**