Toy data set

Given:

- Data about customers buying goods at an online book shop
- Label:
  - Class zero: people have not sent back their books
  - Class one: people have sent back their books

Format:

- We have log files per day (produced by a web server)
- Every line is one order, represented by a json string

Data Set

Data is given as day wise logs

```
➜ data ls -l return-data
total 8184
-rw-r--r--@ 1 pbaier  113584762  110876 Mar  2 09:59 2017-01-01.txt
-rw-r--r--@ 1 pbaier  113584762  110726 Mar  2 09:59 2017-01-02.txt
-rw-r--r--@ 1 pbaier  113584762  110275 Mar  2 09:59 2017-01-03.txt
-rw-r--r--@ 1 pbaier  113584762  110374 Mar  2 09:59 2017-01-04.txt
-rw-r--r--@ 1 pbaier  113584762  110850 Mar  2 09:59 2017-01-05.txt
```

Data Set

Every line of a daily file is one order in json format

```
➜ fraud-data head 2017-01-01.txt
{"transactionId": 6707871407, "basket": [1], "zipCode": 2196, "
{"transactionId": 3459351507, "basket": [2, 1, 5, 4, 2], "zipCo
{"transactionId": 7881605492, "basket": [0, 4, 5, 1, 4], "zipCo
{"transactionId": 8168380925, "basket": [3, 4, 2, 2, 0, 4, 3],
{"transactionId": 4691340970, "basket": [2, 4, 5], "zipCode": 3
{"transactionId": 8555449630, "basket": [2, 4, 0], "zipCode": 4
{"transactionId": 5083761599, "basket": [1, 1, 1, 1, 1, 3, 3, 0
{"transactionId": 6396332618, "basket": [3, 3, 5], "zipCode": 3
{"transactionId": 2771228668, "basket": [5], "zipCode": 8607, "
{"transactionId": 3339586925, "basket": [2], "zipCode": 7840, "
```

Data Set

One of these jsons:

```
➜ return-data cat 2017-01-01.txt | head -n 1 | jq .
{
  "transactionId": 6630251676,
  "basket": [
    4,
    1,
    5,
    4
  ],
  "zipCode": 3798,
  "totalAmount": 484,
  "returnLabel": 0
}
```

Columns

basket: Array of item categorizes that were bought in this order
→ [4, 1, 5, 4]
= customer bought 2 items of cat. 4 and 1 item of cat. 1 and 1 item of cat. 5

totalAmount = sum of all items items in the basket in euro

transactionId = running number for orders in the system

zipCode = zip code of the customer address