**A Needle in a Data Haystack**

**- Ex. 1 -**

Motti Gold - 301260105

Yes:

Itemsets: A = {a}, B = {b}, C = {c}.

Baskets:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Iter. 1 | a – 0.5 | b – 0.6 | c – 0.7 | d – 0.8 | e – 0.5 |  |  |  |  |  |
| Iter. 2 | ~~ab – 0.2~~ | ac – 0.4 | ad – 0.3 | ae – 0.3 | bc – 0.4 | bd – 0.5 | ~~be – 0.1~~ | cd – 0.5 | ce – 0.4 | de – 0.4 |
| Iter. 3 | ~~acd – 0.2~~ | ~~ace – 0.2~~ | ~~ade – 0.2~~ | bcd – 0.3 | cde – 0.3 |  |  |  |  |  |

No itemsets in iteration 4, since any such itemset must include bcd\cde with one more item, which include a subset with confidence lower than 0.3 (that already calculated).

\\ array of rules above the threshold ‘minconf’

We note that:

* + 1. an itemset X is in a basket numbered as j, iff divides j.
    2. for a basket numbered as j, j is also the largest number (item) in j.

example:

We note that .

So, all numbers (items) smaller than 20 are frequent, and all pairs of the form .

We denote by the the largest number (item) of an itemset A. Using the observation in the previous article, we note that:

* + 1. any itemset A s.t. is not maximal since there is a superset B s.t. .
    2. Any itemset A s.t. is not frequent since
    3. For any any item in the itemset is a dividor of , otherwise .

Using this observations, we conclude that the maximal itemsets are:

We can see that only the article c will change, since the mean of each vector will change. Articles a and b, on the other hand, are robust to addition of zeros since they the norm is not changed (b) and since Jaccard only takes the rated values.

I wrote a crawler on <http://www.divers-supply.com>, one of the biggest website for diving\snorkeling\swiming gear.

Since it’s a huge website, I added a variable “products per sub-category” to get a diversity of products within the 50 total products, and the json file includes 50 products with 2 products per sub-category. The code of course designed to crawl all the products of the website.

In addition, I realized that most of the products have no reviews at all, while the products that have review, had only one. On the other hand, the website downloads the reviews with asynchronous ajax requests, which makes the crawler get another several complex request per product, so I didn’t crawl them.

1. About 8 hours.
2. Team members:
   1. Maria Dyshel – [maria.dyshel@mail.huji.ac.il](mailto:maria.dyshel@mail.huji.ac.il)
   2. Sria Louis – [sria.louis@mail.huji.ac.il](mailto:sria.louis@mail.huji.ac.il)
   3. Me – [motti.gold@mail.huji.ac.il](mailto:motti.gold@mail.huji.ac.il)