

implement amazon product crawlers

1. please use following url to query amazon, replace ABC with query from rawQuery3.txt

www.amazon.com/s/ref=nb_sb_noss?field-keywords=ABC

format of feeds file

query, bid, campaignID, queryGroupID

Prenatal DHA, 3.4, 8040,10

2. extract price, product detail url, product image url, category from web page return from url above

3. convert each product to Ads

Note that you need to convert product title to keywords by

a) convert to lowercase

b) tokenize title (split by space)

c) remove stop words (like, a, an, the...)

hint : use lib: `org.apache.lucene.analysis`

```
public class Ad implements Serializable{
    /**
     *
     */
    private static final long serialVersionUID = 1L;
    public int adId;
    public int campaignId;
    public List<String> keyWords;
    public double relevanceScore;
    public double pClick;
    public double bidPrice;
    public double rankScore;
    public double qualityScore;
    public double costPerClick;
    public int position;//1: top , 2: bottom
    public String title; // required
    public double price; // required
    public String thumbnail; // required
    public String description; // required
    public String brand; // required
    public String detail_url; // required
    public String query; //required
    public int query_group_id;
    public String category;
}
```

4. store Ads to file, each ads in JSON format.

hint: use lib `jackson`

for example:

```
{ "category": "Electronics", "query": "home theater system", "description": null,
  "campaignId": 8060, "title": "GPX HT050B 5.1 Channel Home Theater
  Speaker System (Black)", "price": 49.29, "relevanceScore": 0.0, "brand":
  "GPX", "pClick": 0.0, "thumbnail": "https://images-na.ssl-images-
  amazon.com/images/I/41WL3G3OftL.AC\_US160.jpg", "costPerClick": 0.0,
  "bidPrice": 22.0, "query_group_id": 11, "position": 0, "keyWords": ["gpx",
  "ht050b", "5", "1", "channel", "home", "theater", "speaker", "system", "black"],
  "adId": 2005, "detail_url": "https://www.amazon.com/GPX-HT050B-Channel-
  Theater-Speaker/dp/B0084ZYH4I/ref=sr\_1\_7/163-8938683-
  7183664?ie=UTF8&qid=1487468566&sr=8-
  7&keywords=home+theater+system", "rankScore": 0.0, "qualityScore": 0.0 }
```

Bonus points

5. support paging

hint:

you can get 2nd page product by

https://www.amazon.com/s/ref=nb_sb_ss_c_1_6?field-keywords=nikon+d3400&page=2

6. log all exception

grading point

1: 25 points

2: 25 points

3: 30 points

4: 20 points

5: 25 points

6: 25 points