**file\_to\_database.py**

* The purpose of this script is to export Amazon datasets from the website <http://jmcauley.ucsd.edu/data/amazon/> to the application’s SQLite database.
* The script only works with json files, so download the 5-core json you would like to export to the database from the website.
* Move these downloaded files into the /catalog/datasets/ directory.
* Run ***python3 manage.py file\_to\_database*** (note that you should be in project’s directory, meaning you should currently be in the Django\_site/fake\_review\_checker directory so when you ls, you should see the database file).
* Running this command will parse all the json files in the /datasets/ directory, make them into a data frame, and push them to the database.
* If there are any duplicate pk’s, then a message will be sent to the screen stating that the script is resolving the issue.

**minhash.py**

There are 3 methods for detecting fake reviews:

**Step 1: Detection of Duplicate Reviews**

Step 2:Detection of Anomaly in Review Count and Rating Distribution

Step 3:Detection of Incentivized Reviews

* Step 1 is satisfied by running the minHash.py and similarity.py scripts.
* To run minhash.py, use the custom command **python3 manage.py minhash** (note that minHash MUST be run before the similarity.py command, or else it will crash)
* This will create bigram hashes for each review’s text

**similarity.py**

* To run similarity.py, use the custom command **python3 manage.py similarity**

**find\_bigram\_crcs(list *words*)**

***words*** is the actual review text of a Review entity split up into a list of individual words.

**Description:**

Figure 1 shows the steps used in detecting the duplicate reviews.

1. Each review is converted into a set of bigram shingles, which are formed by combining two consecutive words together.
   1. <https://www.elastic.co/blog/searching-with-shingles>

Diagram

Description automatically generated

1. Shingles are given shingle IDs using a CRC32 hash function.
   1. Used for increasing computational efficiency in large datasets.
2. Min-Hash signatures were calculated for each review using a random hash function which prevents from having to explicitly compute random permutations of all the shingle IDs

Diagram

Description automatically generated

**Display.py**

* Amazon.com allowed the sellers to offer products for free or at high discounts in exchange for positive reviews about that product.
* While most of such reviews have a disclaimer that *“the customer received this product in exchange for an honest review”* stated in the reviews, there still exist many reviews written under this condition without including the disclaimer.
* **Incentivized reviewers**, though they claim to be unbiased, tend to give positive and less critical reviews for the products compared with the non-incentivized reviewers.
* The occurrence of these reviews has transformed the review panels into advertising forum. Detecting such incentivized or biased reviews is often more challenging.
* Examples of incentivized reviews:
  + *“I got these at no charge in exchange for an honest review*
  + *I received this product in exchange for a truthful review and I must say that I am overall satisfied with it*
  + *I received this product at a discounted rate in exchange for my fair and honest.”*
* Figure 10 shows the overall steps in the detection of incentivized reviews.

**How are incentivized reviews handled?**

**def detectKeywords()**

**Description:**

* A collection of synonym phrases for a set of key phrases derived from common incentivized reviews are generated using the **Natural Language Toolkit** from **WordNet**.
  + WordNet is a large lexical database which resembles the thesaurus.
  + It groups words together based on their meanings and the specific sense of the words into sets of cognitive synonyms (synsets).
* If a key phrase such as “honest review” is given as input to the function, equivalent phrases are such as “truthful review”, genuine review”, “genuine feedback”, and so on are searched for.
* A dictionary was made of both single and double paired words, for example, “discount” and “no charge”.
* The reviews with these synonyms were identified using the regular expression and the time intervals were also captured for analysis.