https://ci6.googleusercontent.com/proxy/nlUHBEbfZC4jGa6KJwNwgMjf3ygc4ryAjnMSIDja-_rVfm_tYXTgfI-ITMMdo8xeWYVwaC_DhVJ-U_OUTNDJzQ=s0-d-e1-ft#http://feedvisor.com/fv_signature/logo.png

**Proposal for 3rd year project at Feedvisor:**

**Identify Similar Products**

Specification document

**September 2017**

**A. Introduction**

Amazon offers millions of products for sale at any given time. Some of the products are similar to each other in a certain level. Identifying which products are most similar to a source product can be very useful for a seller. This can help in understanding whether the similar products should also be offered for sale by the seller, what is the pricing vicinity the product should be in, what are the keywords that should be used to describe the product, how well is the product positioned in the competitive landscape etc.

Identifying the similar products is not an easy task. Some products can have very similar names, but represent very different products.

For example: “Apple iPhone 8 Plus” which is an expensive cellphone and “Apple iPhone 8 Plus case” which is a cheap plastic case.

We would like the students to develop an algorithm that finds the most similar products to any given source product.

**B. Project scope**

**B.1. Input**

We will provide 3 files as input:

1. XML file with raw data of a large number of products sold on Amazon (tens of thousands) ([**File\_1\_RawProductsFor3rdYearProject .zip**](https://drive.google.com/open?id=0BzpkbIUKxfmydUR5RGVFbF9Sc00)).

2. Text file with the product ids (ASIN) of 100 products for which we would like to find similar products in the large XML file ([**File\_2\_SourceProductsFor3rdYearProject.txt**](https://drive.google.com/open?id=0BzpkbIUKxfmyXzhjYWVUT29kMm8)).

3. Text file with the word frequency found in a very large number of products sold on Amazon ([**File\_3\_WordFrequencyInProductsFor3rdYearProject.txt**](https://drive.google.com/open?id=0BzpkbIUKxfmyVXZjSTc1ckUtdVE)).

For each word we have two values:

A. **products** – The number of products in which the word appears in (at least once).

B. **inverse\_product\_frequency** – Natural log (ln) on the ratio between the total number of products and the number of products in which the word appears.

This data can help determine the relative importance of a word in its product name

**B.2. Algorithm**

We would like the students to develop an algorithm that finds the most similar products to any given source product. For each found product we want a similarity score between 0 and 1 to be noted. A certain threshold between 0 and 1 should be defined, so products that have a similarity grade below that threshold, would not be included in the results as they are not similar enough.

We define a similar product as one whose noun, purpose and functionality are the same/similar to the source product.

Each product is described by many fields, including two specific ones: **Brand** and **Manufacturer**.

We would to conduct **two different searches**. One that can find products with the same brand and/or manufacturer and one that limits the search only for products that belong to different brands and manufacturers.

For example:

Say the source product is “Apple iPhone 8 Plus”. In the first search we would like to find other Apple iPhones, e.g. “Apple iPhone 7 Plus”, “Apple iPhone 6s Plus” etc.

In the second search we would like to find other non-Apple cellphones, e.g. “Samsung Galaxy J7 Prime”, “LG G6 H870DS”, “Sony Xperia XZ Premium” etc.

In both searches we would not like to find products that are a plastic cases, connectors, laptops, t-shirts, oranges etc.

There can be different aspects that reflect the similarity between products, and each can be explored by algorithms from different disciplines.

The quality of the results is probably related to the exploration results of each discipline and their weighted aggregation.

Some of the disciplines we believe are relevant are:

- Text similarity analysis

- Numerical values similarity analysis (weight, width, length, height, price etc.)

- Word frequency analysis in English language and product names, required to identify important words.

- Synonym analysis (e.g. cap/hat, pants/trousers etc.)

- Product image processing (Identify similar/identical images)

**B.3. Output**

The requested output are two txt files (for searches with brand/manufacturer and without) with all of the source products, and for each one a list of the similar products found ordered by their descending similarity grade.

For each product we expect to get about 5-50 similar products, depending on the product.

In the first search that does not limit brand/manufacturer we naturally expect to find more similar products.

**B.4. Input files examples**

1. XML file with raw data of a large number of products sold on Amazon

|  |
| --- |
| <products\_data>  <product\_data>  <GetMatchingProductForIdResponse xmlns="<http://mws.amazonservices.com/schema/Products/2011-10-01>"><GetMatchingProductForIdResult Id="B013HATW1W" IdType="ASIN" status="Success"><Products><Product><Identifiers><MarketplaceASIN><MarketplaceId>ATVPDKIKX0DER</MarketplaceId><ASIN>B013HATW1W</ASIN></MarketplaceASIN></Identifiers><AttributeSets><ns2:ItemAttributes xml:lang="en-US" xmlns:ns2="<http://mws.amazonservices.com/schema/Products/2011-10-01/default.xsd>"><ns2:Binding>Kitchen</ns2:Binding><ns2:Brand>BNYD</ns2:Brand><ns2:Color>Chrome</ns2:Color><ns2:Feature>Keeps bananas off counter to prevent bruising</ns2:Feature><ns2:Feature>Allows for even ripening</ns2:Feature><ns2:Feature>Wipe with damp cloth and towel dry or do not put in dishwasher</ns2:Feature><ns2:Feature>High grade steel wire construction. Chrome finish</ns2:Feature><ns2:Feature>Size: 12"H x 6"W x 6"D</ns2:Feature><ns2:MaterialType>Chromium Steel</ns2:MaterialType><ns2:PackageDimensions><ns2:Height Units="inches">6.00</ns2:Height><ns2:Length Units="inches">13.00</ns2:Length><ns2:Width Units="inches">7.00</ns2:Width><ns2:Weight Units="pounds">0.50</ns2:Weight></ns2:PackageDimensions><ns2:PartNumber>B013HATW1W</ns2:PartNumber><ns2:ProductGroup>Home</ns2:ProductGroup><ns2:ProductTypeName>HOME</ns2:ProductTypeName><ns2:SmallImage><ns2:URL><http://ecx.images-amazon.com/images/I/413gTvCFG1L._SL75_.jpg></ns2:URL><ns2:Height Units="pixels">75</ns2:Height><ns2:Width Units="pixels">66</ns2:Width></ns2:SmallImage><ns2:Title>Banana Hanger, Banana Holder, Banana Stand, Grape Hanger</ns2:Title></ns2:ItemAttributes></AttributeSets><Relationships></Relationships><SalesRankings><SalesRank><ProductCategoryId>home\_garden\_display\_on\_website</ProductCategoryId><Rank>247956</Rank></SalesRank><SalesRank><ProductCategoryId>16353551</ProductCategoryId><Rank>393</Rank></SalesRank></SalesRankings></Product></Products></GetMatchingProductForIdResult><ResponseMetadata><RequestId>cbe61d64-23a1-4955-9737-2e54d8123039</RequestId></ResponseMetadata></GetMatchingProductForIdResponse>  </product\_data>  <product\_data>  <GetMatchingProductForIdResponse xmlns="<http://mws.amazonservices.com/schema/Products/2011-10-01>"><GetMatchingProductForIdResult Id="B00PHUHSLC" IdType="ASIN" status="Success"><Products><Product><Identifiers><MarketplaceASIN><MarketplaceId>ATVPDKIKX0DER</MarketplaceId><ASIN>B00PHUHSLC</ASIN></MarketplaceASIN></Identifiers><AttributeSets><ns2:ItemAttributes xml:lang="en-US" xmlns:ns2="<http://mws.amazonservices.com/schema/Products/2011-10-01/default.xsd>"><ns2:Binding>Health and Beauty</ns2:Binding><ns2:Brand>Nature's Bounty</ns2:Brand><ns2:Feature>Hair, Skin &amp; Nails Support.</ns2:Feature><ns2:Feature>With Biotin.</ns2:Feature><ns2:Feature>Skin &amp; Hair Nutrients.</ns2:Feature><ns2:Feature>Antioxidants Vitamin C &amp; E.</ns2:Feature><ns2:Feature>220 Gummies Strawberry Flavored.</ns2:Feature><ns2:Flavor>Strawberry</ns2:Flavor><ns2:ItemDimensions><ns2:Height Units="inches">6.90</ns2:Height><ns2:Length Units="inches">3.40</ns2:Length><ns2:Width Units="inches">3.40</ns2:Width><ns2:Weight Units="pounds">1.44</ns2:Weight></ns2:ItemDimensions><ns2:IsAdultProduct>false</ns2:IsAdultProduct><ns2:Label>Nature's Bounty, Inc.</ns2:Label><ns2:Manufacturer>Nature's Bounty, Inc.</ns2:Manufacturer><ns2:PackageDimensions><ns2:Height Units="inches">3.40</ns2:Height><ns2:Length Units="inches">6.90</ns2:Length><ns2:Width Units="inches">3.40</ns2:Width><ns2:Weight Units="pounds">1.45</ns2:Weight></ns2:PackageDimensions><ns2:PackageQuantity>1</ns2:PackageQuantity><ns2:PartNumber>na</ns2:PartNumber><ns2:ProductGroup>Health and Beauty</ns2:ProductGroup><ns2:ProductTypeName>HEALTH\_PERSONAL\_CARE</ns2:ProductTypeName><ns2:Publisher>Nature's Bounty, Inc.</ns2:Publisher><ns2:SmallImage><ns2:URL><http://ecx.images-amazon.com/images/I/41VRbcguneL._SL75_.jpg></ns2:URL><ns2:Height Units="pixels">75</ns2:Height><ns2:Width Units="pixels">36</ns2:Width></ns2:SmallImage><ns2:Studio>Nature's Bounty, Inc.</ns2:Studio><ns2:Title>Nature's Bounty Optimal Solutions Hair, Skin and Nails Gummies 220 Count With Biotin Strawberry Flavored</ns2:Title></ns2:ItemAttributes></AttributeSets><Relationships></Relationships><SalesRankings><SalesRank><ProductCategoryId>health\_and\_beauty\_display\_on\_website</ProductCategoryId><Rank>20849</Rank></SalesRank><SalesRank><ProductCategoryId>6939008011</ProductCategoryId><Rank>73</Rank></SalesRank></SalesRankings></Product></Products></GetMatchingProductForIdResult><ResponseMetadata><RequestId>db1694be-caa8-44aa-ae96-199043c408c7</RequestId></ResponseMetadata></GetMatchingProductForIdResponse>  </product\_data>  .  .  .  </products\_data> |

Note that some of the fields that describe a product appear only in some of the products. For example <ASIN> will appear for all products, but <ns2:Studio>, only for the products it is relevant for.

The most interesting fields that appear for most of the products are:

|  |
| --- |
| <MarketplaceId>  <ASIN>  <ns2:Title>  <ns2:Binding>  <ns2:Brand>  <ns2:Manufacturer>  <ns2:ProductGroup>  <ns2:ProductTypeName>  <ns2:URL><ns2:SmallImage> |

Also note that some of the fields have an inner tree structure, for example inside <ns2:SmallImage> there is <ns2:URL> but also the dimensions of the image: <ns2:Width> and <ns2:Height>.

2. Text file with the product ids (ASIN) of 100 products for which we would like to find similar products in the large XML file.

|  |
| --- |
| B013HATW1W  B000XAPLKQ |

3. Text file with the word frequency found in a very large number of products sold on Amazon.

|  |
| --- |
| word products inverse\_product\_frequency  black 1585422 2.6794  women's 1306231 2.8731  for 1262622 2.9071  the 1231785 2.9318  of 1175904 2.9783  size 1055345 3.0864  and 969669 3.1711  with 956674 3.1846  5 930428 3.2124  men's 874906 3.2739  2 828672 3.3282  t-shirt 817992 3.3412  in 760261 3.4144  us 744262 3.4357  m 741967 3.4387  1 705472 3.4892  by 678141 3.5287  white 645619 3.5778 |

**B.5. Output files examples**

|  |
| --- |
| B013HATABC  ========  <ASIN>B013HABCD1</ASIN> = 0.9829  <ASIN>B013HABCD2</ASIN> = 0.7642  <ASIN>B013HABCD3</ASIN> = 0.5569  <ASIN>B013HABCD4</ASIN> = 0.5473  <ASIN>B013HABCD5</ASIN> = 0.5329  <ASIN>B013HABCD6</ASIN> = 0.5125  <ASIN>B013HABCD7</ASIN> = 0.3421  <ASIN>B013HABCD8</ASIN> = 0.3238  <ASIN>B013HABCD9</ASIN> = 0.2844  B005UG0XYZ  ========  <ASIN>B00AIF9A01</ASIN> = 0.8257  <ASIN>B00AIF9A02</ASIN> = 0.7734  <ASIN>B00AIF9A03</ASIN> = 0.6523  <ASIN>B00AIF9A04</ASIN> = 0.6166  <ASIN>B00AIF9A05</ASIN> = 0.5323  <ASIN>B00AIF9A06</ASIN> = 0.4599  <ASIN>B00AIF9A07</ASIN> = 0.3721  . . . |

Note that in this example a threshold of 0.3 was defined, so any product with a similarity score below 0.3 were omitted.