**CHOOSING CATEGORIES**

1. Use human chosen **key words** and their synonyms, variations (i.e. regex pattern) to identify category
   1. Chose broad product categories
      1. Toiletries
         1. Bodywash (bar soap/shower gel/showering cream)
         2. Shampoo,
         3. Razor
         4. Conditioner
         5. Shaving cream
         6. Perfume/cologne
      2. Toys

**IDENTIFYING GENDER**

* + - 1. Word embeddings (Cyril) (scales up with sample)
      2. ADM (Aleksandra) (scales up with sample)
      3. Image recognition (Ignacio) (scales up with sample)
      4. Use the explicit word in title and breadcrumb (scales up with sample)

Assessment against human coded random sample of 200 products

**CLEANING UP PRICES**

Does not automate with scaling up

Cleaned up “by hand” – and rescaling the products so they are on the same scale, ended up with 5 types of pricing:

1. Units
2. Kilograms
3. Litres
4. Metres
5. Sheets

**HOW DO WE NARROW SCOPE?**

1. focus on just one store
2. limit to a specific brand:

**POTENTIAL ADDITIONAL ROUTES:**

1. Analyse time trends to see if female products are more manipulated with discounting than male

* Need to verify if loyalty and discount prices appear in toiletries product category