# Report 1

SNumbers: u264332, u264443, u264202

Names: Levente Olivér Bódi, Riccardo Zamuner, Giada Izzo

## Introduction

## Code

* Section 1:
  + Nltk stopwords
  + Necessary imports
  + Preprocessing functions according to part1, assignment1
* Section 2: we keep in mind, mention possible drop of crawled\_at, images
* Section 3 – copy text from code here + indicate why we don’t merge the fields
* Section 4
  + Read in dataset
  + Plot distributions, analyze them
  + Check missing values – 855 amount impute discount with 0, 777 amount impute actual price, 2 drop bc no selling price
  + Clean seller – no avg rating, and keep avg rating column for now as it is

### Part 1

We would like to reflect on what we did in the code.

#### Section 1

For Section 1, we imported the necessary libraries first, alongside the nltk stopwords package. After, we made 2 functions, one for preprocessing textual data, and one for preprocessing numerical data. These are combined in the preprocess\_document function, which preprocesses with the help of the functions described above. In the preprocess\_text function we tokenize the text, convert it to lower case, eliminate stopwords, stem it, remove punctuation, just as described in the assignment.

#### Section 2

The dataset contains 2 columns, namely crawled\_at, and images, which the assignment does not mention in the compulsory return fields, however, for now, we keep them. The fields are read into a pandas dataframe in Part 2.

#### Section 3

Pros of keeping separate: end user can easily filter products (i.e. if category / sub-category / brand fully matches user filter, product is relevant, otherwise it's not)

Cons of keeping separate: more computation to check general relevance of products

Pros of merging: every aspect of the product helps with relevance (e.g. user can query (not filter) for brands, and the most relevant products will be the one with the user requested brands, but other products, which might still be useful for the user, could still be relevant)

Cons of merging: you lose the option to filter

Maybe some fields could be saved in the inverted index tree using one-hot encoding, to try a hybrid approach, depending on how one wants to build its retrieval engine

We choose not to merge fields. It is understandable that it costs more computationally, however, we aim for better precision this way.

#### Section 4

We reflect on the section for Part 2, after doing the exploration.

### Part 2

* Clean seller – no avg rating, and keep avg rating column for now as it is
* In the exploratory data analysis we read in the dataset and observe it.
* TODO: Plot distributions, analyze them

After checking the plots and exploring the data, we found out that the dataset contains 855 fields which do not have a discount, 777 fields which do not have actual price, and 2 fields which do not have selling price. We impute 0 for the discount, because we hypothesize that if there is no data on discount, there is no actual discount existing. After that, we impute the selling prices to the missing actual price fields, since if discount = 0, then selling price = actual price. The 2 rows which do not have a selling price we drop, since price is crucial information.

Another column which has a lot of missing records, namely 2259, is average rating. This we will leave for now, since we do not know yet what we need it for and we do not think it is sensible either to drop, or to impute it with zeroes or averages of other ratings.

* TODO: reflect on other cleaning which Riccardo did