**Use case Selected**

*“Dentro del marketplace existen productos similares o idénticos entre sí (son*

*productos vendidos por distintos sellers, en la api puedes obtener y descargar los*

*títulos e incluso las imágenes!). ¿Cómo buscar dichos ítems para agruparlos y*

*volverlos comparables entre sí? Esto permitiría mejorar la experiencia ante muchas*

*opciones similares.”*

**TitleSimilarityMeLi**

Project of title cosine similarity between products in MeLi

**Exploration with multiple analysis**

(fixed examples, only to explore, not to be run again):

* Tf-Idf and Cosine similarity
* Word Embeddings (word2vec) and Cosine similarity
* Plot titles proximity (of title embedding) in 2D with PCA, TSNE, (with colors both from KMeans and original Categories)

**Implementation with selected analysis**

(meant to be run multiple times as a train/test solution, to get new examples):

* Selected Algorithm: Word Embeddings and Cosine similarity (due to simplicity of implementation, more robust algorithm and fastness/smaller cosine matrix)
* Selected Plot: 2D with PCA and TSNE (only original Categories, for better understanding)

**Pickles**

* Historical products for comparison and to minimise computation time
* PCA for plot purposes, explain model importance and way of working to line of business

*Observations: word embeddings are downloaded from web, since the file is too heavy to upload to git repo*

**Problem:**

The problem to be solved is to find similar products for the current one the potential customer is looking at. Products both in the same category and 2 alternative categories, that might have some value to explore, to maximise the sale.

Hypotheses:

* If the customer has access to very similar products (more than 80% of cosine similarity according to initial explore, or with a maximum of X fixed selected items) in the same category might be helped in the selection of current article giving the potential customer access to slightly different articles
* If the customer has access to products with high some similarity from other categories might be tempted to buy other articles that have some relation to the one he/she is currently looking at.

**Information:**

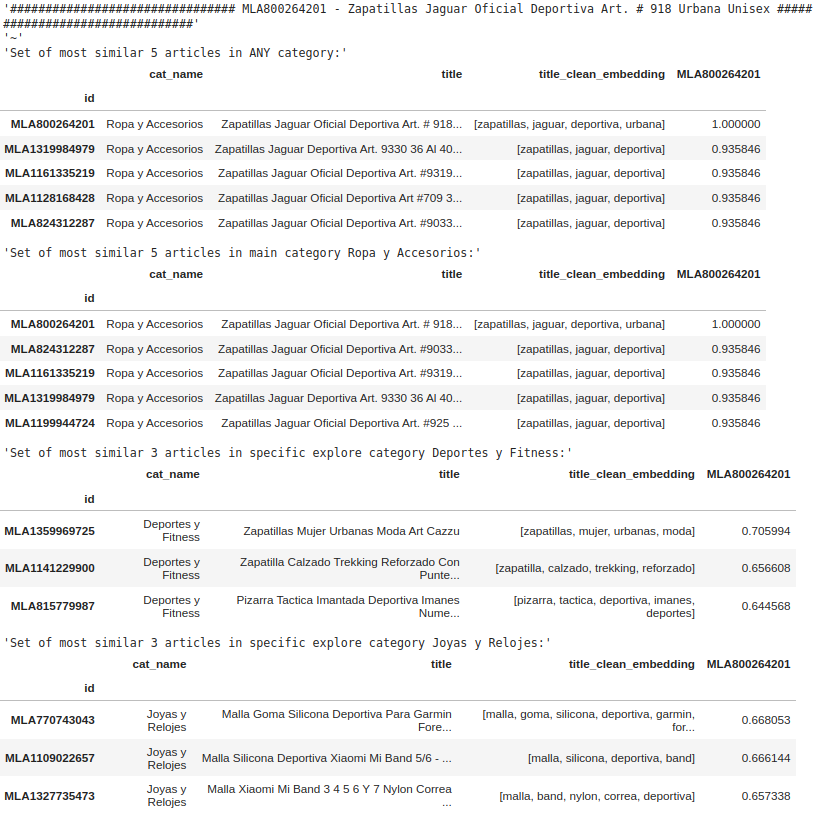
The products selected were around 1000 per category.

* Main category 'Ropa y Accesorios'
  + Hypothesis: clothes is a big category (variety of products different products for women, men and children) that could be combined with others
* Alternative categories 'Deportes y Fitness' + 'Joyas y Relojes'
  + Hypothesis: that might have some relation with the main category. ( “women clothes → women sports accessories + women watches”)

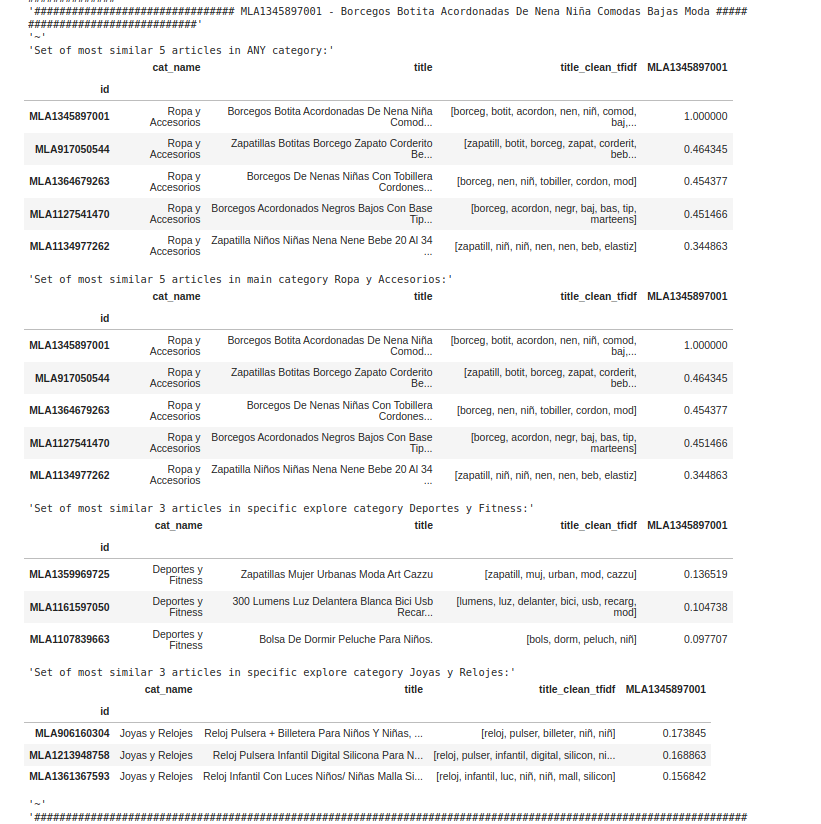
**Insights:**

* Products related to some areas like: women, men, kids, urban, sport, medical, work, gave interesting combinations with alternative categories.
* 'Deportes y Fitness' examples:
  + men trainers → football shorts
  + bra →women sport tights
  + women sporty trainers → cycling women accessories
  + kids jacket or boots → kids sleeping bag or kids bikes accessories
* 'Joyas y Relojes' examples:
  + sporty trainers → sporty watches
  + urban trainers → urban watches
  + kids clothes → kids watches
  + kids clothes → mother jewellery
* Very specific or rare products, or products made from other very specific materials, might not get as interesting mixes with other categories (ex: “rain pilot”).
* Figure:
  + In top the article analysed, (between “#”)
  + After 5 most similar articles in all 3 categories (almost always from original category)
  + After 5 most similar articles in original category (Ropa y Accesorios)
  + After 3 most similar articles in 1st alternative category (Deportes y Fitness)
  + After 3 most similar articles in 2nd alternative category (Joyas y Relojes)
* It found similar products (trainers) ordered by cosine similarity (descending) in same category (slightly different) and other less similar associated to sport category since there were sporty/urban trainers, and sporty watches too.

Ex1: Sporty trainers



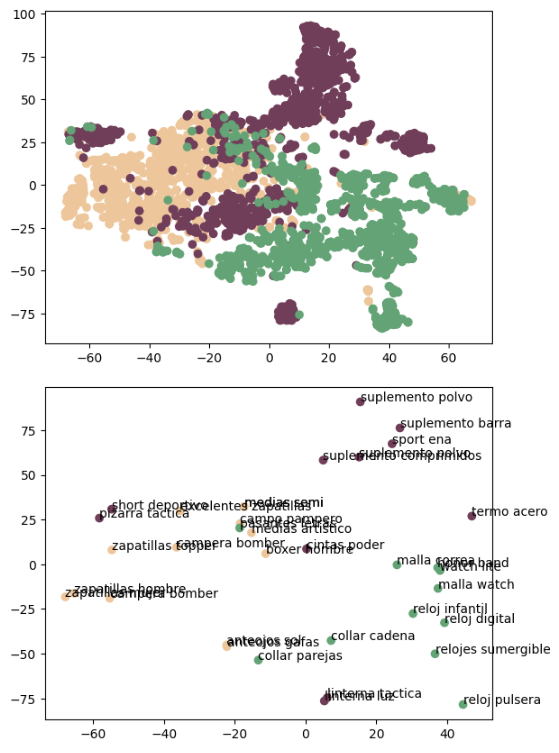
Ex2: Kids Boots

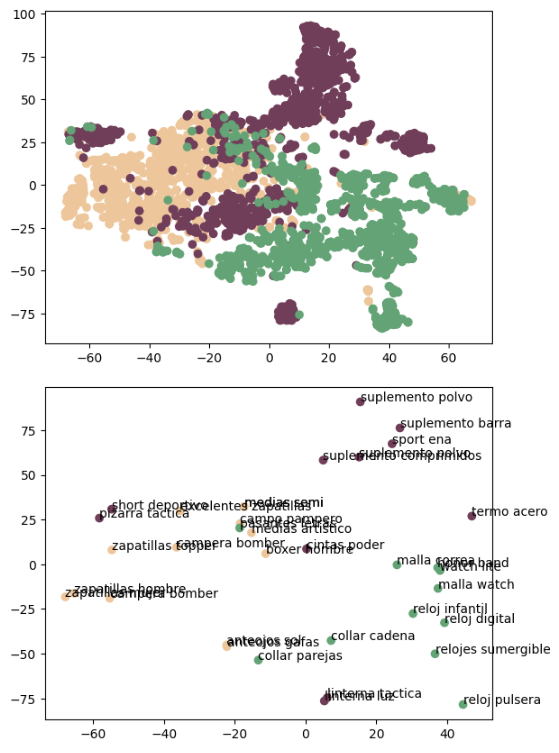


***Methodology and Algorithms applied:***

* *Title preprocessing of words: lower case, remove symbols, remove stop words, remove words with no value in similarity context (related to the delivery or colours)*
* *Word importance and meaning:*
  + *Tf-idf was used to check word importance. This is a simpler approach but slower and too fixed to training data → not selected for final solution*
  + *Embeddings were used to understand word meaning. Faster approach (smaller vector representation) and simpler implementation for a variety of words (captures similarity between different words that might not be part of original training ex. “Mujer”, “Dama”, .. ) → selected for final solution*
* *Title similarity:*
  + *Word vectors (from each title) were summed up into one vector to represent the “product title representation” as a simple approach in both cases*
  + *Cosine similarity between product vectors to order from most to least similar product*
* *Plots for better Line of Business Explainability*
  + *PCA and TSNE over product vectors to reduce variables and understand weights in each products, and plot results in 2 dimesions*
  + *Kmeans over PCA of product vectors to try to find different groups of categories, given more time and training might have found better clusters.*
  + *Original Categories used to show intersections between different categories*

**TSNE** plot with colours of **3 different categories** analysed to show intersection between categories. And some examples (first 2 words of title used in embedding) to show how similar products are close to each other in a very graphical way (2D representation).





**Final solution, metrics and conclusion:**

Since this solution has a very **unsupervised** **approach**, there is not a specific metric. From exploratory analysis a good measure of cosine similarity could be 80%-100% to be on the safe side. Anyway there are products very similar but lower scores. 80% minimum similarity could be an initial standard to suggest to the line of business to start a pilot with, or a fixed amount of similar items with most similarity.

The final solution is trained on 80% of the data, and tested on 20% never seen by the model. It generates a historical database of train product representation (embeddings).

The idea is to get the best similarities of new products (test) among themselves and with historical products (train). These similarity scores are ordered from highest to lowest, in general and by categories to give both true similar products and other category similarities the customer might be interested in.

This approach is meant to be fast on new products.

**Next steps:**

This **solution could be piloted** and with some **feedback** from customers (clicks/buys of items suggested) a **predictive model** could be trained, **to refine % of similarity** by category.

Also, **more specific clusters** (jackets, trousers,.. or women, men, kids, .. etc) could be created for other purposes.