E-commerce Market Analysis for Zalando using Machine Learning Techniques

Nithin Vincent Thayyil
Dublin City University
21261508
nithin.thayyil2@mail.dcu.ie
https://github.com/nithin8145/Zalando_machine_learning

ABSTRACT

Social media has become a vital tool in our day-to-day life and by-default e-commerce is also playing a huge role in humans' life. People can purchase things through any website with just a click of a button. In this assignment, DCU Computing department (Machine learning course) in collaboration with Zalando company has provided a dataset to compare their products with their competitive company (About you) and eventually find the similarity and match products with the help of F1 score using various Machine learning techniques such as fuzzy way logic, TfIdf, and cosine similarity to determine the similarity so as to provide competitive prices to their customers and to drive revenue growth.

Keywords- ***TfIdf, cosine similarity, Machine learning, Zalando, similarity, F1 Score.***

1- INTRODUCTION

As we all know, social media and the Internet have been a boom, especially with Covid and people doing most of the work on the Internet. E-commerce is one such platform for which every customer is demanding more than the current trend.

In this assignment, we have taken the leading online platform Zalando as our e-commerce company and compared it with relevant European competitors (e.g., About you) to find the similarity between the product and identify the product matches. The assignment was in collaboration with Zalando to identify which products match About you's products and simultaneously find the similarity so that the company can offer competitive prices for their products to their customers.

Although many companies focus on what they can do to improve their products by comparing with other competitive company's products based on unique barcode systems like EAN which is not always available (so the process is a little tedious)

So, focusing on the current disadvantages the dataset provided to us in this challenge is based on the 'title', 'color', 'price', and so on due to which similarity can be easily calculated based on the columns provided in the dataset. The dataset contains the products of both Zalando and About you

The primary objective of this paper is to find which products in Zalando match the products in About you with the help of F1 score and to find the similarity score for the same using 'fuzzy way logic', TfIdf, and cosine similarity

The rest of the paper is organized as section 2 discusses the related work in the current field and the methodology is provided in section 3, the approach through which the project is executed is provided in Section 4, Results and analysis are presented in section 5, and finally, Section 6 concludes the paper.

2- RELATED WORK

Much work has been done in the field related to Machine Learning and NLP techniques. One of the papers [1] proposes how to empower Indian women entrepreneurs in e-commerce clothing based on the Machine Learning algorithm in which they have used SVM and REPTree classifiers to classify the customer review on women's clothing e-commerce and has managed to achieve an accuracy of 91.43%

In the [2] paper, the authors propose how the ML algorithm can be used to discover patterns between the data using 'Fuzzy way logic' and generating a set of rules to improve the searches on e-commerce websites and preferably suggest users what to buy based on the rules.

In [3] paper the authors have described their work performed in a competition. They have used the F1 score to judge the prediction results which were analyzed in a dataset named 'Tmalls' (the largest B2C online retail platform in China). Due to the massive amount of data, the authors had processed the dataset in MapReduce and obtained an 'F1 Score' of 3.13% for the rule-based model, with precision being 2.72% and the recall rate being 3.69%.

3- METHODOLOGY

3.1 Fuzzy Wuzzy Logic

Before getting to know about Fuzzy Wuzzy logic, we must first understand what is Levenshtein distance calculation. Levenshtein distance basically means a metric which can be used for fuzzy string matching by calculating the distance between 2 strings (in simple terms, how different 2 text are) [4]

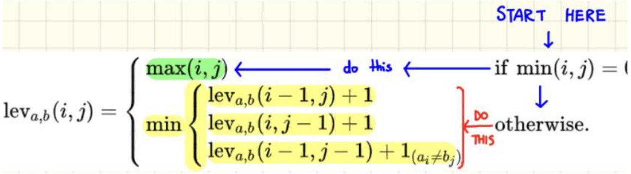

START HERE
do this
DO THIS
$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j)+1 \\ \operatorname{lev}_{a,b}(i,j-1)+1 \\ \operatorname{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$

Figure 1. Levenshtein-distance Equation [4]

So, with the help of Levenshtein distance, Fuzzy Wuzzy returns the similarity score with respect to the 2 texts. There are more functions in Fuzzy Wuzzy that can perform string matching such as Partial Ratio, Token Sort Ratio, and so on [5].

3.2 TfIdf and Cosine Similarity

TfIdf (Term frequency- Inverse document frequency) is basically used in Information Retrieval. Tf generally means that in a sentence how often a term appears and Idf generally provides us with the insight about the uniqueness of a term in the sentence [6]. It is used in many different ways such as Machine Learning and Natural Language Processing, in Bag of Words, and so on. Although its biggest advantage is that it is simple to calculate but it cannot prove the meaning (the content) of the sentence.

When it comes to cosine similarity, its almost similar to Fuzzy Wuzzy Logic and TfIdf, but instead of a text, cosine similarity measure the similarity between 2 documents [7].

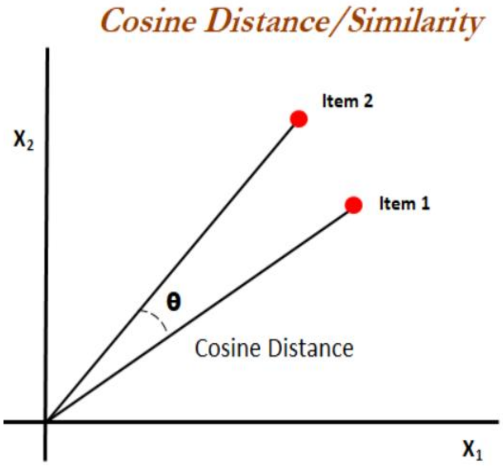


Figure 2. Cosine Similarity Explained [7]

3.3 F1-Score

F1-Score is an important concept with respect to the assignment due to it being the determining factor whether the products are matching or not. F1 Score is the mean of recall and precision and its range is between 0 and 1, 1 being the best score and 0 being the worst score [8].

To understand F1 Score we first have to understand what Confusion matrix is and the related terms with respect to that. Confusion matrix is a performance measurement which consist of 4 combinations namely: 1-True positive, 2-False Positive 3- True negative 4- False negative [9].

4- APPROACH

The following approach has been taken for this assignment and they are as follows

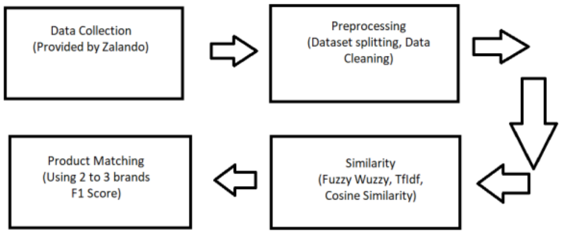

Data Collection
(Provided by Zalando)
Preprocessing
(Dataset splitting, Data Cleaning)
Similarity
(Fuzzy Wuzzy, TfIdf, Cosine Similarity)
Product Matching
(Using 2 to 3 brands F1 Score)

Figure 3. Approach Taken

(Steps involved in the proposed solution)

4.1- Data Collection

3 types of dataset were provided to us by Zalando for the assignment namely

1- offers_training.parquet
2- offers_test.parquet
3- matches_training.parquet

In this assignment, the 1st dataset was used for the processing and further analysis. The goal was to find the similarity with the training dataset and then proceed with the 2nd dataset and the 3rd dataset to verify whether the similarity of the 2 companies are matching or not

The dataset consists of over 100000 rows of data and 10 columns describing the 'title', 'color', ' price', and so on about each product from both companies.

4.2- Preprocessing

As the dataset was provided by the Zalando company and it was private (no one outside the company or from DCU could have accessed it) there was minimal data cleaning work for this assignment but data preprocessing is an important step before analyzing it through Machine learning algorithm cleaning of data had to be done.

Firstly, necessary packages were installed and imported for the code to run. Then the parquet file was called in using a jupyter notebook and preprocessing steps started. We began by splitting the dataset into 2 datasets namely 1- Zalando, 2- About you

After splitting the data, we combined a few columns (‘title’, ‘color, ‘brand’) that were required for further analysis and dropped the columns which weren’t useful. Then we cleaned the data by removing special characters from the combined column and making all the values in the dataset lowercase and making the dataset available for similarity and product matching.

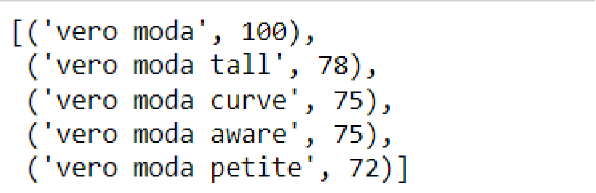
For Fuzzy Wuzzy Logic and TfIdf & Cosine Similarity, we haven’t split the data but proceeded with the training data by cleaning it and removing special characters and columns which aren’t required.

5- RESULT & ANALYSIS

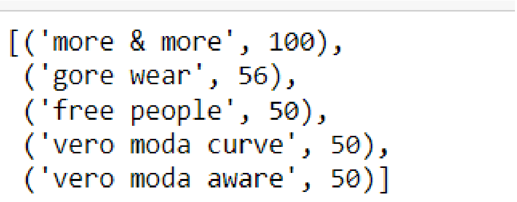
5.1- Fuzzy Wuzzy Logic

For Fuzzy Wuzzy Logic training dataset without splitting is taken into consideration for determining the similarity of the text.

Here 2 brands have been considered for Fuzzy Wuzzy Logic namely, 1-vero moda and 2- more & more

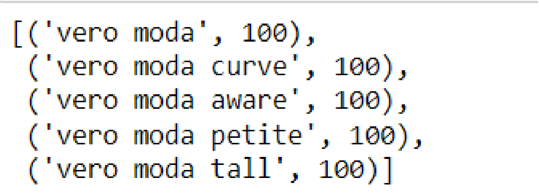


```
[('vero moda', 100),
 ('vero moda tall', 78),
 ('vero moda curve', 75),
 ('vero moda aware', 75),
 ('vero moda petite', 72)]
```

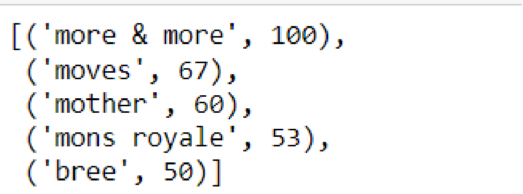


```
[('more & more', 100),
 ('gore wear', 56),
 ('free people', 50),
 ('vero moda curve', 50),
 ('vero moda aware', 50)]
```

Figure 4. Brand ‘vero moda’ and ‘more & more’ similarity using fuzz.token_sort_ratio



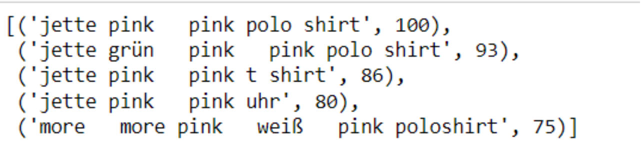
```
[('vero moda', 100),
 ('vero moda curve', 100),
 ('vero moda aware', 100),
 ('vero moda petite', 100),
 ('vero moda tall', 100)]
```



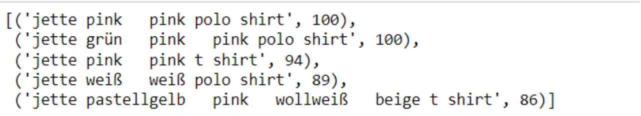
```
[('more & more', 100),
 ('moves', 67),
 ('mother', 60),
 ('mons royale', 53),
 ('bree', 50)]
```

Figure 5. Brand ‘vero moda’ and ‘more & more’ similarity using fuzz.token_set_ratio

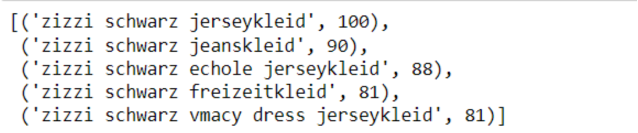
And considering the combined columns we have determined the Fuzzy Wuzzy for few texts in combined column also



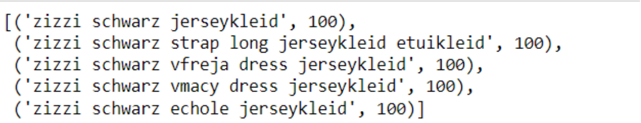
```
[('jette pink   pink polo shirt', 100),
 ('jette grün   pink   pink polo shirt', 93),
 ('jette pink   pink t shirt', 86),
 ('jette pink   pink uhr', 80),
 ('more   more pink   weiß   pink poloshirt', 75)]
```



```
[('jette pink   pink polo shirt', 100),
 ('jette grün   pink   pink polo shirt', 100),
 ('jette pink   pink t shirt', 94),
 ('jette weiß   weiß polo shirt', 89),
 ('jette pastellgelb   pink   wollweiß   beige t shirt', 86)]
```



```
[('zizzi schwarz jerseykleid', 100),
 ('zizzi schwarz jeanskleid', 90),
 ('zizzi schwarz echole jerseykleid', 88),
 ('zizzi schwarz freizeitkleid', 81),
 ('zizzi schwarz vmacy dress jerseykleid', 81)]
```

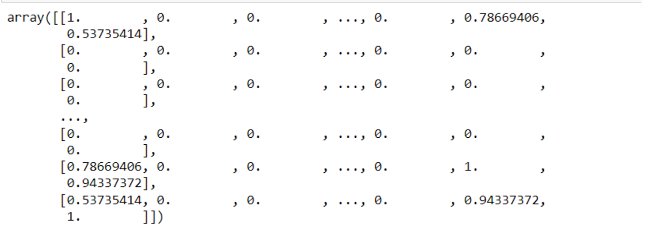


```
[('zizzi schwarz jerseykleid', 100),
 ('zizzi schwarz strap long jerseykleid etuikleid', 100),
 ('zizzi schwarz vfreja dress jerseykleid', 100),
 ('zizzi schwarz vmacy dress jerseykleid', 100),
 ('zizzi schwarz echole jerseykleid', 100)]
```

Figure 6. Updated column with ‘jette pink pink polo shirt’ and ‘zizzi schwarz jerseykleid’ similarity using fuzz.token_sort_ratio and fuzz.token_set_ratio

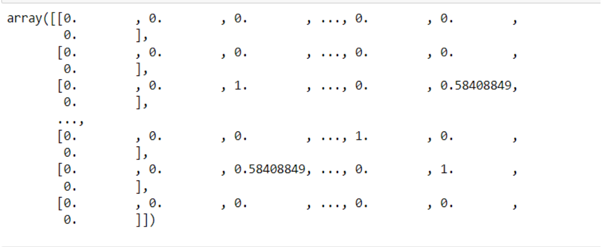
5.2- TfIdf and Cosine Similarity

2 Brands were considered for determining the similarity (If we considered the whole dataset with all the brands then the kernel was crashing so had to minimize the dataset for the process) namely 1- more & more and 2- camano In this the stopword chosen is ‘german’ with the ngram (neighboring sequence of items) ranging from 1 to 3. For this we have used the following packages namely, CountVectorizer, TfidfTransformer, FunctionTransformer and Pipeline.



```
array([[1.        , 0.        , 0.        , ..., 0.        , 0.78669406,
        0.53735414],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.78669406, 0.        , 0.        , ..., 0.        , 1.        ,
        0.94337372],
       [0.53735414, 0.        , 0.        , ..., 0.        , 0.94337372,
        1.        ]])
```

Figure 7. Brand ‘more & more’ cosine Similarity



```
array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 1.        , ..., 0.        , 0.58408849,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 1.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.58408849, ..., 0.        , 1.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

Figure 8. Brand ‘camano’ cosine Similarity

5.3- F1 Score

F1 score for 3 brands have been considered for this project namely- 1- more & more, 2- free people, and 3- Ellesse. From sklearn package MultiLabelBinarizer and f1_score packages have been imported to analyze the dataset.

The 2 dataset (Zalando and About you) was merged with required columns using ‘tmp’ and then after merging using the following packages, F1-Score was determined

Brand	F1-Score
More & more	0.4629
Free people	0.5721
Ellesse	0.4174

Table 1. F1 Score for various brands

From the table, we will come to know that the brand ‘free people’ has the highest F1 score with 0.5721.

6- Conclusion

In this paper, we have compared 2 online fashion websites named ‘Zalando’ and ‘About you’ to find similarities between the products and to match the products using the F1 Score. Few brands were considered from the dataset for the completion of the project and achieved an F1 accuracy of 0.57211 for the ‘free people’ brand

Although the whole dataset was specified by Zalando, we could consider only small clusters from that dataset such as brands. In the future, we will try to consider 2 or 3 brands rather than a single brand at a time for analysis and attempt to achieve a better F1 score than the current one achieved.

7- ACKNOWLEDGMENTS

I would like to thank Dr. Tomas Ward (Dublin City University) for giving me this chance to learn Machine Learning and the algorithms related to the project through this project and David (Zalando) for giving me this opportunity and the exposure to solve a real-world problem using ML algorithm.

8- REFERENCES

[1] P. Hamsagayathri and K. Rajakumari, "Machine learning algorithms to empower Indian women entrepreneur in E-commerce clothing," 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020, pp. 1-5, doi: 10.1109/ICCCI48352.2020.9104111.

[2] J. Albusac, L. M. López-López, J. M. Murillo and J. J. Castro-Schez, "Supporting Customer Searches in E-marketplaces by Means of Fuzzy Logic-Based Machine Learning," 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008, pp. 892-896, doi: 10.1109/WIIAT.2008.167.

[3] H. Dong, L. Xie and Z. Zhang, "Research on statistics-based model for E-commerce user purchase prediction," 2015 10th International Conference on Computer Science & Education (ICCSE), 2015, pp. 553-557, doi: 10.1109/ICCSE.2015.7250308.

[4] https://medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0.

[5] https://towardsdatascience.com/fuzzy-string-match-with-python-on-large-dataset-and-why-you-should-not-use-fuzzywuzzy-4ec9f0defcd.

[6] https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/

[7] https://medium.com/web-mining-is688-spring-2021/cosine-similarity-and-tfidf-c2a7079e13fa

[8] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[9] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[10] https://vitalflux.com/accuracy-precision-recall-f1-score-python-example