

Sentiment analysis and product recommendation on Amazon's beauty products.

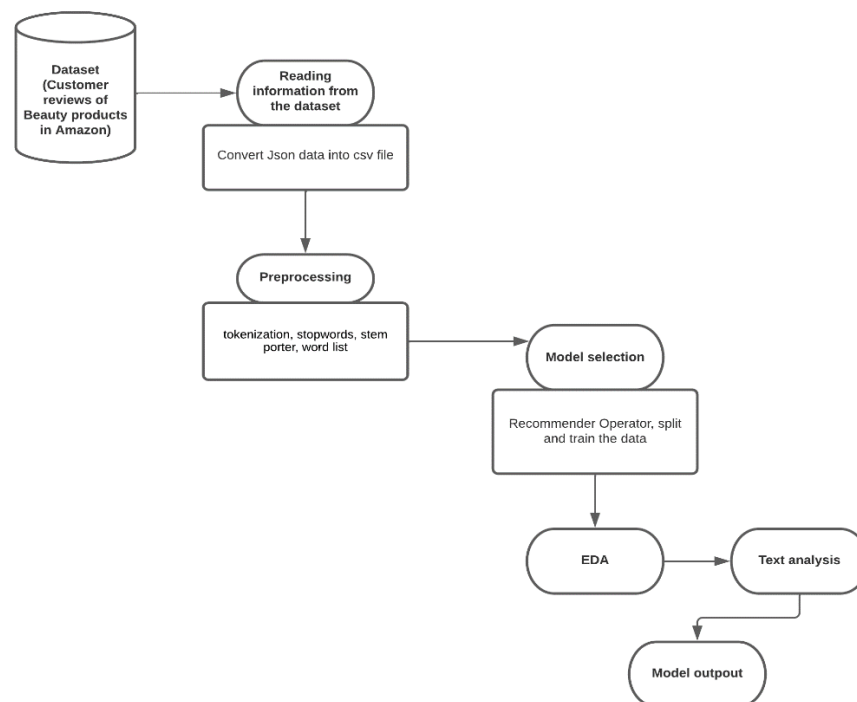
Diana Patricia Carvajal Rozo

1. Introduction

With the rise of digitalization, e-commerce applications have developed an additional advantage for customers, allowing them to purchase desired products based on website reviews. Reviews are undoubtedly a crucial factor for a buyer to consider when purchasing a product. According to a study conducted by Amazon, in 2017, it was shown that more than 88% of online shoppers trust reviews as if they were personal recommendations. As a result, a product with many positive reviews will have higher credibility, while a product with many negative reviews or few reviews can cause skepticism among customers and a reduction in sales. (Amazon, 2018)

Thus, in the decision-making process, consumers want to find relevant reviews as quickly as possible. Therefore, models capable of predicting the user's rating from the text review are crucial and by understanding the overall review, consumers on the one hand could improve their experience and on the other hand businesses could increase sales and if required improve the product by understanding the consumer's needs. (Uma Maheswari Raju, 2020)

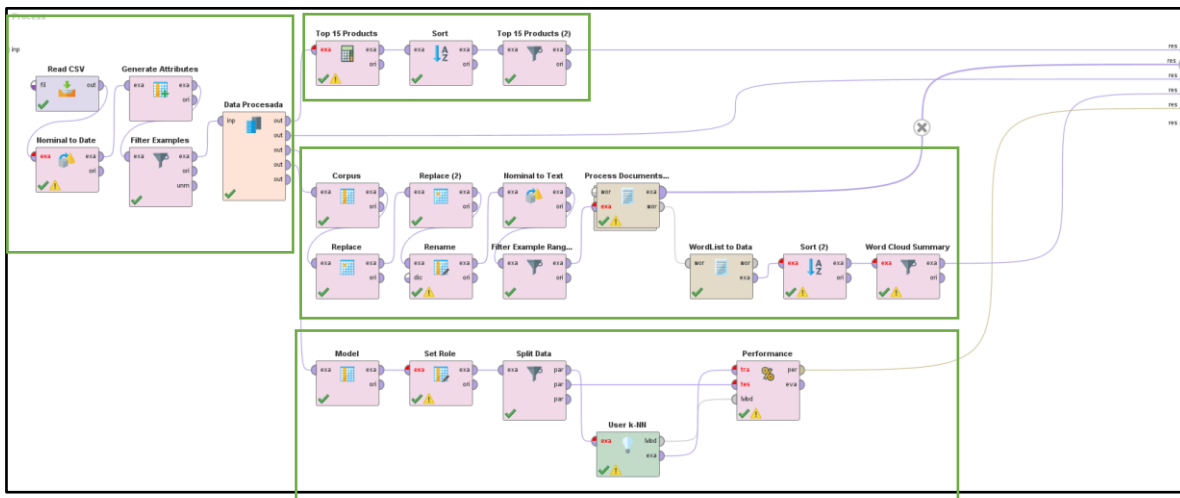
This project's goal is to develop a model to predict user ratings, the usefulness of reviews, and recommend the most similar articles to users based on the recommender and its structure will be followed in this way:



2. Sample, data and corpus

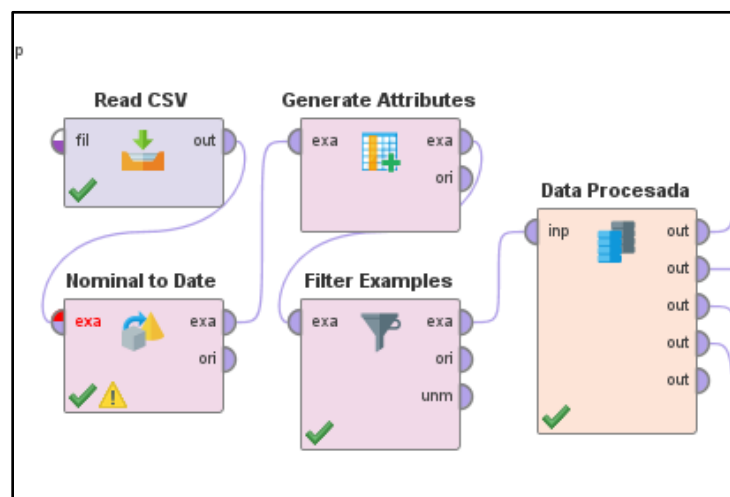
Through the RapidMiner tool, we interacted with a database containing comments made on beauty products sold through Amazon from 2002 to 2014. The beauty dataset consists of reviews (ratings, text, usefulness votes) and product metadata (descriptions, category information, price, brand, and image features).

In general terms, the following diagram represents each of the processes that were carried out for this analysis.



Based on this diagram, it can be seen that four main sections or processes were taken into account for the analysis.

2.1 Preprocessing and information reading section:



This first section takes into account the raw data from Amazon, where it is imported through a csv file which was briefly processed through Python as follows:

```
import pandas as pd
import json
✓ 0.4s

path = "Data/Beauty_5.json"

with open(path,"r") as d:
    components = d.read()
    d.close()
✓ 0.9s

df = pd.read_json(components,lines=True)
✓ 5.3s

df.to_csv("Data/data_amazon.csv")
✓ 5.8s
```

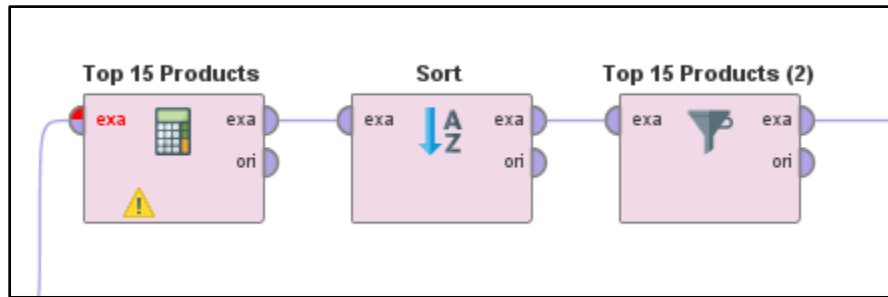
This portion of the code was necessary since RapidMiner directly doesn't allow to import locally json data. For this reason, the data was read with Python, it was processed as a pandas dataframe and finally written as a .csv with the transformed data.

Once transformed, it enters the flow previously presented and is read through Read_csv, as no previous cleaning was performed, this section within the flow seeks to purify the information for subsequent steps.

Among the processing performed, the date was transformed from text type to Date type, additional attributes such as Year, GeneralReview, Lengh_Summary and Lengh_Text were generated to simplify the exploratory analysis.

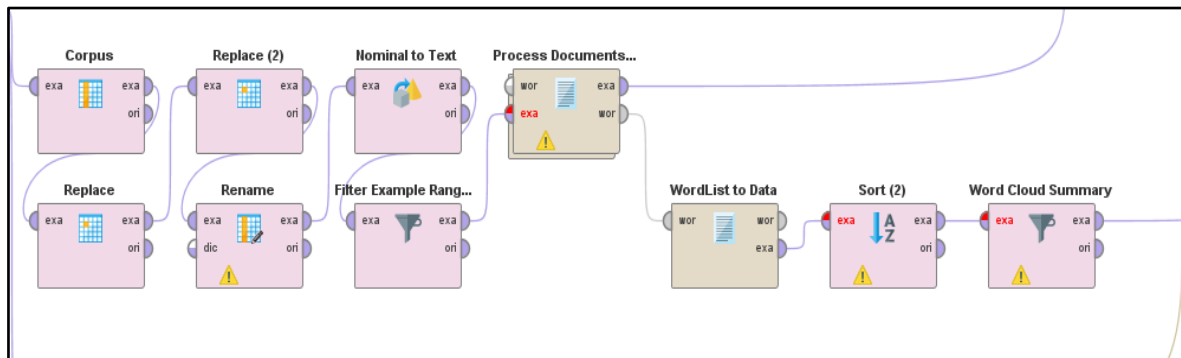
A filter was made on the variable Year where it was null or empty, since we do not have information on this comment, but mainly because it is not a considerably high participation, this same filter was applied in overall, where the comment itself was empty. The Lengh_Summary and Lengh_Text variables are just a count of the number of characters used by the user in both the summary variable and the detailed text of the review. Finally, an Operator is implemented to use the result of the previous processes for multiple purposes.

2.2 Top Products for Exploratory Analysis section:



An additional small section was built to take the previously processed data and generate the top 15 products, as this was required to determine the most commented products or those with the most reviews and also to obtain relevant metrics on the number of products and their historical interaction.

2.3 Text Analysis section:

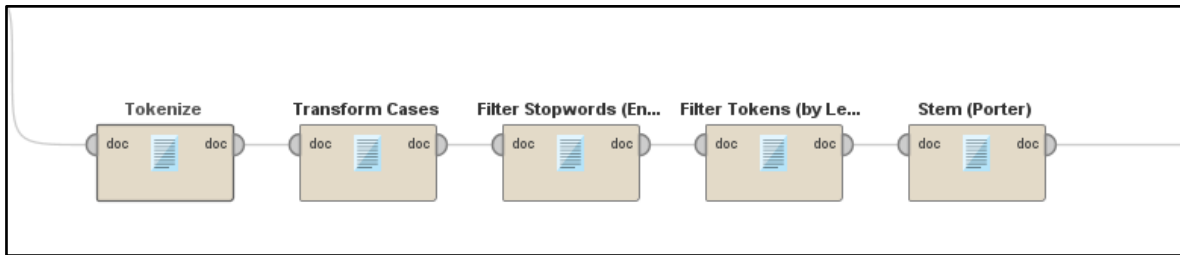


For this section we took into account the two main text variables available (summary and reviewText), so the analysis was performed on these two variables, it is noteworthy that analyzing the relationship between computational difficulty and results it can be said that the reviewText variable has a component of greater difficulty given the length of the text, therefore, tests were performed in order to ensure a conclusion in line with the message that the user wants to give. Although the summary variable may omit certain details, it can drastically simplify the processing time of the analysis. On the other hand, it should be noted that the overall variable has the rating given by the user, which allows, with respect to the summary variable, to have a double validation of the sentiment of the comment.

Therefore, the decision was made to perform the analysis through the summary variable, where this attribute was first selected, filtering out those symbols that do not contribute to the analysis, such as punctuation marks, spaces and special characters.

After this, it was decided to generate a filter of the data due to the fact that in total there are about 200K comments, it was decided to test a filter of 50% of the data to increase the processing speed of the analysis, however, it is guaranteed that the result makes sense with previous tests that took more time and showed that the top of words does not vary drastically with respect to the sample.

Previously this data enters the pure text processing process through a subprocess defined with the following operators.



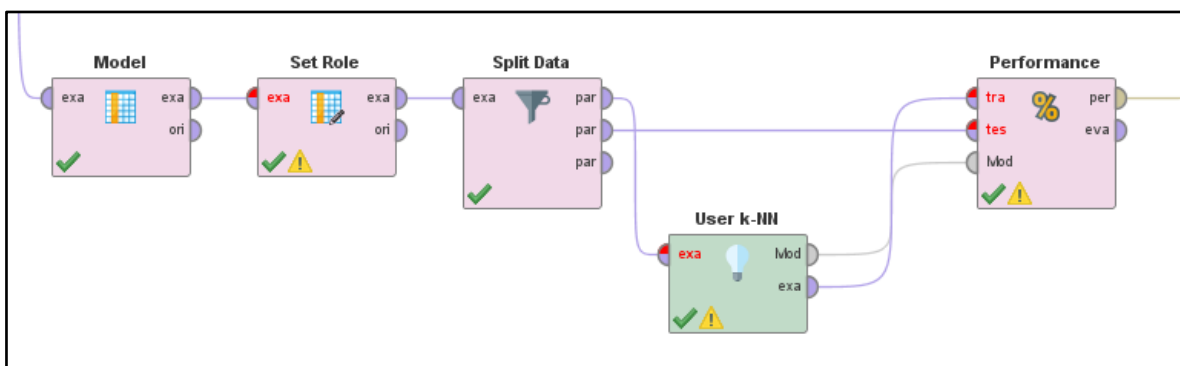
This flow shows how tokenization processes are performed for the input data, which is nothing more than splitting the data in such a way that each word is a row within a new data table, then the transformation of each word from containing words in lowercase and uppercase to only have them in lowercase is performed.

After performing this transformation, the data filter is generated on the StopWords that are those connectors and words that have a high frequency but lack a deep level of knowledge about the subject of the database.

Ultimately, a data filter is created where if the length of each token is less than 4 or more than 15 characters it will be filtered. Finally, by means of Stem (Porter) that seeks to eliminate the most common morphological and inflectional endings of English words. Its main use is as part of a term normalization process that is generally performed when configuring information retrieval systems. This normalization reduces the variability of the corpus and therefore when calculating the frequency of the data, it will be much lower.

Given this subprocess a WordList type object is obtained which by means of the WordList to Data Operator will allow to finally build the word cloud. Before building the word cloud, two additional steps are generated where the results are sorted by the total number of occurrences in descending order and finally the top 20 are taken to build the cloud.

2.4 Model section



This section has the analysis for the model where based on the Recommender Operator the model is taken to create the product recommender with respect to the user's name and the ID

of the product commented. The target of this model is based on the user's overall experience with the product.

A split of the 80/20 data was made (80% for training and 20% for validation) and the model used was a closer neighbor model based on the user's decision which took into account the target previously mentioned, in other words, the user's rating of the product.

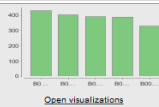
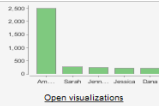
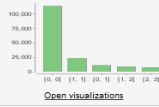
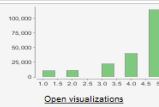
Finally, the Performance operator allows to validate the results with respect to the model and the test dataframe.

In this case a simple architecture was built for the recommendations, however, if want to be much more rigorous in the recommender, it is possible to build an assembly of models where, in addition to models based on the most popular products, a random recommendation is taken into account to give greater variability to the recommendations, if the database of the products is available later it is possible to generate another model where the characteristics of the products are analyzed to make the recommendation. This assembly can also have a weighting on the number of recommendations generated in such a way that each user has a pull of n recommendations that guarantees the variability of products on different characteristics for these.

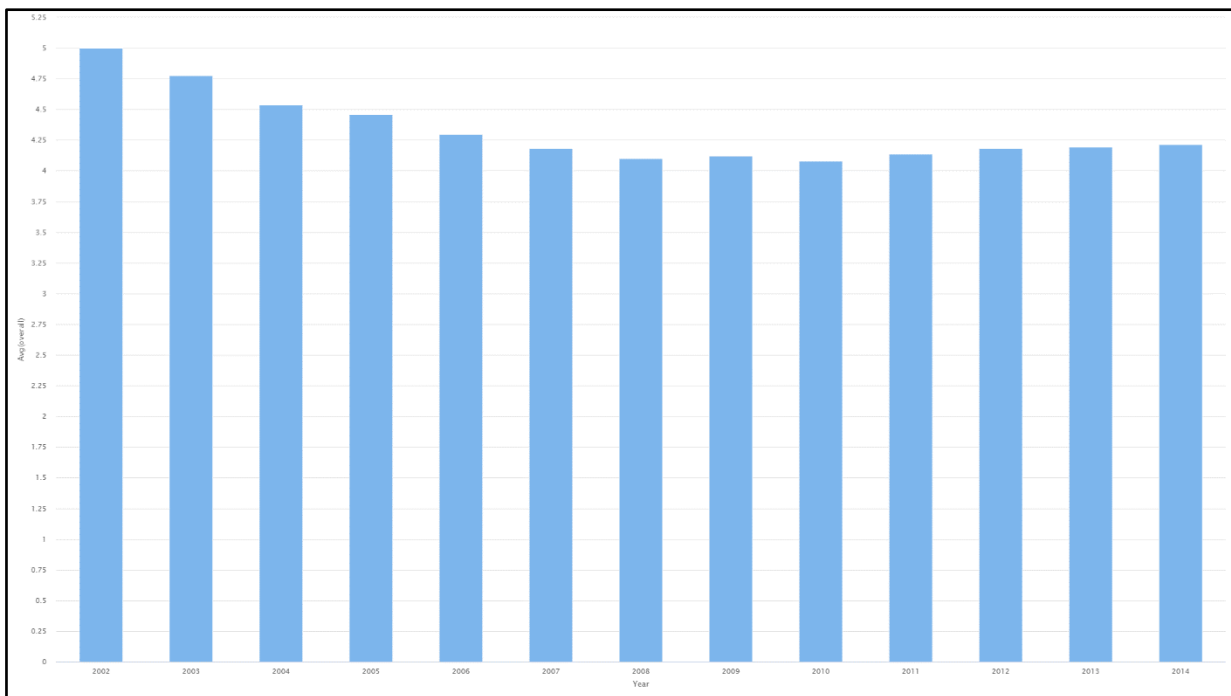
3. Exploratory Analysis (EDA)

Row No.	att1	reviewerID	asin	reviewerNa...	helpful	reviewText	overall	summary	unixReview...	reviewTime	Year	GeneralRevi...	Lengh_Sum...	Lengh_Text
1	0	A1YJEY40YU...	7806397051	Andrea	[3, 4]	Very oily and ...	1	Don't waste y...	1391040000	Jan 30, 2014	2014	Bad	22	154
2	1	A60XNB876K...	7806397051	Jessica H.	[1, 1]	This palette ...	3	OK Palette!	1397779200	Apr 18, 2014	2014	Bad	11	155
3	2	A3G6XNM24...	7806397051	Karen	[0, 1]	The texture of...	4	great quality	1378425600	Sep 6, 2013	2013	Good	13	528
4	3	A1POFF6SAJ...	7806397051	Norah	[2, 2]	I really can't l...	2	Do not work o...	1386460800	Dec 8, 2013	2013	Bad	22	171
5	4	A38FVHZTN...	7806397051	Nova Amor	[0, 0]	It was a little ...	3	It's okay.	1382140800	Oct 19, 2013	2013	Bad	10	334
6	5	A3BTN14HIZ...	7806397051	S. M. Randall ...	[1, 2]	I was very ha...	5	Very nice pal...	1365984000	Apr 15, 2013	2013	Good	18	693
7	6	A1Z59RFKN0...	7806397051	tasha "luvely1...	[1, 3]	PLEASE DO...	1	smh!!!	1376611200	Aug 16, 2013	2013	Bad	6	237
8	7	AWUO9P6PL...	7806397051	TreMagnifique	[0, 1]	Chalky, Not Pl...	2	Chalky, Not P...	1378252800	Sep 4, 2013	2013	Bad	66	225
9	8	A3LMILRM9O...	9759091062	?	[0, 0]	Did nothing f...	2	no Lightening...	1405209600	Jul 13, 2014	2014	Bad	43	271
10	9	A30IP88QK3...	9759091062	Amina Bint Ib...	[0, 0]	I bought this ...	3	Its alright	1388102400	Dec 27, 2013	2013	Bad	11	253
11	10	APBQH4BS4...	9759091062	Charmmy	[0, 0]	I have mixed f...	3	Mixed feelings.	1400544000	May 20, 2014	2014	Bad	15	573
12	11	A3FE8W8UV...	9759091062	Culture C Si...	[0, 0]	Did nothing f...	1	Nothing	1392681600	Feb 18, 2014	2014	Bad	7	108
13	12	A1EVGDOTG...	9759091062	Jessica "Anar...	[0, 1]	I bought this ...	5	This works	1390435200	Jan 23, 2014	2014	Good	10	338
14	13	AP5WTCMP6...	9759091062	Layla B	[0, 0]	This gel did ...	1	Does nothing	1389398400	Jan 11, 2014	2014	Bad	12	121
15	14	A21IM16PQW...	9759091062	mdub9922	[0, 1]	i got this to g...	5	it works	1392681600	Feb 18, 2014	2014	Good	8	207
16	15	A1TLDR1V4...	9759091062	Mickey O Neil...	[0, 0]	I used it for a...	2	burns	1396742400	Apr 6, 2014	2014	Bad	5	113
17	16	A6F8KH0J1A...	9759091062	SanBen	[2, 4]	I order this cr...	5	Did work for ...	1379116800	Sep 14, 2013	2013	Good	15	255
18	17	AXPKZAT7UZ...	9759091062	Shirleyyy	[2, 4]	Good product...	4	excellent	1382054400	Oct 18, 2013	2013	Good	9	203
19	18	A2SIAYDK7G...	9759091062	theredtranny	[0, 1]	I didn't use it ...	3	weird smell	1383284000	Nov 1, 2013	2013	Bad	11	140
20	19	A1QV5IH6HD...	9788072216	armygirl	[24, 24]	I haven't been...	5	Love the sme...	1316390400	Sep 19, 2011	2011	Good	23	485

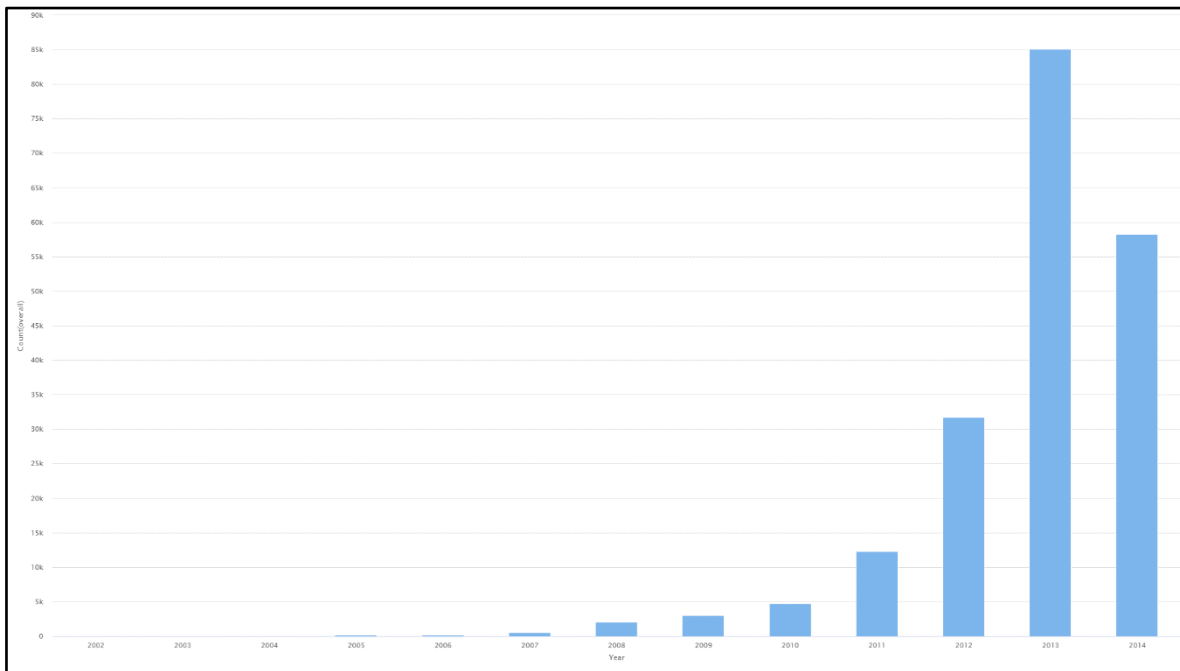
Within the data processing performed in the tool, valuable information was extracted that allows understanding the information in the following way:

Name	Type	Missing	Statistics	Filter (11 / 11 attributes)
asin	Polynomial	0	 <p>Least: B00BSG5YUG (4)</p> <p>Most: B004OHQR1Q (430)</p>	<p>Values: B004OHQR1Q (430), B0043OYFKU (403), B0069FDR96 (391), B000ZMBSPE (389), ...[12097 more]</p> <p>Details...</p>
reviewerName	Polynomial	1384	 <p>Least: the anne (1)</p> <p>Most: Amazon Customer (2496)</p>	<p>Values: Amazon Customer (2496), Sarah (265), Jennifer (242), Jessica (221), ...[19761 more]</p> <p>Details...</p>
helpful	Polynomial	0	 <p>Least: [99, 106] (1)</p> <p>Most: [0, 0] (112740)</p>	<p>Values: [0, 0] (112740), [1, 1] (22696), [0, 1] (10496), [1, 2] (8761), ...[1380 more]</p> <p>Details...</p>
reviewText	Polynomial	27	<p>Least: Most of [...] h it! (0)</p> <p>Most: Love it (4)</p>	<p>Values: Love it (4), great product (4), ...[198136 more]</p>
overall	Integer	0	 <p>Min: 1</p> <p>Max: 5</p> <p>Average: 4.191</p> <p>Deviation: 1.167</p>	

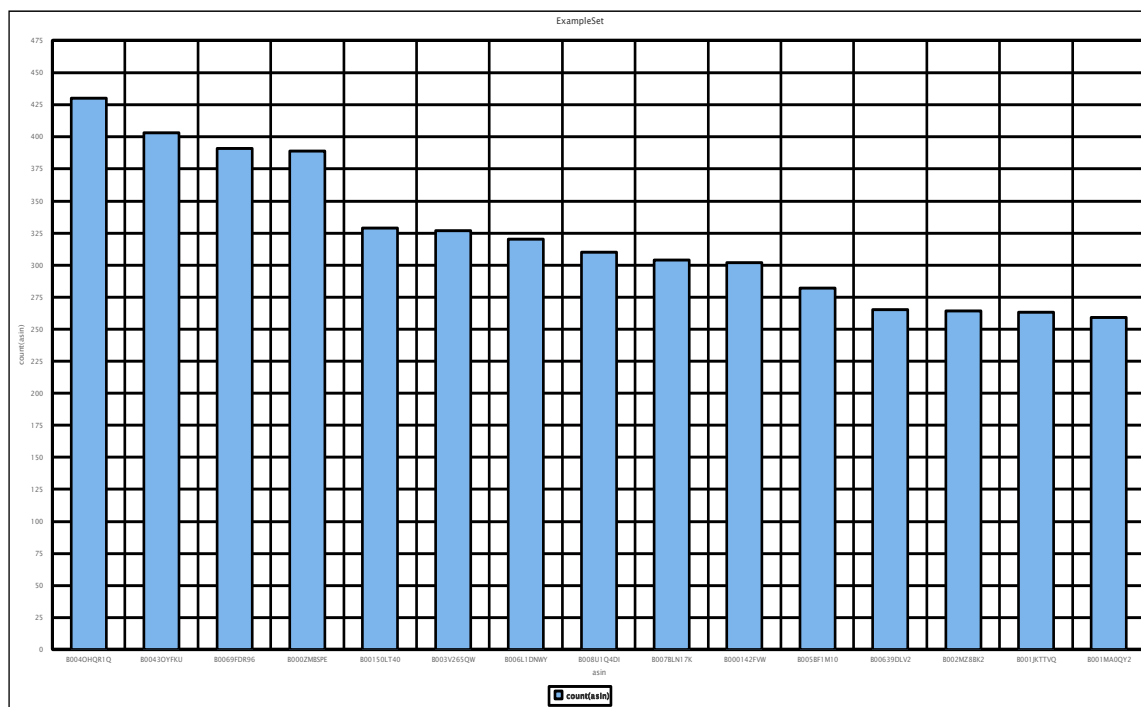
Among the missing values, it was found that for the variable reviewName there are 1384 comments without the name of the person who made the comment, and for the variable reviewText there are 27. On the other hand, one comment was previously filtered out, which had no value overall and therefore, since it was only one observation, it will not be taken into account for the analysis given its null relevance over the total number of comments available. Previously for the other variables, subsets will be generated for different analyses, such as the construction of the model or the sentiment analysis of the comments made.



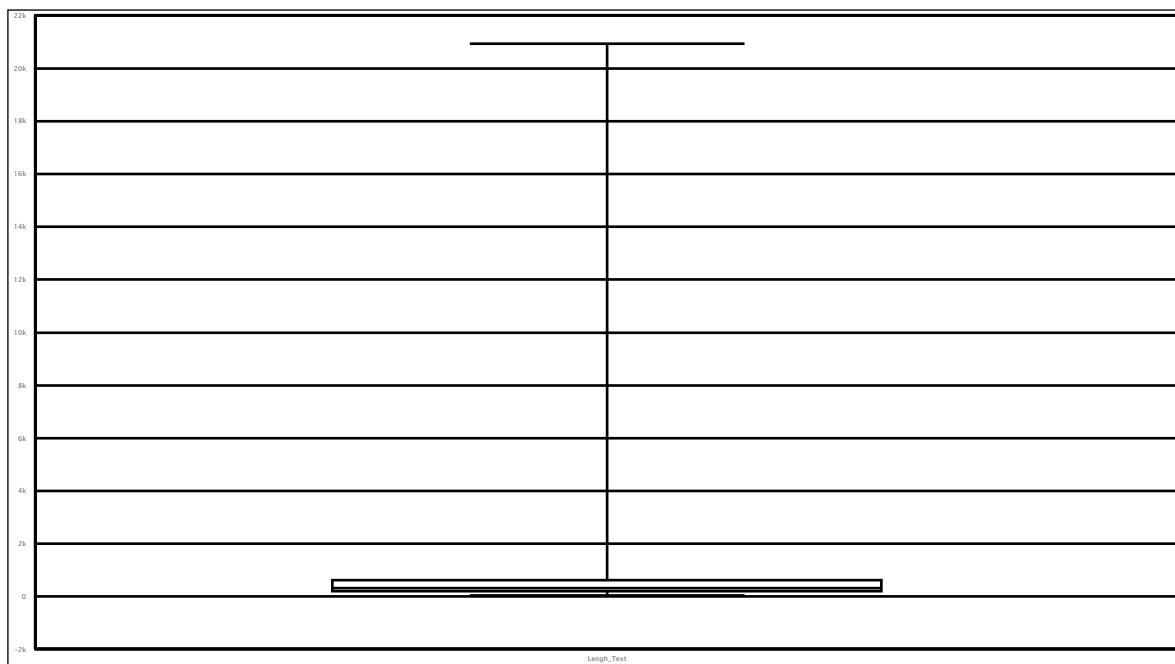
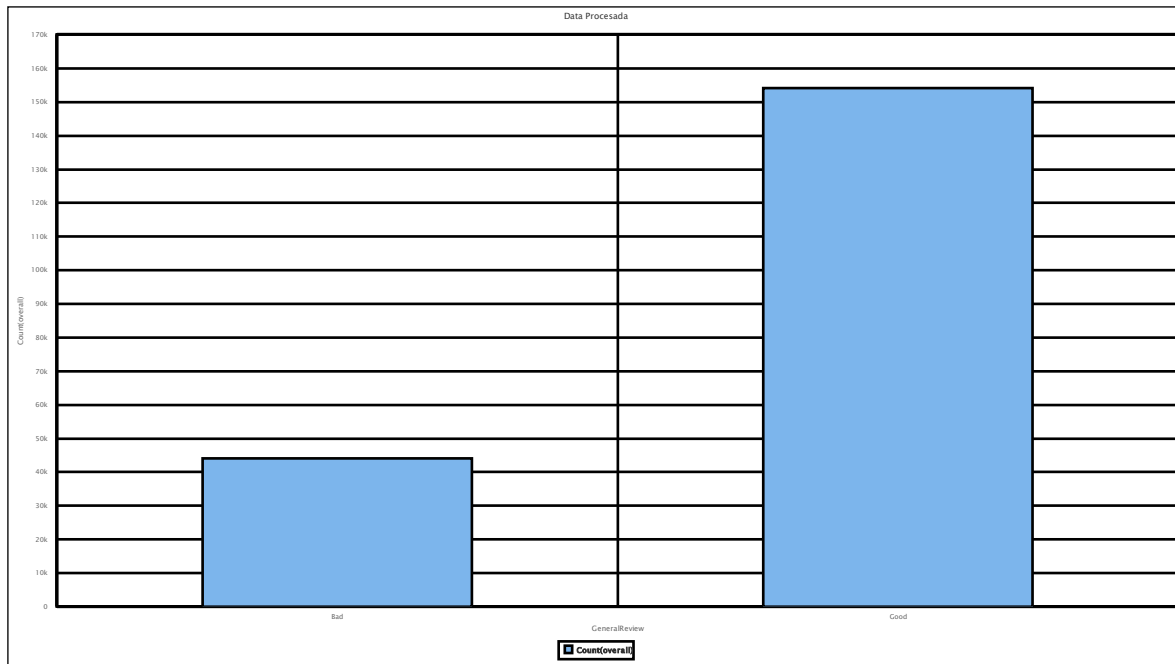
In this graph it can be identified that there is a decrease in the average overall from 2002 to 2014, this is mainly due to the growth of reviews since 2002, going from 4 in 2002 to 58K in 2014, in 2015 there was the highest number of reviews with a total of 85K.



With respect to the products, it can be seen that there is a great variety of products, in the database with the previously explained filters there are a total of approximately 12 products, of which the top 15 products with the highest number of reviews are presented below. On average, taking into account all the products, a total of 16 reviews per product are available.



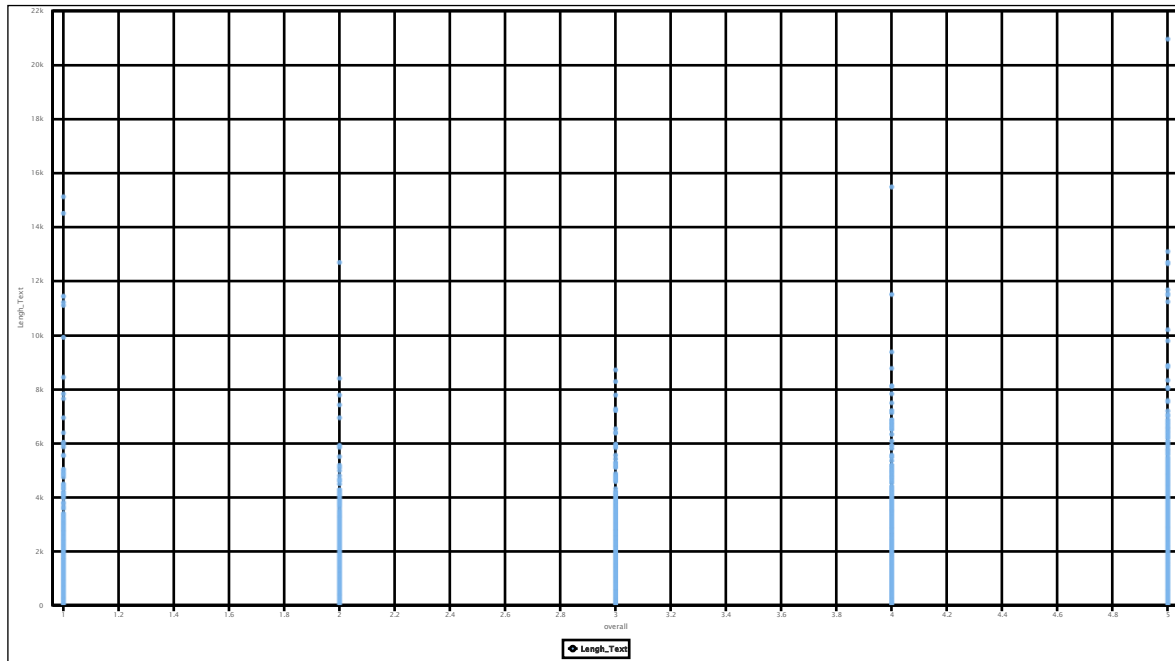
With respect to the overall variable, a new variable was determined that would take into account those ratings higher than 3 to determine a category as "Good" and those lower than this would be categorized as "Bad". This exercise resulted in a very high percentage of comments with scores above or equal to 3 or good ratings. This means that in general the products rated have a good reputation.



With respect to the length of the comments, it can be seen that there are certain comments that have up to 20K characters, which is considerably high considering the 75th percentile

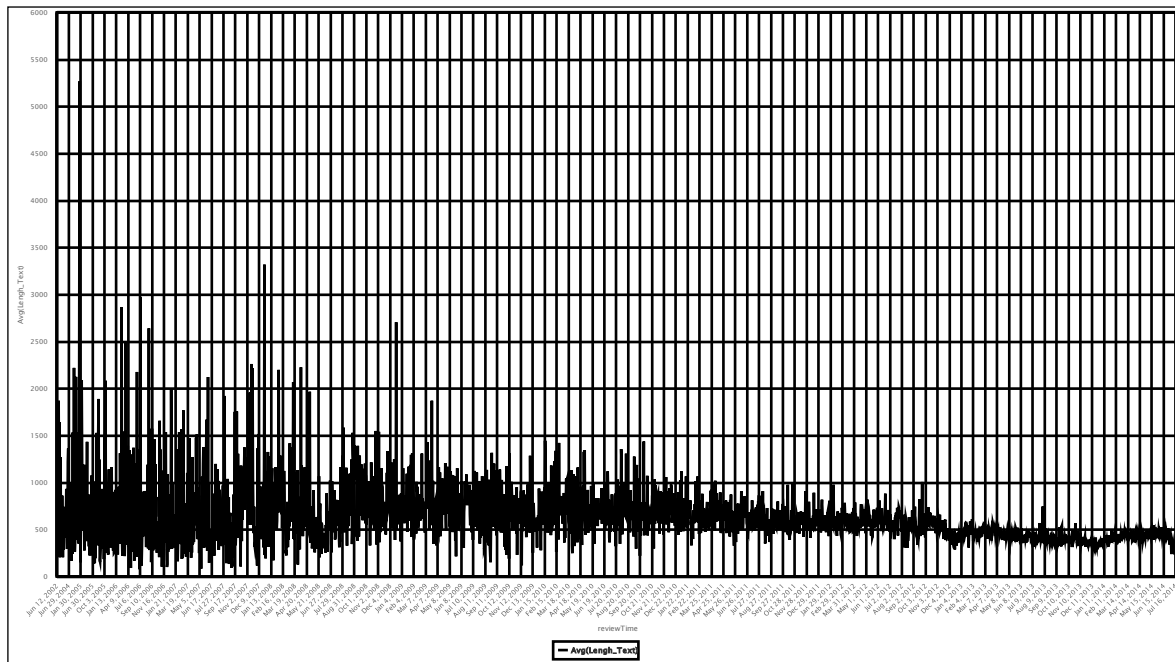
that barely reaches 582 characters, so it is possible that the users who used more characters had a bad experience with the products they received.

To validate the previous hypothesis, the results of overall were also taken into account to corroborate it with the following graph.



With this graph it was possible to identify that there is no relationship as previously stipulated, as can be seen, the highest comment also had a vote of 5. This means that the user decided to give a very detailed review of this product. Comparing the results with respect to other ratings it can be concluded that it is irrelevant to take into account the length of the comment to determine the user's rating, there is an equal variety of comments.

However, we were able to identify those users have changed over time, as they have gone from commenting with many characters to commenting with fewer words each time.



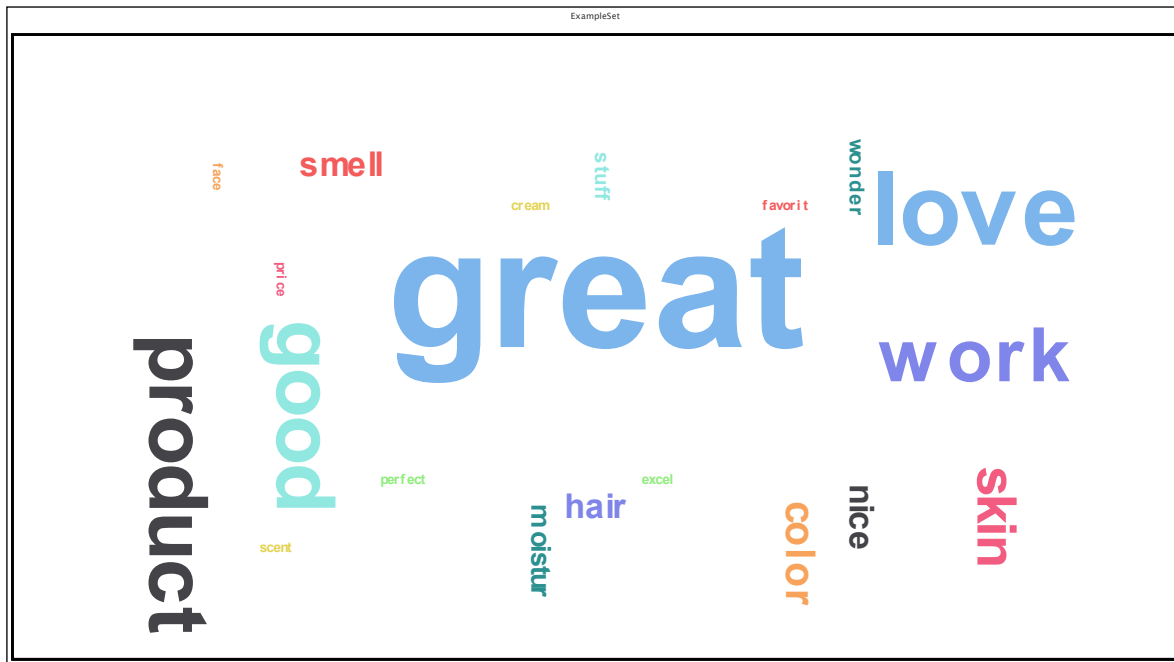
This graph also shows that the user who made the comment with the largest number of words did so in 2004. In fact, it is possible to identify peaks of comments with the largest number of characters practically every 6 months, with the trend mentioned earlier of using fewer and fewer characters in the comments.

4. Text analysis

For the text analysis process, the following steps were identified as necessary within the information processing in order to subsequently apply a sentiment detection technique to generate some kind of idea or prior conclusion about the feelings generated by those products when the customer received or used them.

Firstly, for the two variables available with text (Summary and reviewText) it is possible to generate a word cloud analysis for each one, in principle both variables allow to detect patterns on the type of words and the intention in these. However, their main difference is that the first variable contains a smaller length of words, so computationally it will be much faster to process the information.

Constructing the word cloud, it was determined that the most relevant words agree with the good ratings given, since words such as "great" "good" "love" are mentioned in the top 20 words in many of the comments, which could be related to the good perception of the users towards the product.



5. Model output

As mentioned in the model explanation section, a nearest neighbor algorithm was used to build the model and the results yielded the following performance metrics:

PerformanceVector

```
PerformanceVector:  
AUC: 0.604  
prec@5: 0.029  
prec@10: 0.021  
prec@15: 0.018  
NDCG: 0.179  
MAP: 0.047
```

This indicates that taking into account the AUC, it can be concluded that the recommendation results obtained contemplate a 60% very close to a random level of recommendations, however, for a model that is based on data with a high component of good product recommendations. It is possible that these results are related to the fact that there is low user voting on many products, so that products that are less recognized and consumed by other users have had an important influence on test recommendations. This is why it is important to generate multiple models to try to balance the possible results.

It should be noted that although this recommender may have had a result that could be considered low, it should be taken into account that this type of models should always have a bit of randomness so that products that have low recognition or have little visibility can appear and thus become more popular and have more information about them.

References

- Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, & Antonio Feraco. (2017). *A Practical Guide to Sentiment Analysis*. Springer.
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, & Harshit Surana. (2020). *Practical Natural Language Processing*. O'Reilly Media.
- Uma Maheswari Raju. (22 de May de 2020). *Sentiment Analysis and Product Recommendation on Amazon's Electronics Dataset Reviews - Part 2*. Obtained from Towards Data Science: <https://towardsdatascience.com/sentiment-analysis-and-product-recommendation-on-amazons-electronics-dataset-reviews-part-2-de71649de42b>
- Uma Maheswari Raju. (22 de May de 2020). *Sentiment Analysis and Product Recommendation on Amazon's Electronics Dataset Reviews -Part 1*. Obtained from Towards Data Science: <https://towardsdatascience.com/sentiment-analysis-and-product-recommendation-on-amazons-electronics-dataset-reviews-part-1-6b340de660c2>