Understanding Consumer Trends with images

Andrea Rafanelli[1]

[1]andrea.rafanelli96@gmail.com

1 Business case

1.1 Global level insights

The Dashmote_data_set_case.json contains 10,000 data with several variables, such as: the hairstyle cluster, the number of likes, the number of comments etc.

Through the analysis of this dataset, some information are extracted.

First of all, we discovered that the photos were taken and posted in the first 3 months of 2017: January, February, and March (see Fig.1). Then, there are users who have posted more than one photo, for example there is a user who posted 55 photos (id 4232945658), another who posted 28 (id 40019921) etc. Overall, 15 users posted more than 15 photos. We can therefore affirm that the dataset has 10,000 data but there are only 7,953 users.

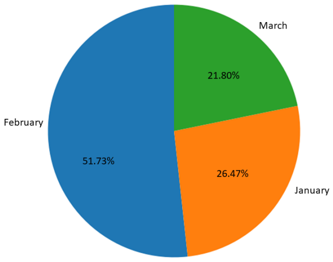

Figure 1: Months of posted photos

In the dataset there are 124 clusters, that is, the visual recognition algorithm has recognized 124 possible hairstyles. Among these 124 hairstyles the 20 most common are represented in the Fig. 2[1].

The most common hairstyle is *foils* with 724 recognitions in the photos, followed by *balayagedandpainted* with 700 recognitions. *mediumhair* is the twentieth cluster with 173 appearances in the dataset. The other hairstyle clusters all have less than 150 photos.

Regarding the likes, only 55 photos have more than 400 likes, 26 more than 450 and only one more than 500.

Among the photos with the most likes, there are 3 photos posted by users who have posted other photos. We refer to user 424624797 (465 likes and 7 comments, 208 likes and 5 comments, and 70 likes and 3 comments), user 3992690338 (465 likes and 24 comments and 170 likes and 2 comments) and user 3063561 (458 likes and 5 comments and 267 likes and 8 comments).

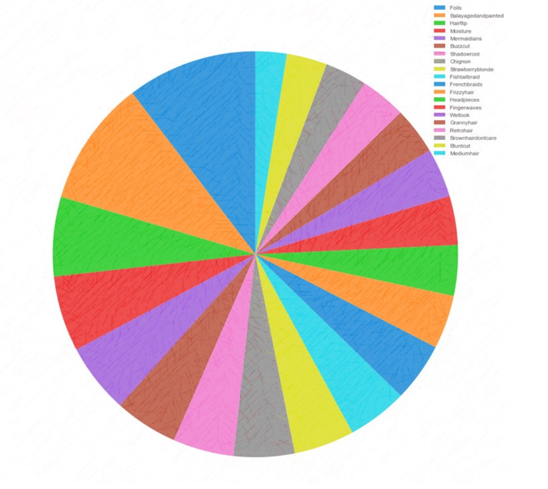

Figure 2: The 20 most common hairstyles

It can be seen that in the three cases described above, users present a photo with many likes and other photos with medium likes. Consequently, we cannot speak of influencers.

The user with the most popular photo is the one with the id 355426846. The photo has 2808 likes and 19 comments. Most likely it is a profile with many followers. This cannot be tested as there is only one photo of this user in the dataset.

1.2 Hairstyle level insights

The 10 photos with the most likes (Fig.3) have different types of hairstyle: the first photo, 2808 likes, is clustered with the **greyhairdontcare** tag, followed by winterhair with 499 likes, **effortlesshair** with 499 likes, **hairponytail** with 496 likes, **babyblonde** with 496 likes, **retrohair** with 495 likes, **colormelting** with 493 likes, **wetlook** with 490 likes, **babyblonde** with 486 likes and **brondehair** with 481 likes.

It can be seen that the **babyblonde** cluster is present twice in the top ten photos with the most likes.

As for the 10 photos with more comments (Fig.4), the clusters are positioned as follows: **babyblonde** with 2311 comments, **fishtailbraid** with 101 comments, **longhairs** with 94 comments, **frizzyhair** with 83 comments, **glitterroots** with 74 comments, **longhairs** with 66 comments, **mermaidians** with 67 comments, **longhairs** with 66 comments, **mermaidians** with 66 comments, **shoulderlengthhair** with 64 comments

[1] The pie chart represents the proportion with respect to 6970 photos considered.

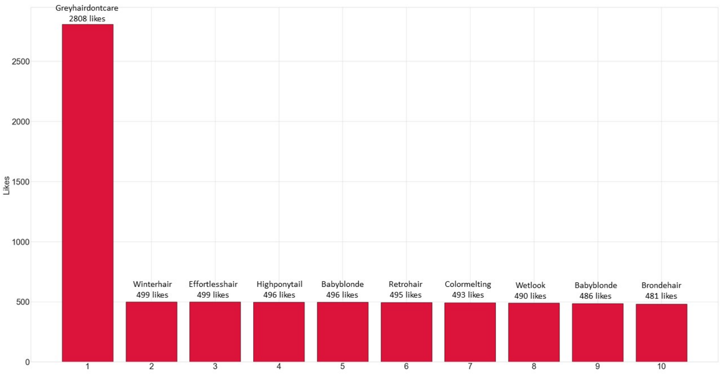

Figure 3: The 10 most liked photos

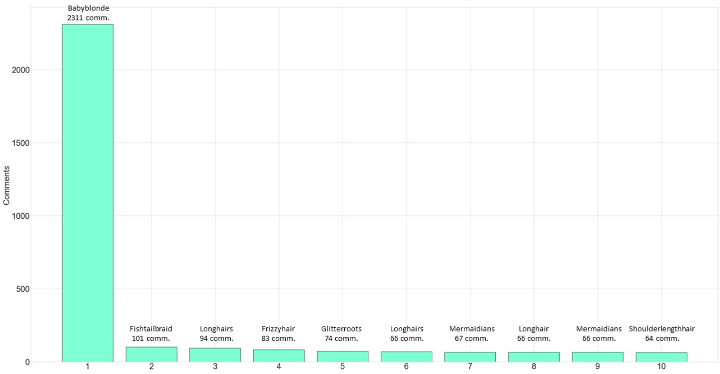

Figure 4: The 10 most commented photos

and **retrohair** with 61.

With reference to the two Figures (3 and 4), it can be seen that **babyblonde** is the cluster that receives the most interactions, positioning itself at the top among the comments and likes.

Despite the popularity of some clusters, it is interesting to ask which are the ten clusters that receive the most likes and comments on average (Figures 5 and 6).

On average hunhair receives more likes, followed by flatwaves, denimbluehair, colortrak, ashbalayage, effortlesshair, mermaidians, greyhairdontcare, **traditionalhaircut** and finally ladywithlocks with an average of 91.1 likes. As for the comments, the **babyblonde** cluster stands out with 39 average comments.

Thanks to the analyzes made in this section, it is possible to state that the hairstyles that receive the most interest from people are those that concern the blonde color, the gray color and some hair coloring techniques such as balayage and "mermaid's hair".

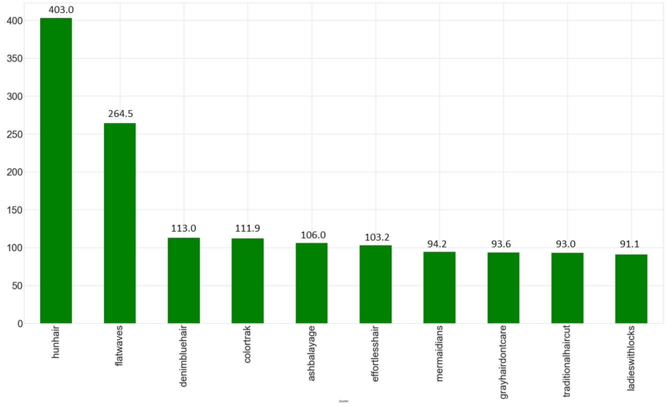

Figure 5: The hairstyles with more AVG likes

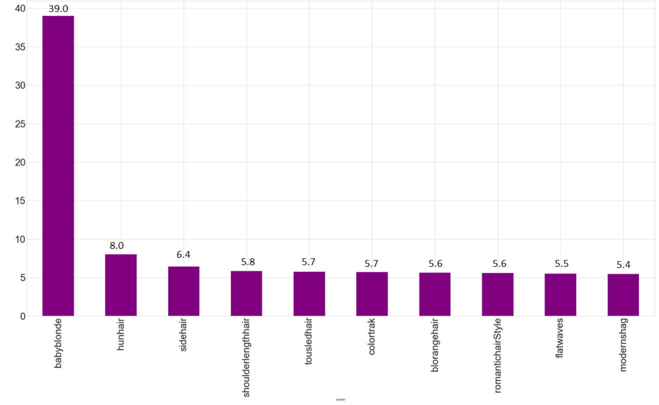

Figure 6: The hairstyles with more AVG comments

Likes	Hashtags
2808	#UniGirls #StarLaceFrontWig #UniWigsTrendyTopSocialStar
499	#instafashion #hairstyles #winterfashion #winterhair #olaplex #loreal #hairstylist #glamhair #healtyhair #hair2017
499	#ramireztran #ramireztransalon
496	#hair #hairstyles #haircrown #beauty #beautiful #everydayhair #photooftheday #highponytail #2017trends #hairstyleoftheday
496	#vipbodyhair #teamVIP #RickyMoica #MagicomMoica #NeiaSoares #AdrianaBill #AdrianaBillVIP #babyblonde #babyblond #BabyLigh

1.3 Hashtags

First of all, the 15 photos with the most likes and those with the most comments are re-analyzed, to understand the combinations of hashtags used.

Then two word clouds are created to better understand the hashtags most used by users.

With reference to the Figure 7 it can be seen that among the most used hashtags we find `#blonde` and `#balayage`; `#haircolor` is also present, confirming the trends found in the previous section. Moreover, some information can be deduced from Table 1.3. For example, Olaplex and L'oréal are two brands of products that are inserted inside the tags. Note that Olaplex is also present among the tags in the Figure 7.

The words hair, hairstyle and hairstylist are used with different tags and these are also present among the most common tags in the Figure 7.

Comments	Hashtags
2311	N/A
101	#champagneatshannons
94	N/A
83	#happybirthday #birthday #bday #curls #CurlCrush #curlyhair #curlsonfleek #curlybeauties #BerryCurly #naturalhair #naturalcurls #myhaircrush #kinkyhair #hairporn #haircrush #haironfleek #afrohair #frizzyhair #TeamNatural
74	#nye #spacebuns #glitterroots #diy

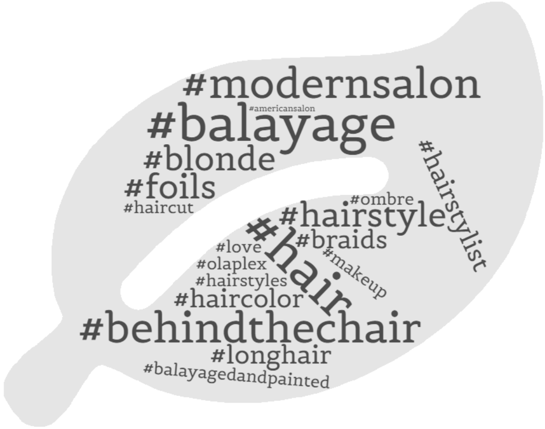

Figure 7: The most common hashtags

1.4 Conclusion

To conclude, the company under consideration needs to focus on products for blonde customers or on hair coloring products. Curly hair is also often mentioned through tags (see Table 1.3) or through clustering: frizzyhairs. Another strategy would therefore be to focus on curly hair products as well.

Surely, there is a need for a greater collection of data, since these data are only concentrated in the first 3 months of 2017 and perhaps on the collection of comments and not just the number of comments, to better understand people's impressions with respect to the look proposed in the photo.

In the analysis (see Section Global level insights) only one photo with more than 500 likes and only one photo with more than 500 comments was noticed.

Another important step for the company would be to create partnerships with high profile influencers or in any case to collect more data from this kind of profiles. This would guarantee the company to better understand the trends and the hottest fashions, as this kind of profiles has many internal interactions that allow to obtain important information.

1.5 Future analysis

An analysis to consider would be that which consists of a more in-depth analysis of the hashtags.

For example, it would be interesting to carry out an analysis similar to the one that is done on tweets, that is to understand if within hashtags there are positive, negative or neutral opinions (we speak of sentiment analysis). Or, it would be stimulating to be able to cluster not only based on images (as has already been done by the algorithm) but also according on hashtags.

Furthermore, given the presence of some brands such as Olaplex and L'orèal but also some salons (for example in the Table 1.3, Ramirez salon is indicated) and some people (again in Table 1.3), it would be important to do a NER (Named Entity Recognition) analysis to understand if famous people, the locations and the brands are indicated in the hashtags. So, for example, famous people could be contacted to advertise the company in question, the locations could be a future partnership for the company, the brands could be future products to be sold by the company etc.

2 Additional questions

2.1 Pipeline

The first thing I would do in case I had to download photos from an online source to build a dataset, would be to go to the site and look at the DOM. This way I could understand the HTML tags that the images refer to (e.g. <img>) and other tags that identify the images and the necessary information.

Then I would start by defining a path for the images and the data and creating a function to download them in XML or JSON files.

Later, I would use a Web Scraping API service to send HTTP requests to the site. Referring to the tags above, I would use those tags to identify all the images on the site, or those of my interest, and I could also use other inputs that allow me to download other information, for example comments, likes, profile url, number of followers etc. To do all this, I must first understand the position within the page of the various elements, the identification tags etc. in order to make precise requests to the API.

As soon as the various requests are implemented, the files are downloaded and saved in the defined path.

Once the dataset has been built, it will be necessary to develop the data cleaning part, for example in our case, dropping duplicates, imputing or dropping null data or if you have the comments apply some text mining technique such as tokenization, lemmatization, filtering etc.

Before inserting the images into the Convolutional Neural Network to create clusters, it is necessary to apply some data transformation technique, for example it may be necessary to normalize some variables or to discretize others, to transform some objects in such a way that the network recognize them in the right way as input etc.

After all these steps, the data is entered into the visual recognition algorithm to be clustered.

2.2 Premium photography identification

The first step of the process would be to take 100k of professional photos and 100k of amateur photos. This photo will be the initial input of the visual recognition algorithm.

To discriminate between the two types of photos I would first consider some photographic aspects that professional photographers tend to respect. For example, these aspects could be the light exposure of the photo, the contrast, and the shadow values within the image. Also, I would consider the degree of focus of the photo. Many of these values could be recognized by the algorithm thanks to a histogram.

An histogram is able to give values that recognize if the photo is overexposed or underexposed, has too closed shadows or too bright lights. These aspects tend to discriminate a professional photographer from an amateur photographer and could be inserted within the algorithm in order to differentiate the two types of photos. In addition, the photo pixels values could be recognized, in order to understand if the photo was taken with a professional camera or not.

Additionally, some changes could be made to the photos, such as various enlargements to understand the levels of blur (normalization of pixels) of the image, changes of color filters (black and white, RGB). Moreover, one could also think of increasing the size of the dataset adding distortions to be able to see if the algorithm can classify distorted photos as non-professional.

For the validation of the algorithm, we take another 100k photos in which 30k are professionals and the others are of amateur photographers and we try to understand if the algorithm is able to discriminate well between the two types of photos. If this is not successful, the network parameters could be modified: change activation layer, add dense layers, change optimization algorithms etc.