Trend Analysis in Harper's Bazaar

Sophie Kushkuley, B.S.
Brandeis University
415 South St
Waltham, MA 02453, USA
sophiek@brandeis.edu

Abstract

Topic modeling of fashion trends were analyzed using the MALLET toolkit. Harper's Bazaar magazines from 1860-1899 were used (freely available online). This resulted in 20 topics with 4 characterizing words each. Trends over time were analyzed in several different ways using 100-topics and 20-topics.

1 Introduction

Using trend analysis to extract topics from a fashion magazine may finally put to rest the question of the cyclicality of fashion. Entire issues of Harper's Bazaar from the 19^{th} century are freely available online [1]. Preliminary data analysis using the NLTK toolkit in Python [2] did not yield promising results. Bigram collocations were extracted for the first three years of data. The collocations were extracted on a monthly basis, however the bigrams were extremely non-specific and contained no relevant information. Topic modeling is the natural choice for large amounts of historical data, so this was the strategy implemented for the second attempt. Extracting this data and applying the MALLET toolkit [3] for topic modeling provided a good start and a novel way of looking at fashion trends.

2 Data Extraction

For every year from 1867 through 1900 roughly 52 volumes (one volume per week) per year of Harper's Bazaar are available in text format online. Most volumes from 1867 - 1899 contain an article entitled 'New York Fashions'. For consistency, this is the article that was scraped from every volume for the purposes of topic modeling. Volumes from year 1890 only go up to April 28th and do not contain the 'New York Fashions' article, hence 1890 was excluded from the analysis. 1867 was a short year as well, volumes started in November. A Python script was used to extract the articles from every volume, however, the data is not entirely uniform and contains errors. Despite significant post-processing, noise in the data has caused some imprecision.

3 Topic Modeling

MALLET uses latent Dirichlet allocation [5] to produce a topic distribution over any given text. Stop words were removed automatically, and a distribution of a user specified number of topics was produced together with a user specified number of topic keys associated with every topic.

3.1 100-Topic Model

First, 100 topics were distributed over all the data, each with 4 topic keys. Many of the resulting topics were uninformative (e.g. 'cost made black ladies' and 'good made make great'). Below is a sample of the topic keys for the first 20 topics:

black satin green jet; trimmed satin skirt dress; skirt made skirts hips; silk black long jet; brown gray blue dark; white tulle dress low; worn ladies young made; de black white soie; flannel worn warm skirts; dress costumes white blue; blue pink white pale; satin velvet lace brocade; price shown centre set; plain figures shown designs; yard cents sold cost ;

Workshop on Computational Linguistics for Literature, pages 84–87,
Montréal, Canada, June 8, 2012.

capes back long jackets; cambric tucks linen french; girls years white children; costumes costume style trimming; collar high pointed front.

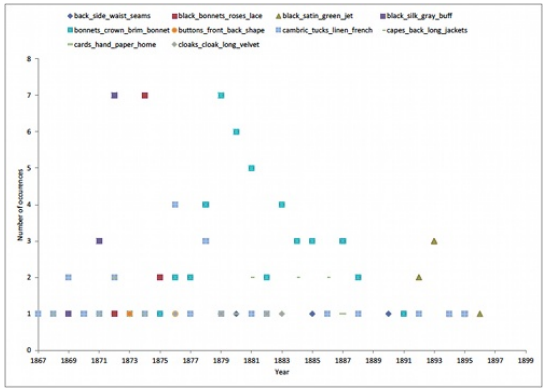

Figure 1: Occurrences of topic by year: first 10

At first, the topic with the highest percentage for each volume was chosen, along with its topic keys. Then every year for which this is the number one topic was found. Some topics will be associated with the same year a number of times, because each year contains roughly 52 volumes each of which has a topic distribution. The number of times each year shows up in each number one topic was counted and plotted (Figure 1). This allows one to see which topics are strongly correlated with a particular year or set of years, and from here it may be possible to deduce which topics are most representative of trends by year.

Some keys (e.g. 'cost made black ladies') are very general, uninformative and infrequent. Others occur several times in one year and not in any other year (e.g. 'costumes costume style trimming'). Then there are topics that occur frequently over a short range of years (e.g. 'dresses skirt plain wool'). And still other, such as 'fur seal skin black' that occur infrequently over a vast range of years, vs those that occur infrequently over a short range of years, 'flannel worn warm skirts'.

The most relevant topics are the ones that occur frequently over a short range of time. These are topics that may be representative of trends, and it would be helpful to plot the distribution of such topics over time.

3.2 20-Topic Model

Due to the large amount of data generated by the 100-topic model, a 20-topic model was generated to provide a more thorough analysis of the data. Twenty topics are more easy to manually look at, and 20 is not too few topics but it's also not an overwhelmingly large number of topics. For each of the 20 topics 4 topic keys were generated. Below is the list of all 20 topic keys:

fur seal black long; made gown skirt gowns; hair ladies made worn; yard price cents suits; velvet cloth red made; made girls waist dresses; crape black mourning worn; white dress dresses silk; white lace blue made; satin lace black jet; skirt front back side; silver gold diamonds large; black blue color brown; designs wood cost small; worn suits white black; stripes colors blue designs; text px hearth td; bonnets crown hats brim; silk long made back; back waist sleeves skirt.

It is evident that one topic ('text px hearth td') is due to bad post processing, so it has been thrown out from further analyses. Topics in the 20-topic model have much higher counts than those in the 100-topic model. This might lead to less fine-grained topics for the data. Some topics clearly stand out as having a very high frequency for a short range of years (e.g. 'made gown skirt gowns').

Next, the distribution of a sample of topics was plotted by year. This was done by following each topic separately over time; the percent of the topic in the distribution for each year was plotted by year. Since every year has many topic distributions (corresponding to each of the 52 volumes per year for most years) a bootstrapping sampling method was performed to determine the percentage associated with a specific topic and year. Bootstrapping provides the advantage of a weighted average that correctly preserves the original distribution. A percent from the topic percent-pool for the year was chosen randomly 1,000 times and averaged to produce one single percent associated with that topic and year.

The most frequently occurring topics are 'Silk long made back'; 'Skirt front back side'; 'Velvet cloth red made'; 'White lace blue made'; 'Yard price cents suits'; 'Made gown skirt gowns'; 'Back waist sleeves skirt'; 'Black blue color brown').

Four of these topic percentages are plotted by year

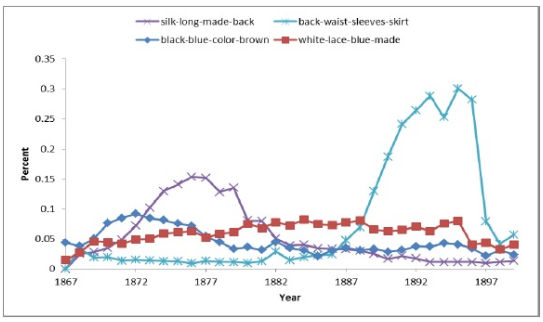

Figure 2: Topics by year

(Figure 2). This yields a view of the topics in which specific trends can be spotted based on frequency (e.g. long and silk items peaked around 1872-1882, velvet and red characteristics peaked around 1882-1887). In addition, a certain amount of cyclicality can be seen in Figure 2. Two of the topics have a topic key in common, 'back'. The first topic 'silk long made back' peaks around 1877, and 'back waist sleeves skirt' peaks around 1895, and these two topics have a topic key in common, 'back'. Given a context, meaning can be found in such trending topics.

3.3 Project Refinement

It appears that the topic keys generated by the 100-topic model are more relevant and contain more information than those generated by the 20-topic model. On the other hand the 20-topic model is simpler to analyze. Therefore, inspired by the work on Martha Ballard's diary [4] another model was generated with 20 topic keys for each of 20 topics.

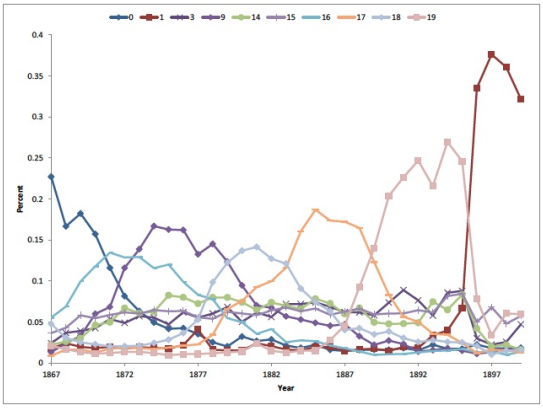

Figure 3: Trend-predicting topics by year

These labels were plotted by year, and topics with the most movement are shown (Figure 3). There are trends here that clearly stand out, but there does not seem to be enough data to follow the trends through to the end. Certain topics rise in popularity and fall again, however, it is impossible to know from 30 years of data if they rise again further in time. Similarly to the 20-topic model with 4 topic keys, many topics stand out due to their cyclic nature. However unlike the 4-topic keys model, rises and dips in frequency are more dramatic, suggesting that more topic keys leads to a more thorough analysis of trends.

An additional analysis was performed using not only the most highly probable topic associated with every year but the top three such topics. The number of occurrences of the top three topics of every month were counted and divided by the total number of months to obtain a percent of occurrence for the year. A heat map was then generated for this data (Figure 4). Some specific trends are discernible, and even a potentially cyclic topic (topic 7 is frequent in 1867 and resurfaced in 1896).

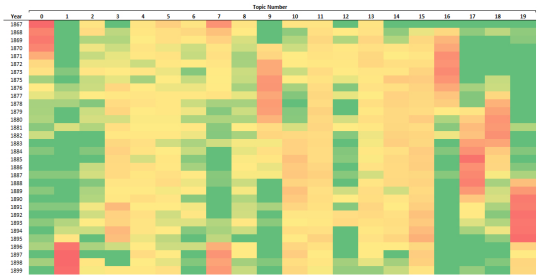

Figure 4: Heat map of topic occurrence by year

The high frequency topics are presented below. Types of garments, accessories, colors, materials and fabric stylings have been highlited:

0 **black** yard **silk** wide **dress** price worn half gros sold **fringe** folds **trimming** grain cents inches yards trimmed centre amp

1 made **gown** gowns **lace black silk** worn **white** style wear put **satin** trimmed waist fashion effect great year smart women

7 **skirt** front **belt** skirts waist cut narrow side back long material short **ruffles** bands band **flounce trimmed** trimming full upper

9 **silk black skirt dresses basque** back long front

side **pleating** trimmed pleated **flounces** pleatings dress **apron** polonaise plain trimming sleeves

16 **suits** worn made ladies cost **gray** suit blue stylish amp imported fashion shown **linen brown** garment soft yard cents **braid**

17 **skirt dresses** made waist front full side plain **red** dress **drapery** foot **wool** skirts line **gathered** surah **pleats basque vest**

18 **satin red lace gold** plush colors beads **velvet** large made small **brocaded** plain **trimmings embroidery** imported great shown shades dark

19 **black** sleeves waist **gowns skirt silk** back **ribbon satin** full front **white** collar large **green bodice belt yellow** foot gown

These topics include many descriptive terms (in bold) that could depict broad scale fashion trends. Highlighted topic keys describe real trends and provide a starting point for further analysis. Not surprisingly, frequent topics coincide with the most highly peaking and oscillating topics in Figure 3. These frequent topics also depict more fashion trends, as defined by the categories mentioned above. For example, below is a non-frequent topic, and it does not include a single topic key that falls into the categories defined above:

2 hair head price natural cents large fancy pretty long ladies hand lady box made cards dolls water children good dressed.

4 Conclusion

Topic modeling for 19^{th} century magazine data does not automatically yield relevant topics that can be plotted and analyzed. Extensive post-processing and generalization is necessary for useful results. It is important to correctly classify topic keys and identify useless topics.

It seems to be the case that more topics and more topic keys yield better results; which may then be obtained by carefully sorting through the 100-topic model and categorizing topic keys into topic labels, then plotting them by year to analyze trends. It may be possible with more detailed analyses to deduce topic keys that are cyclic in nature when put in context (e.g. 'back' in Figure 2). A heat map can be a good way to weed out uninteresting topics. It also provides an excellent visualization method for the rise and fall of topics, as well as topic cyclicality. Based on these highlighted topics, it is interesting to group the topic keys by characteristic (e.g., color, article of clothing, construction, technique, material etc).

Trends have been discerned in this analysis, and with this wealth of freely available data specific fashion trends can be searched for and analyzed.

5 Future Work

The fashion trend analysis of the 19^{th} century according to Harper's Bazaar presented here is incomplete. Further refinement of the topics will yield the identification of more specific trends that can then be analyzed over time.

Honing in on specific topic categories (e.g. articles of clothing, materials, colors and styles) can help illuminate trends. Topic analysis can then be performed for every category for a cross-sectional view of trends.

Post-processing of the keys is another necessary step to focus in on trends. For example, stemming the topic keys is necessary to avoid repetitive keys.

Additional data will also be helpful to fully analyze trends and assess cyclicality. Shared topic keys among the topics may provide insight into the problem, if a method of linking them contextually can be deduced.

This paper provides an initial insight into fashion trends using topic modeling with MALLET, but it also leaves room for further directed analyses.

References

1. Albert R. Mann Library. 2012. *Home Economics Archive: Research, Tradition and History (HEARTH).* Ithaca, NY: Albert R. Mann Library, Cornell University. http://hearth.library.cornell.edu (Version January 2005).
2. Bird, Steven. *NLTK: Natural Language Toolkit.* http://www.nltk.org/.
3. McCallum, Andrew Kachites. 2002. *MALLET: A Machine Learning for Language Toolkit.* http://mallet.cs.umass.edu.
4. Cameron Blevins. 2010. *historying. Topic Modeling Martha Ballard's Diary*, http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.
5. Blei, David M. Ng, Andrew Y. Jordan, Michael I. 2003. *Latent Dirichlet Allocation.* Journal of Machine Learning Research 3 (2003) 993-1022.