ב"ה

Advertisements 1840-1957 in North America

What did they sell and how?

Team info: Amitai Gispan - amitai.gispan@mail.huji.ac.il, Eitan Rosenfelder – eitan.rosenfelder@mail.huji.ac.il

Teacher: Professor Dafna Shahaf

Course: Data mining, 47717, Hebrew University of Jerusalem, Israel

אור לכ"ד אב ה'תש"פ

ב"ה

Table of continents

ב"ה

Problem description:

We wanted to understand how ads changed over time, what was advertised and how in different periods, including during wartimes.

Data

We built two data sets based upon two major collections of ads. We studied each data set on its own, and we merged the two sets for some of our research. Both databases are built from digital collections of advertisements which we collected by crawling from the Duke University site. One collection is comprised of 3,186 advertisements and publications dating from 1840 to 1920, illustrating the rise of consumer culture and the birth of a professionalized advertising industry in the United States. The second collection has 7,219 advertisements covering five product categories - Beauty and Hygiene, Radio, Television, Transportation, and World War II propaganda - dated between 1911 and 1957.

For each advertisement, we collected the following meta data about each ad, as provided in the digital collection:

Title given by the collector; year of publication; first words of the ad (when provided by the digital collector); publishing company; product being marketed; prior collector (donor to Duke University); subject of ad; location of ad and medium used (when provided), and the link to the digitised ad.

Solution

In this project we used Python to achieve the following:

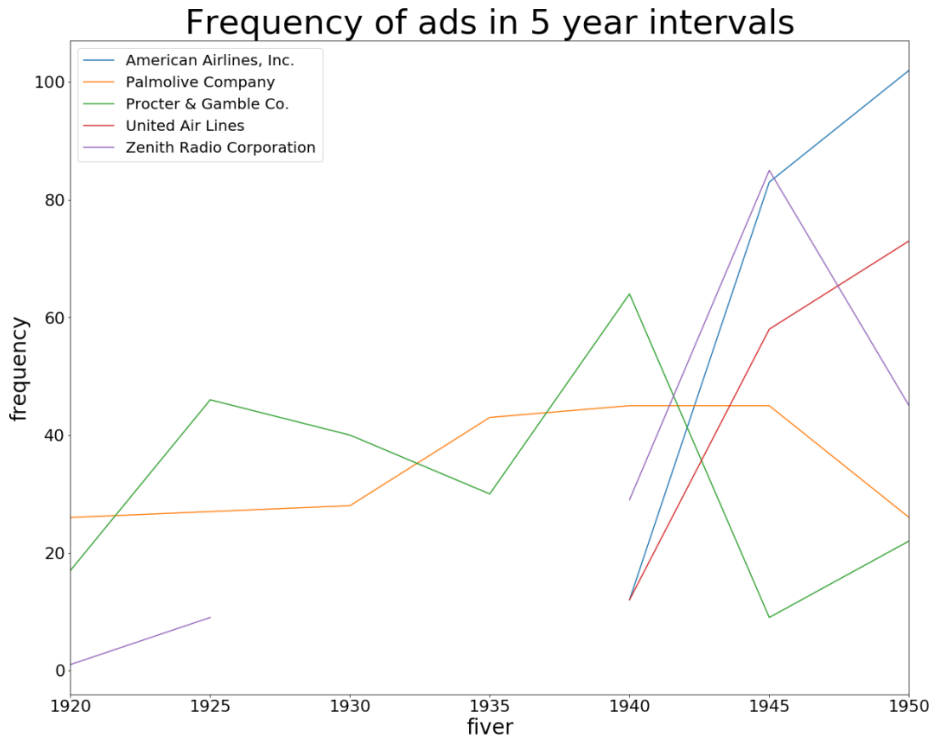
1. Data Collection, using crawling methods from two online collections in the Duke University.
2. Descriptive analysis, looking at the data and showing various interesting items, using four methods:
   a. Word clouds, to show the titles and subjects over years.
   b. Bar plots, to show the frequency of several topics in the data.
   c. Line graphs, to show the frequency of advertising over time, of the leading companies in certain areas.
   d. K-means classifier, to cluster data about advertisements related to children and to tobacco. We rated the accuracy of the classifier macro.
3. Predictive analysis with K-means, including performing a confusion matrix.

Evaluation:

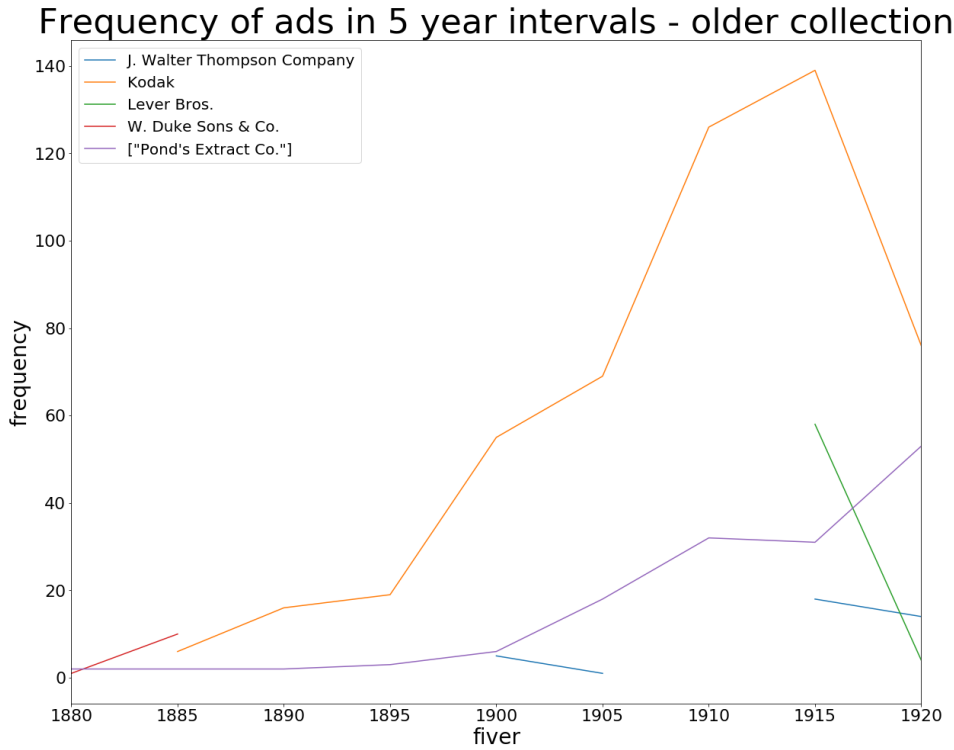
Descriptive analysis

- Top 5 companies in each data set:
  The data we collected came from two different collections, therefore we checked the 5 companies with the most advertising in each data set. Merging the data sets would cause the mistaken observation that some companies, such as Kodak, stopped advertising in 1920. The reason for this mistaken impression is that the second collection of ads is limited to a few specific topics, therefore Kodak ads are not in that collection. By studying each collection separately, the following picture appears:

  1. Frequency of advertising of 5 main companies, in 5-year intervals (1911-1957)
     American Airlines and United Airlines both started advertising during WWII, with the number of ads increasing after the war. Zenith Radio advertised in the early 1920s, stopped in the early twenties and then returned to advertising during WWII, with far fewer ads after the war. The other major two companies are Palmolive and Procter & Gamble, both of whom ran a moderate number of ads throughout this period.
     This graph demonstrates that during wartime, Zenith invested in a lot of advertising. However, after the war ended, they cut back on advertisements. United and American Airlines began advertising in the early 40's, and have continued to advertise throughout the time frame. At the same time, advertising of products used in day to day life is more stable, whether in wartime, economic depression, or peacetime prosperity.

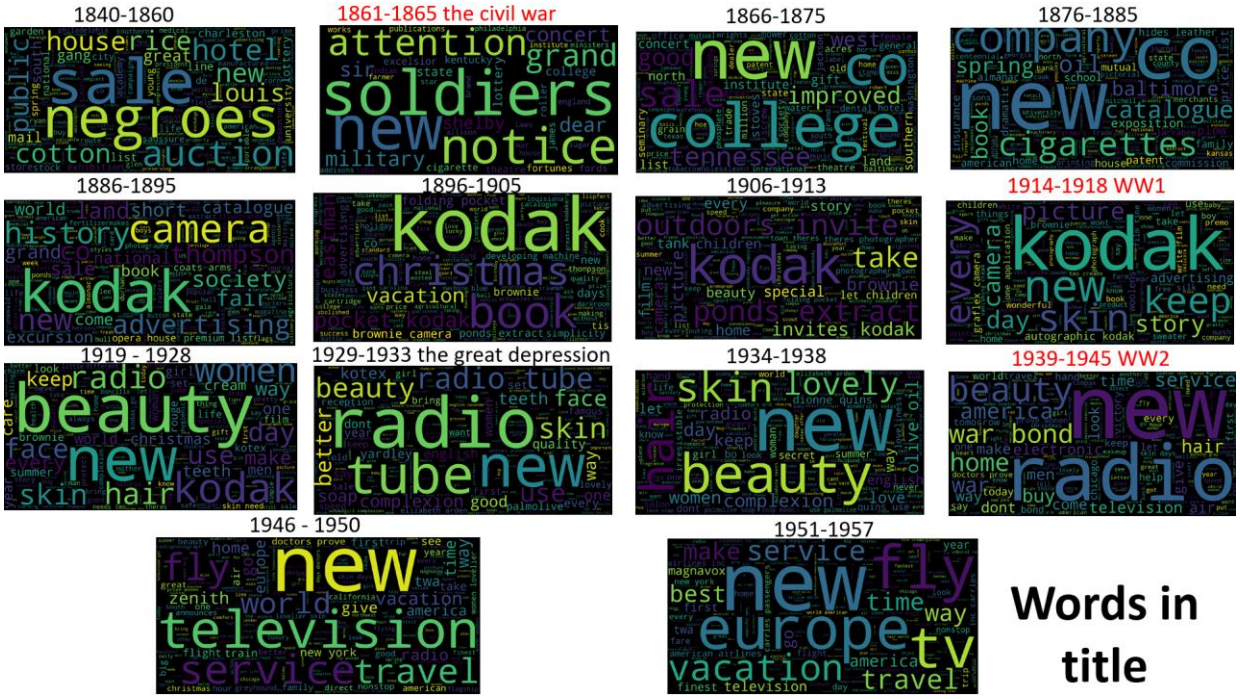


2. Frequency of advertising of 5 main companies, in 5-year intervals (1840-1920)
   Of the five main companies in this collection, two market hygiene products. The first, Ponds Extract, ran very few ads in the years 1880-1900, and increased their advertising significantly since then. The other, Lever Bros., was founded in 1917, and therefore only began advertising towards the end of World War I. The other 3 leading companies in this collection are: J. Walter Thompson, an advertising company which published ads in two periods, 1900-1905 and again in 1915-1920; the W. Duke Sons and Co. tobacco company, which advertised for a short time in the early 1880's; and the company which advertised the most in this time period, Kodak camera and accessories company (founded in 1888). Kodak had a big increase in advertisements throughout this period, indicating company growth. The change in advertising frequency, as shown in this collection, paints a picture of public interest in products used in day to day life, alongside the ads for the advertising company, which was presumably in order to attract more companies.



- Changes over the year as reflected in the word clouds:
  Looking at two main aspects of ads over the years - the title given for each ad (a few words from the ad, usually from the top of the ad, selected by the Duke University librarian) and the ad's subject (as described by the Duke librarian). For each of these topics, we plotted a word cloud for specific time periods. In the case of data from before 1860, the 93 ads were grouped together, as there were too few to split them. During periods of significant change, i.e. during the Civil War, WWI, the Great Depression, WWII, the ads are grouped by those years. Additionally, ads from the years 1946-1950 and 1951-1957 are in their own groups. The other word clouds describe ten-year intervals. War periods are labeled in red.

  Word Clouds of ad title words:

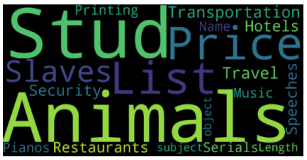


  - Ads run before 1860 are very different than all the other types, with some ads that seem to deal with sales of people as slaves, something which was stopped after the Civil war.
  - During the Civil War and WWII, ads relate to the military and the war. This is as opposed to WWI, when war vocabulary does not appear as frequently in the titles. It appears that since the Civil War and WWII were fought on American soil, Americans were more concerned about what was happening and it was more noticeable in day to day life. On the other hand, North American soldiers fought in WWI, but by leaving the continent to other parts of the world, and therefore the ads preferred not to make a deal of it.
  - Kodak mentioned their name in the title of big percentage of their ads ever since they existed. However, Kodak ads do not appear in the second Duke U collection, and therefore the word clouds from the second collection do not contain their name. In the first collection, Kodak advertised a lot, and this is reflected in the word clouds.
  - In the 19th century, colleges and cigarettes were a major topic in the text of many ads.
  - The word "women" appeared frequently in ads from the early 20s, which may be related to the fact that women got the right to vote in the United States in the year 1920.
  - "New" is a term that appears in almost every decade. This phenomenon will be discussed in depth later.
  - Looking at the years past WWII, in the divisions of 1946-1950 and 1951-1957, one sees similarities in that ads for flights, TV and vacation are very common. However, three main differences between the two time periods are apparent:
    1. "Europe" becomes a major term, which indicates that flights to Europe became more common in the early fifties.

2. "Vacation" is much more noticeable, implying that going on vacation became more common.
3. The third point is amusing: In 1946-1950, the term used is "television," while in 1951-1957 the term used is "TV." This points to common use of slang words, such as TV for television.

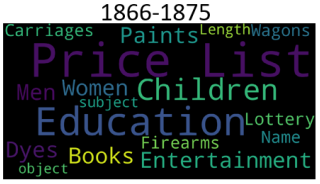
Word clouds of ad subjects:

1840-1860



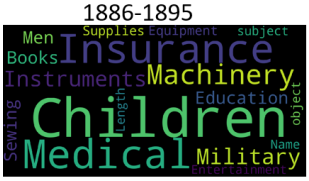
1861-1865 the civil war

1866-1875

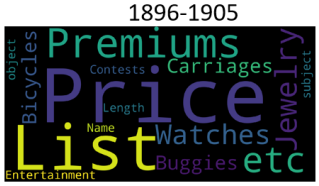


1876-1885

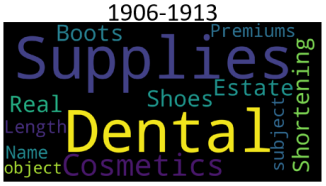
1886-1895



1896-1905

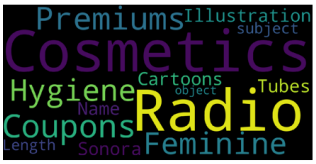


1906-1913

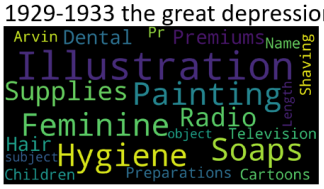


1914-1918 WW1

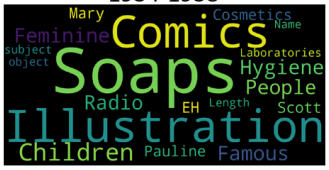
1919 - 1928



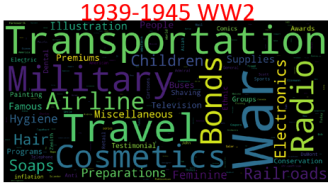
1929-1933 the great depression



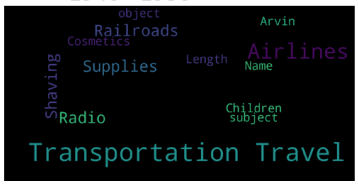
1934-1938



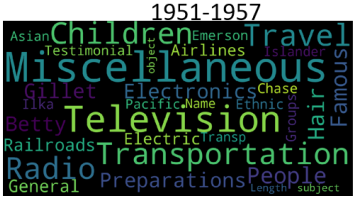
1939-1945 WW2



1946 - 1950



1951-1957

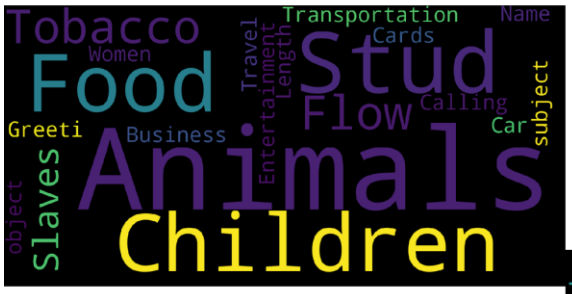


subjects

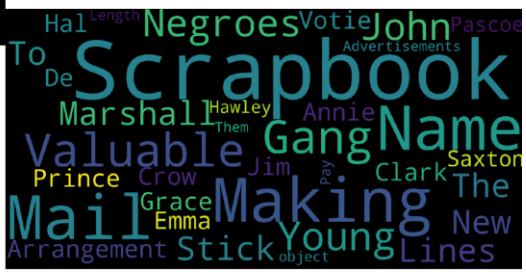
- Ads dealing with military or war appear only in 4 periods: the Civil War, 1886-1895, WWI and WWII. The 1886-1895 period was not familiar, but it matches the time of the Hawaiian rebellions from 1887 till 1895, which took place in Hawaii.
- As mentioned in the titles, in the first time period (1840-1860), ads for slaves were common. From the subject, it appears that slaves, studs and animals are the main topics which were advertised.
- From the ads published in the 19th century, it is clear that price lists were common in those ads. This reflects a difference in how items were advertised: while ads used to publicize price lists in the 19th century, these seem to have disappeared in the 20th century.
- The topic of transportation became popular from WWII and onwards, which may mean that ads for transportation became very common from that period and later, or that the collection of such ads only started then. These possibilities should be considered regarding ads from airlines starting in 1940 or 1943.

❖ Ads from the 19th century:

The subjects in the data from 1840-1920



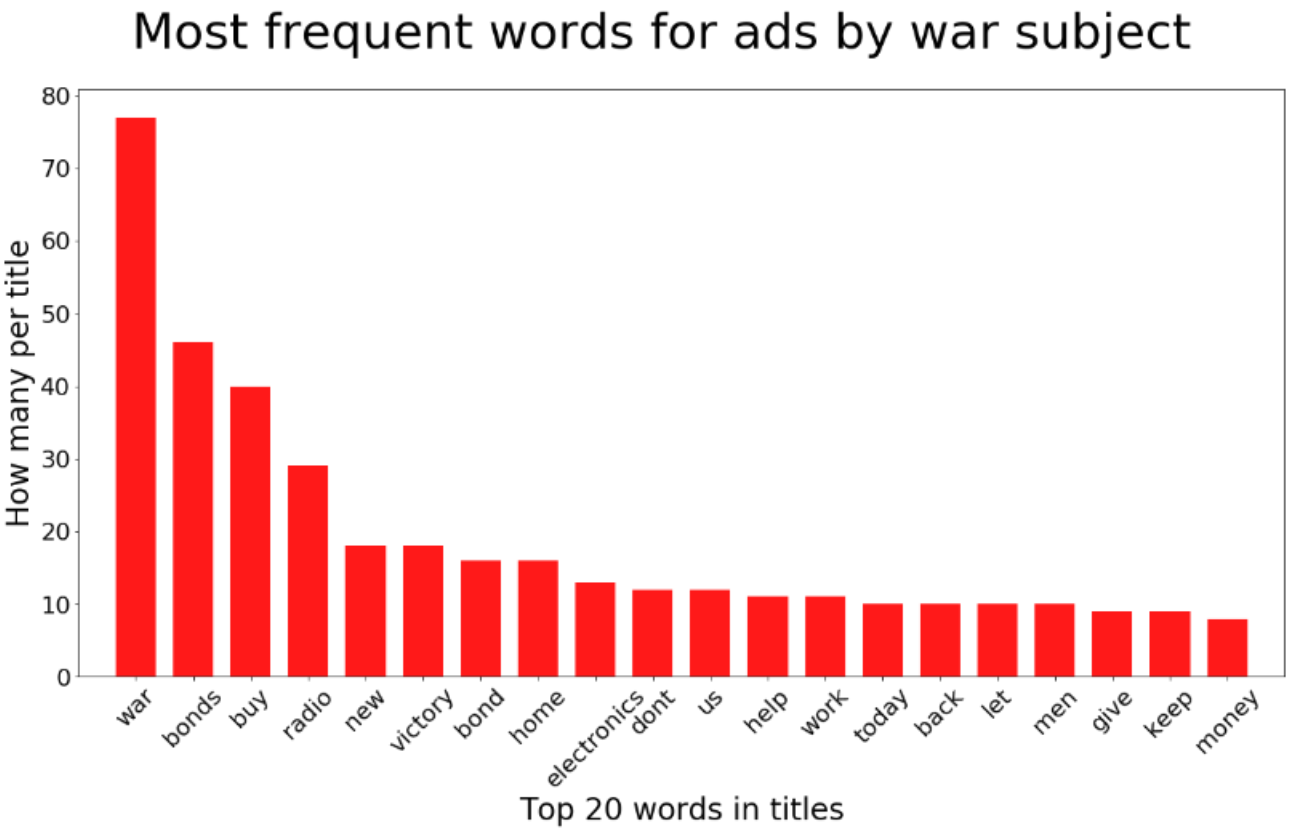
The titles in the from 1840-1920

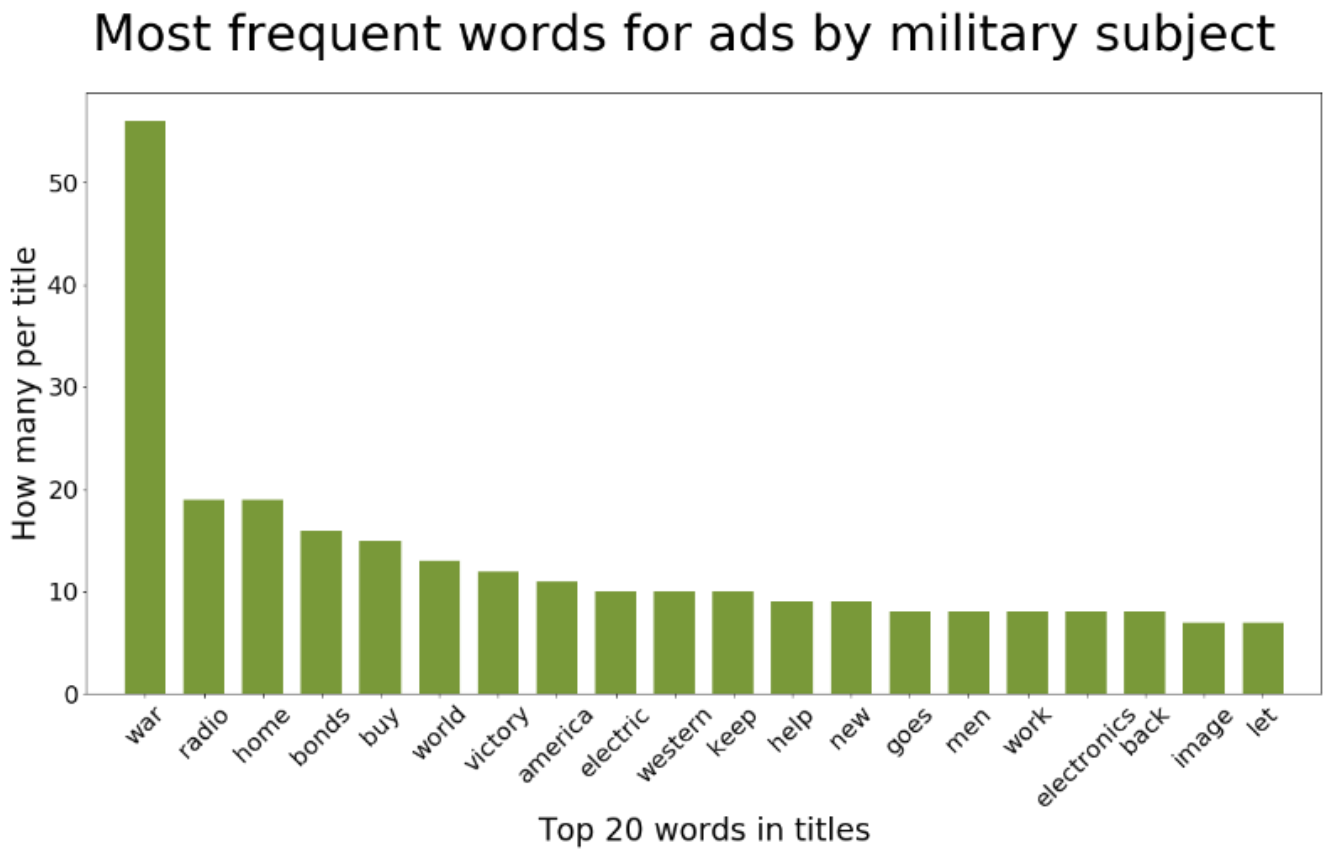


Two interesting items are seen from the data:

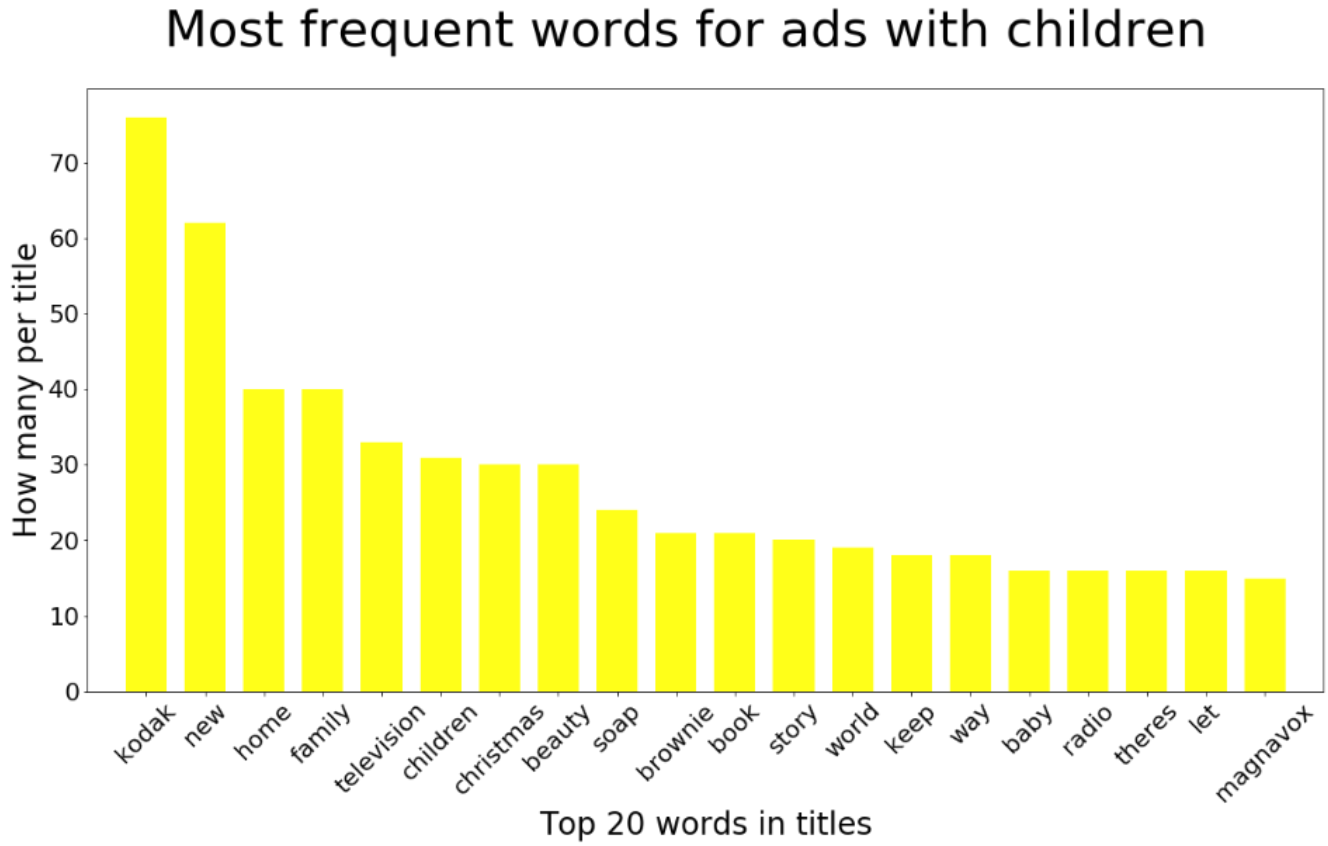
- The title very often includes names of people.
- The subjects animals and animal-related products, food, and children.

❖ War and military: The most frequent words regarding war and military topics are basically the same, but there is a slight difference in how frequent each of the top words are. For example, "bonds" are more common in war topics than in military topics.

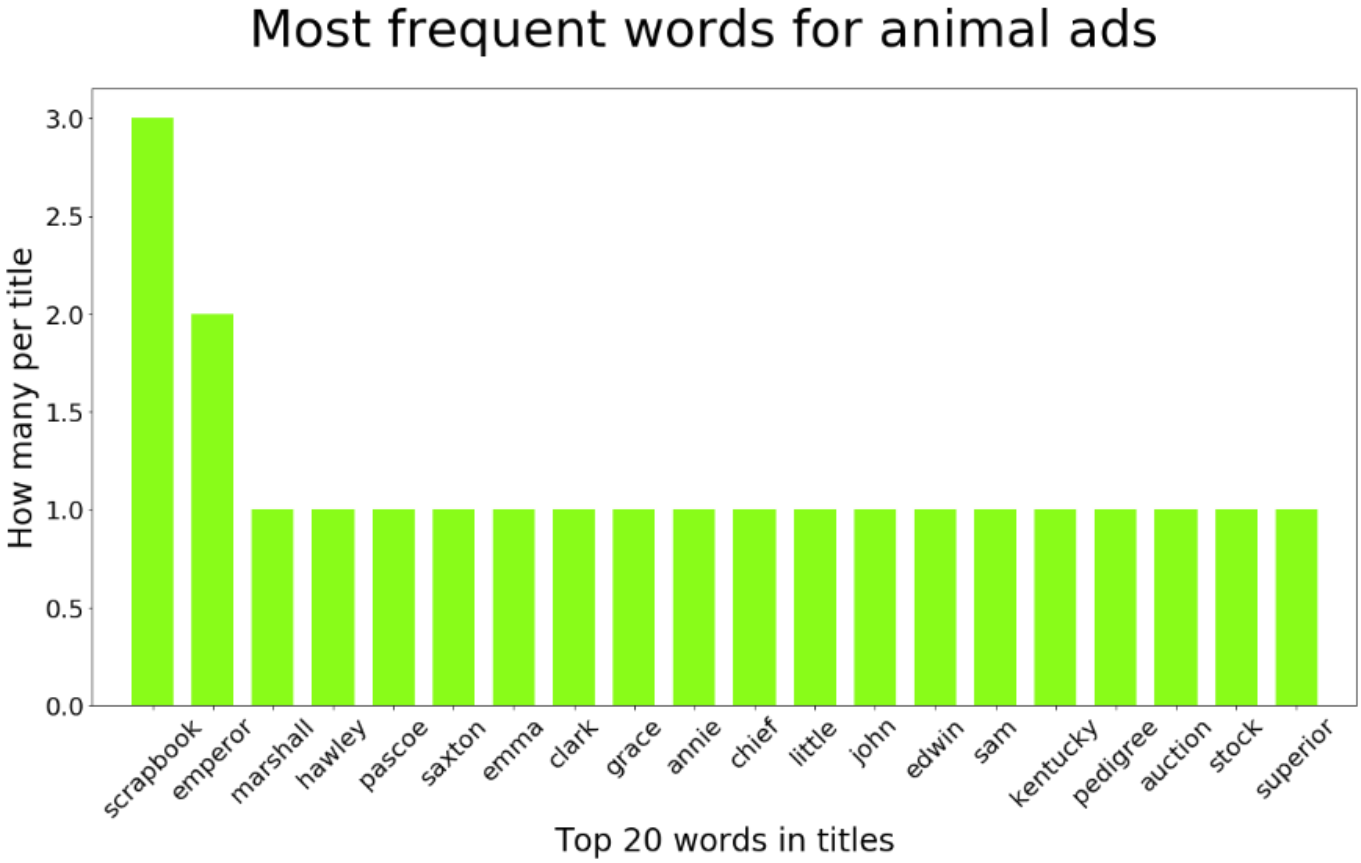




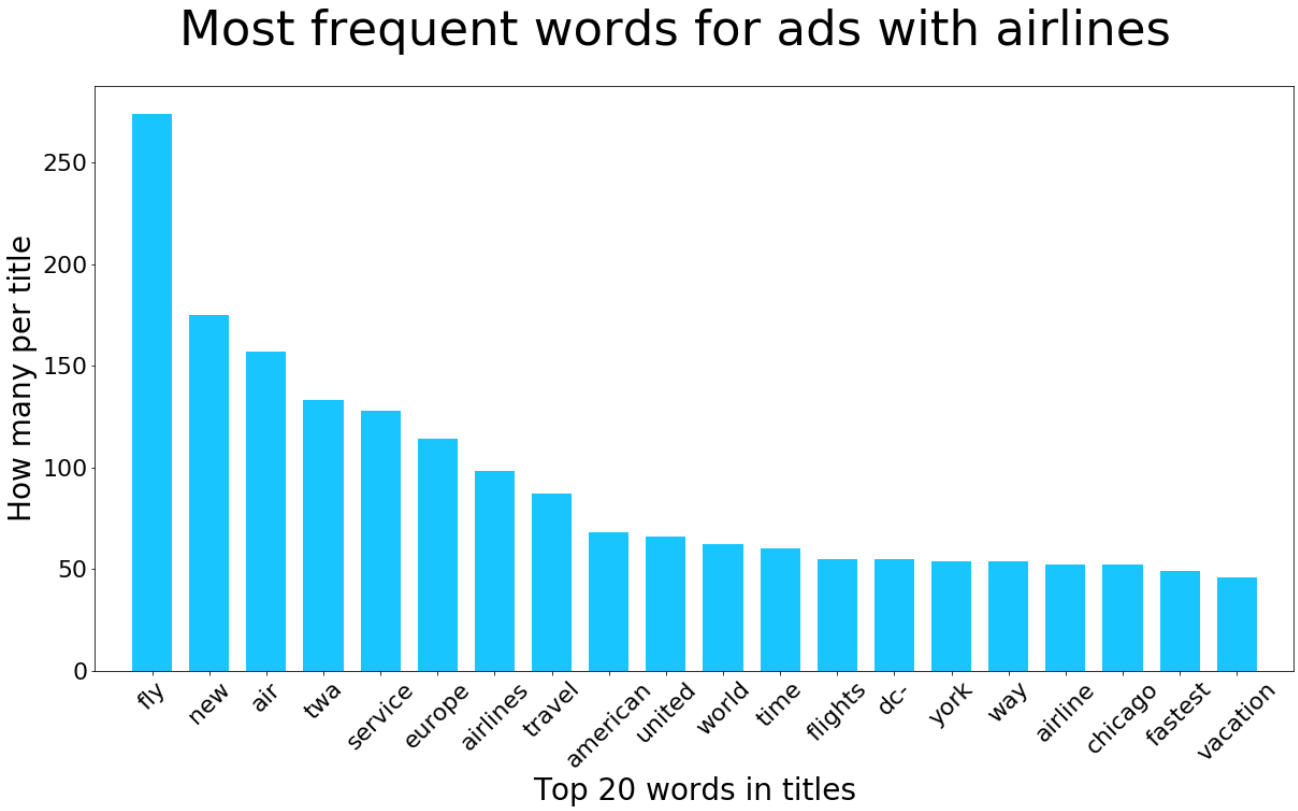
- Children: Ads which target children are mainly the ads from Kodak as well as ads for home- and family-related items, which helps to draw a pastoral picture for ads that are for family products.



- Animals, there we see scrapbook, stud Marshall are common which can explain why those topics were relevant in the ads from the 19th century data.

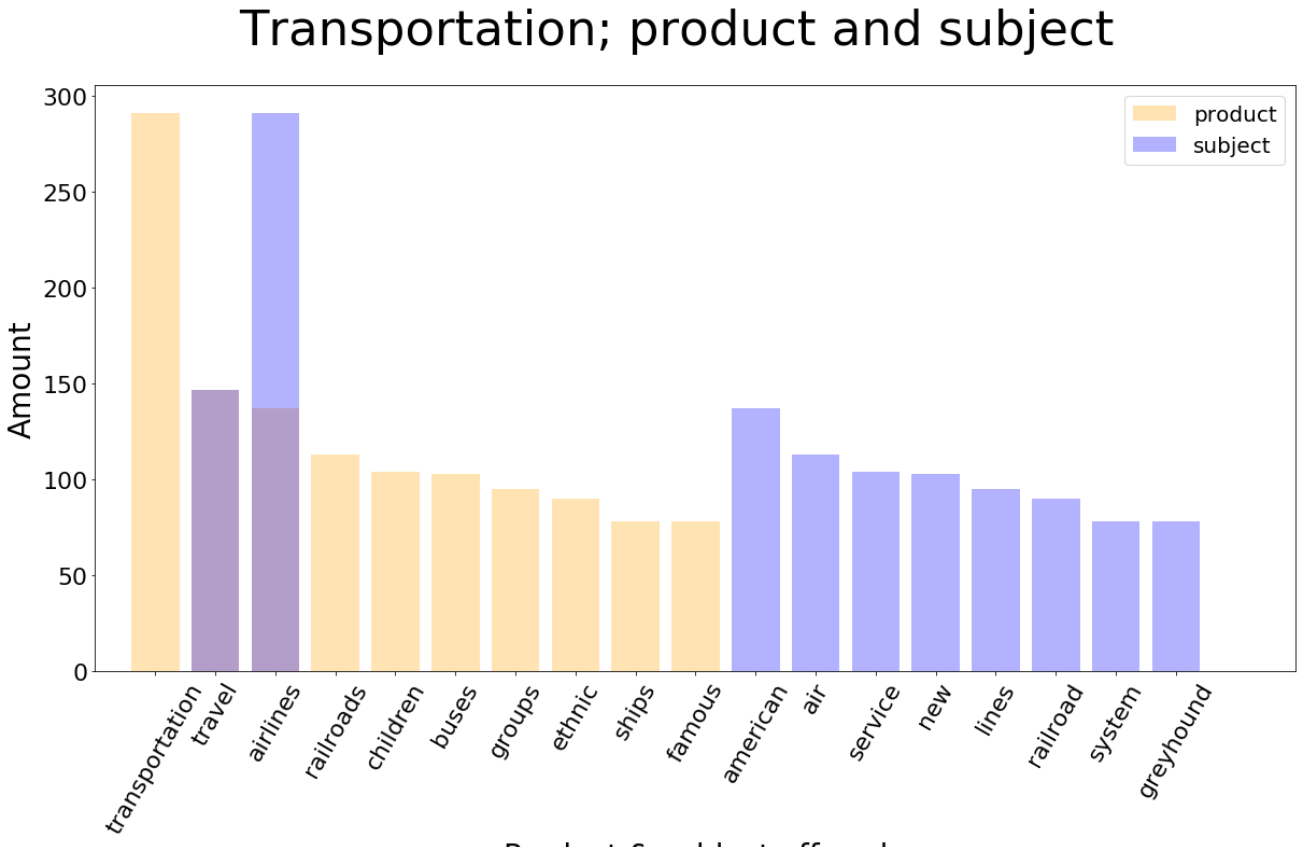


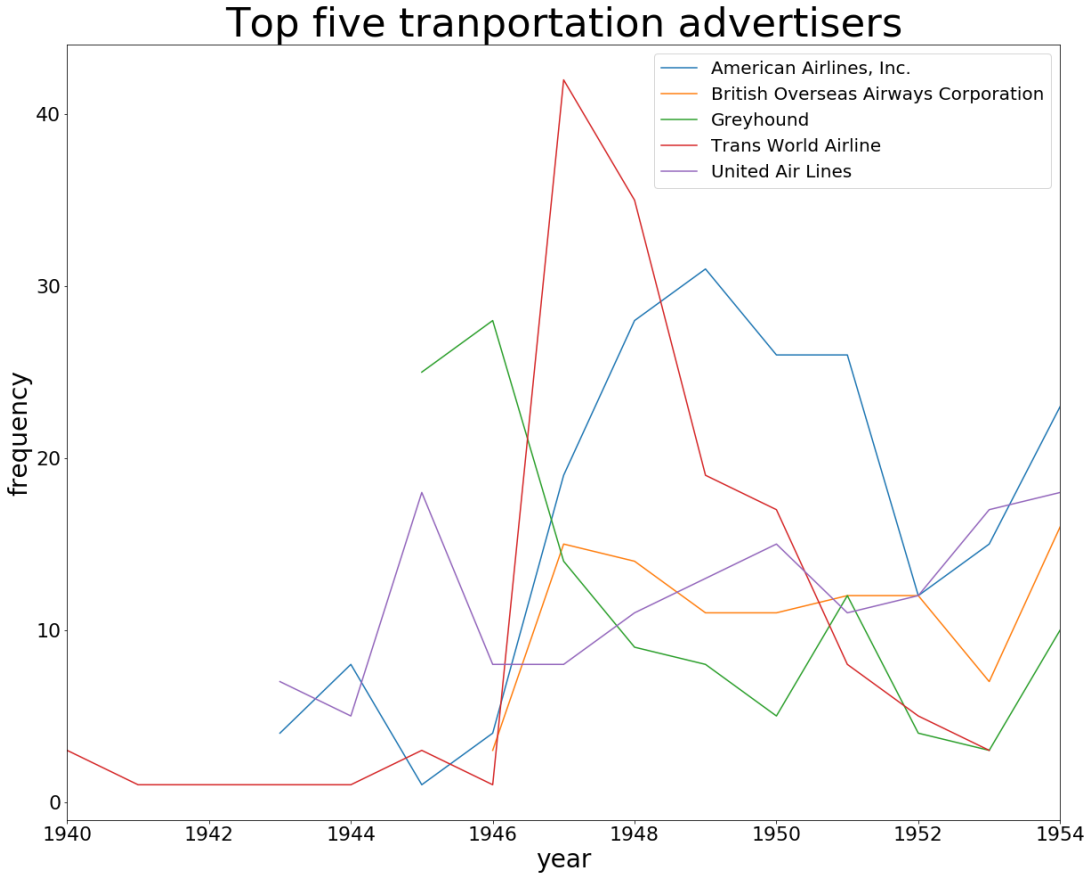
- Flights: The most frequently-used words are "fly", new names of airlines, and the destanations and services airlines offer.



- Transportation: looking at the transportation closely we got to see a few interesting points:
  - ✓ Airlines were the biggest advertisers, we have some ads for railroads, ships and busses, but none of those companies are in the top 3. The top 3 are: American Airlines, Trans World Airline and United Airlines.
  - ✓ The top subjects were the combined subjects, where the collector gave a general topic.

✓ Looking at the top 10 products and subjects, there is some overlap, with a glimpse to the main transportation ideas from then: Airlines and flights, railroad, ships and busses.



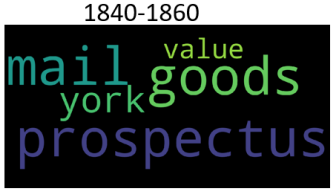


❖ New:

✓ During the Civil War there were new laws.

- ✓ During the great depression there were a lot of hygiene products.
- ✓ New appears often as part of a name of a place. (New York, New Orleans, new England and New Memphis) which shows that "new" is usually meaning less for the ad, and it is just part of the name of the place.

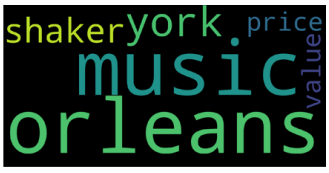
1840-1860



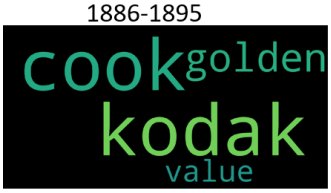
1861-1865 the civil war

1866-1875

1876-1885



1886-1895

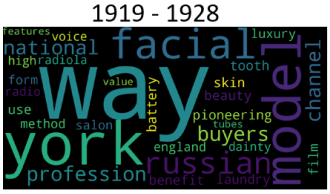


1896-1905

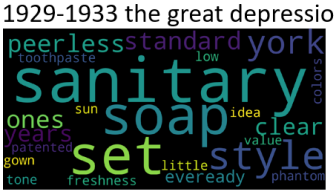
1906-1913

1914-1918 WW1

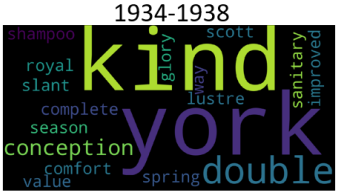
1919 - 1928



1929-1933 the great depression

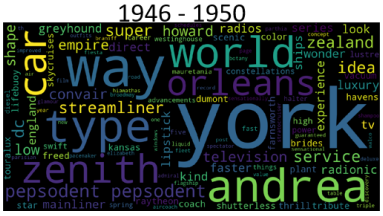


1934-1938

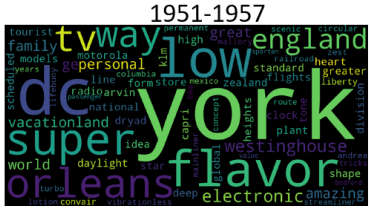


1939-1945 WW2

1946 - 1950



1951-1957



Words after new

- ❖ K-means Clustering:
  We used the k-means algorithm on two merged databases, the titles from the data with the subject of children and the titles from the data from the tobacco category. These two were selected since they should be different and therefore, we can check if the K-means algorithm can classify these subjects based on the titles.

Results from K-means on the words in the ads with the subject of children vs. the tobacco collection, we did once with 10 top terms and once with 15 top terms.

Top 10 terms per cluster:

Cluster 0: probably tobacco	Cluster 1: probably children
Image, actresses, holidays, series, jokes, caporal, sweet, americans, coins, nations	Kodak, new, family, children, christmas, home, story, television, beauty, brownie

Top 15 terms per cluster:

Cluster 0:	Cluster 1:
Kodak, new, family, children, Christmas, home, story, television, Beauty, brownie, tobacco, tv, Book, business, cigarettes	Image, actresses, holidays, series, Jokes, caporal, sweet, americans, Coins, nations, great, flag, Doings, terrors, national

This analysis shows that the clustering for the 10 top terms was successful, where the split seems to match the topics from the data set. However, when clustering the top 15 terms, no words moved within groups, but we found out that the results did not match the topic subjects in the data set. This result may be due to an imperfection in the K-means clustering words, and more so when discussing how words were used 100 years ago.

Predictive analysis:

Prediction with K-means:

Using the title given by the Duke Library for ads, the K-means clustering was meant to predict what the subject of the ad was. For this, we used ads with the subject/ category "children" or "tobacco."

The resulting confusion matrix:

	Children from the data	Tobacco from the data	Total predicted
Predicted children	933	188	1121
Predicted tobacco	7	228	235
Sum total in data	940	416	

Accuracy level is 77%.

Children positive is 99.25%, Children negative is 0.75%

Tobacco positive is 54.8%, Tobacco negative is 45.2%

The unexpected result was that the prediction was quite good, even though the k-means seemed to prefer choosing the bigger group, therefore the children topic which was much bigger was almost always predicted correctly, while the tobacco was predicted a bit more than half out of the times.

Impediments:

1. Difference between data sets: When we started to look at the data in graphs, we observed that each data collection has its own topic, and therefore some observations are not historical but are affected by the choice of the collector. For example, Kodak vanished from our data in 1920 because the second data set doesn't have ads from all topics. In order to deal with this difficulty, we made use of three different databases, one of the merged data, one for the old data and one for the new data, and we used each one for the appropriate research questions.
2. When looking in depth at the data, we saw this "boardwalk" category which we didn't understand what those ads were, so we opened the site and looked at the "boardwalk" ads and realised that these were pictures from of boardwalks. We decided to remove that category from the databases since that category isn't ads and therefore not informative.
3. Dealing with the words appearing immediately following the word "new" in ads: Here we had two main challenges to deal with, using regex so we can find the words and erasing all the "nan" and other problematic strings that were placed by Python instead.

Future work:

More research can be done on the data with the following questions:

1. Trying to find specific words for set products/subjects or companies.
2. Checking about prediction of clusters based on years.
3. K-means on more categories with more clusters.
4. Are we able to classify different companies with K-means or PCA for example.
5. Checking if hierarchical clustering can be applied here.

Conclusion:

We can see that occasions can change how ads would be worded and what will be advertised and there is a difference of the ads over time. Titles can be predictive for the subject, and different category will have different wording in the advertisement.