Overview and Motivation:

The motivation for this project is to use local data. People say shop local, support local, etc. We believe that this also applies to data. Incorporating Utah newspaper data is the way can be local and use data visualization to connect to Utah's history and past, making this project a little more meaningful. The Marriott Library has scanned over 36 million Utah newspaper documents whose data is accessible via their public API, this data is what we are visualizing.

Related Work:

Media has a rich history of utilizing visualizations. National newspapers like the New York Times and local newspapers like the Salt Lake Tribune and the Deseret News use visualizations extensively. However, little work has been done to visualize the newspaper data itself. If you can find anything outside of a simple word cloud, let us know.

In terms of word analysis, the idea of "stop words" is something we came across when cleaning text data. Here is an NLP perspective on stop words that we came across and inspired us to make a "stop list":

https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

Questions:

Initial Questions:

What words are most frequently used in a paper?

What does the archive itself look like in terms of content distribution?

How do different newspapers compare when covering certain topics at a certain time? (An example would be the Magna Earthquake of 2020)

Due to the limitations of OCR, and how text is scanned digitally, the third and first questions may not be as feasible. (More on this in the data section)

Current Questions:

What words are most frequently used in a paper?

What does the archive itself look like in terms of content distribution?

What counties and cities have the most scanned documents in the archive?

(This perhaps answers if there is a “gap” in coverage in the archive)

What is the date range of content within the archive?

How can we use a map of Utah to visualize this data?

Data:

Source, scraping method, cleanup, etc.

Data Source:

Our primary data source is the Utah Digital Newspapers (UDN) archive API that is provided to the public by the Marriott Library. We acquired supplementary data from the Utah Geospatial Resource Center (UGRC), a Utah state-run repository of GIS data. Lastly, publication location data was not part of the data set, so we had to manually search for this data using a combination of searching with Google or looking through the scanned documents contained in the UDN archive.

Scraping methods:

Initial scraping methods were made in Python using the standard curl library. We simply called the API and stored the response as a JSON object. Our project is intended to be dynamic, meaning that it leverages the public API on the go and does not hold any static data. However, there are potential problems with this. Relying on a third party for data can be risky, so we store some data locally.

Data that we chose to store locally:

With so many newspapers, it is hard to visualize all of them. For static and consistent data in the chance that the API we relied on is closed, we chose to store three newspapers published on specific dates from these four publications. Each publication serves a distinct area of Utah.

We analyzed these papers:

Salt Lake Tribune (Salt Lake City area)

Garfield County News (Garfield County, southwestern Utah)

The Herald Journal (Logan area, northern Utah)

Vernal Express (Vernal area, eastern Utah)

Across these three dates:

1-1-1995

4-1-1997

9-12-2001

Resulting in 12 total papers stored locally.

In addition to this data, we also opted to hardcode some data. We have a list of all publications, their document counts, publication locations, and the range of dates in the archive stored in JSON format.

Cleanup:

Data cleanup was necessary for the OCR (optical character recognition) data acquired from the UDN archive and for some geographical data obtained from the UGRC. The data cleanup for the UGRC data was minimal. It consisted of fixing a few spelling errors in the data, such as “Summit” county being spelled “Summitt.”

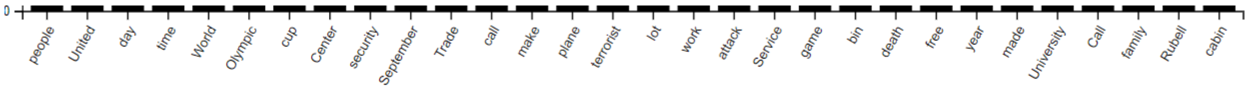
OCR data is text scanned by a computer, all newspaper text data in the archive, such as articles, advertisements, and classifieds, were transcribed to digital text using OCR. OCR, unfortunately, is incredibly inconsistent. Different typefaces, human-made typos, the darkness of ink, scan lighting, and equipment all impact how accurate OCR is. In some instances, OCR transcribes every word perfectly, in other transcriptions, it is impossible to read. Fortunately, OCR is virtually the only data we had to clean, unfortunately, it offers some of the most interesting data as it is the literal content of a newspaper.

First, for some context, we created a publications object that holds all the data of a certain publication as well as a word frequency array of that publication A word frequency array is an array that denotes the number of occurrences of a word, here is an example [{The,90}{Cat,14}], this array is derived from the text data of a publication.

In order to store good data in this array and not meaningless data, we used regex to strip the data clean of punctuation characters such as “,.’” In order to analyze word usage and further refine this data, we employed a stop list technique (see the related work), and imported a dictionary to check if the words we scanned were valid. The condition of counting a word and adding it to the word frequency array is as follows: We add a word if the words were not in the stop list but are in the dictionary. We do lose data when we clean up; however, it is necessary.

This type of cleanup requires a good stop list, a JSON dictionary, and the thoughtful use of data structures and basic algorithms. If the cleanup is implemented naively, the code can run too slow, we tried a simple naive method for cleanup but realized we had to optimize as our performance was noticeably poor, the dictionary, publication text, and stop word list can easily exceed 100,000 elements. To illustrate why this cleanup is necessary, we leave you with two X axes of a bar chart derived from the same publication. One where minimal cleanup was taken and one where cleanup was employed. It should be apparent which one can better distinguish the word choice of a paper.

0
the
to
and
a
in
of
for
is
on
that
i
The
with
was
at
from
be
"
as
said
are
it
have
Salt
will
—
or
Lake
by
an



A Data Sidenote: Limitations of responsible API usage and poor API design

The API we use is not efficiently designed. For instance, finding a paper that was published closest to a date can take 30-100 API calls. The API does not offer the closest paper when using their find paper by publication and date get method, rather it returns if a paper exists or not, forcing us to increment the date and call again until we find the next issue of the publication. This presents a problem, we are unable to efficiently consume the API. In addition, the analyses we want to conduct (such as exploring the use of a certain word over time in a newspaper and its publication) could take thousands of API calls. While this isn't a lot on paper, this API is a public resource. Consuming or hoarding compute power is poor practice and is something that we think heavily about while using it.

Exploratory Data Analysis:

We initially started only by analyzing word frequency but have now moved on to analyzing the archive itself and what data it doesn't have (see the questions section). We initially used simple bar charts for our visualizations. These get the job done in terms of comparing quantities but lack a visual appeal. They are easy to implement, however, and are a time-tested, familiar design.

Additionally, since the data in the archive is incomplete (it does not have the most recent publications from newspapers), we have provided information about the date ranges for each publication in the archive. This can be viewed either by using the map or the timeline. With the map, it can be seen on a per-city basis. Upon clicking on a city on the map, a table will be filled listing each newspaper from that city and the date range for which the papers' publications are contained within the archive. Through the timeline, users can see at a glance approximately how long each newspaper is contained within the archive. They can furthermore hover over a line for a newspaper to see specific details below the timeline.

Design Evolution:

Barcharts, while simple, are boring (at least in our opinion). As a result, we have also decided to incorporate bubble charts. While this type of visualization is flawed, it seems to work well in datasets where there is an extreme outlier compared to a bar chart. Drastic "size" differences can be easily compared. However, these types of visualizations make it hard to compare things that are close in value to each other. Our brains struggle to see the differences in slight area differences in bubble charts compared to bar charts where the bars have the same baseline. Additionally, we added a map which is useful for seeing the geographic location a newspaper is published in. Maps provide a unique perspective on the data that is lacking in bar charts or bubble charts, and we feel that they help users relate to the data in ways that purely categorical or quantitative data visualizations lack.

We opted to add a line chart to show the use of words in a newspaper over time.

Why a line chart?

Many visualization techniques work well in mapping changes over time. However, a line chart offers two distinct advantages for our visualizations, multiple lines or word usage can be visualized in a single chart, and easy visualization of the slope(how word usage has changed with time). Adding multiple data sets for the purpose of comparison does not work particularly well with a bar chart, nor does it work with a heat map calendar, while adding multiple sets is possible with a scatter plot, our brains would just be connecting the dots together to form lines anyways, this is also called connectedness and is a gestalt principle.

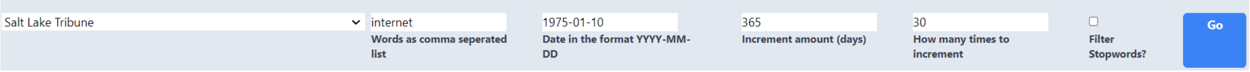
A Timeline for Historical Perspective

Lastly, we gathered data to implement a timeline of newspapers contained within the UDN archive. While the archive is incomplete, this can still provide users with a unique perspective of how long newspapers have tended to exist in Utah. Timelines are useful for providing a unique historical perspective that can be difficult to achieve using other methods. Additionally, using a timeline can provide valuable information regarding the usefulness of the archive. If, for instance, users wanted to view information about the COVID-19 pandemic from the Salt Lake Tribune, they can quickly discover that this is not possible since the newspaper is not contained in the archive past December 31, 2004.

Implementation:

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

The core visualization is a line chart, that shows the use of words in a newspaper over time. The user can supply the newspaper, words, and date range they want to visualize. Granularity in the form of increment amount(how many other days to get a paper), and times to increment(how many distinct dates the user wants to get) is given to the user. A user may only want to visualize every Christmas paper in a 5-year range, the granularity we give allows for this. Granularity is also needed to control the cumbersome size of data that needs to be pulled from the API. 30 days of newspaper data can get up to 1000 MB data, the user may not have the bandwidth, local storage, or time to deal with this size of data. Below are the input fields as mentioned above that the user interacts with to produce a line chart.



Salt Lake Tribune

internet
Words as comma seperated list

1975-01-10
Date in the format YYYY-MM-DD

365
Increment amount (days)

30
How many times to increment

☐ Filter Stopwords?

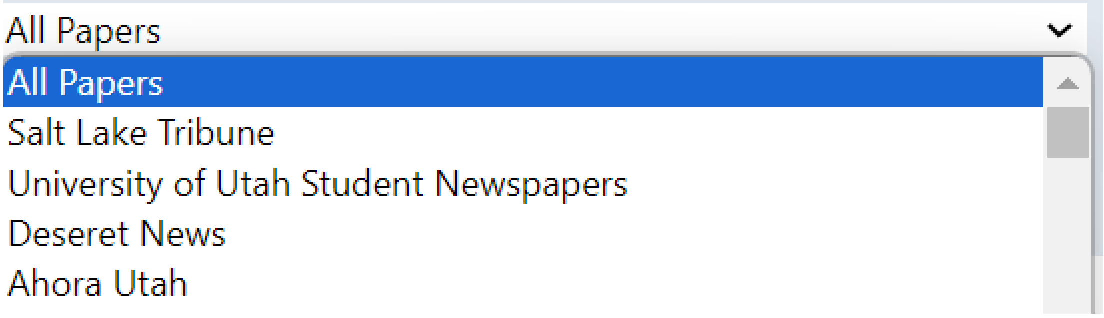
The rest of this section will go into greater depth on important aspects of the interaction.

Stopwords and dictionary optional:

We allow the user to determine if they want to apply a stop word filter or not. Many words seen in newspapers such as slang, typos, and names are not found in the dictionary, typically these words get filtered out, by removing the stop word filter, the user can search for any word of any spelling such as ‘y2k’. However, this does come at a cost of storage as many unimportant words such as “the”, “and” etc are stored and counted. The cost of storage is not trivial when removing a stop word filter, applying a stop word reduces memory usage by about 50%, thus the user should consider if what they are searching for necessitates the removal of the stop word filter.

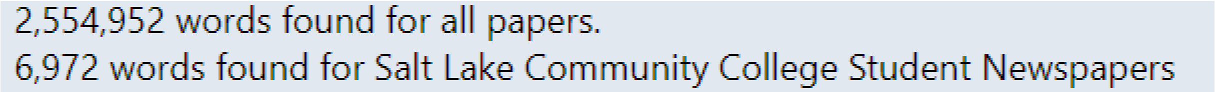
Analyzing all papers as a whole:

Users can choose to see the word frequency line for all papers found in their date range or just their specific paper. This is useful if the user wants to analyze all newspapers published in Utah as a whole and is not focusing on a specific paper. This all-papers selection is the default in the newspaper selection dropdown menu



Showing the number of words analyzed in a user query

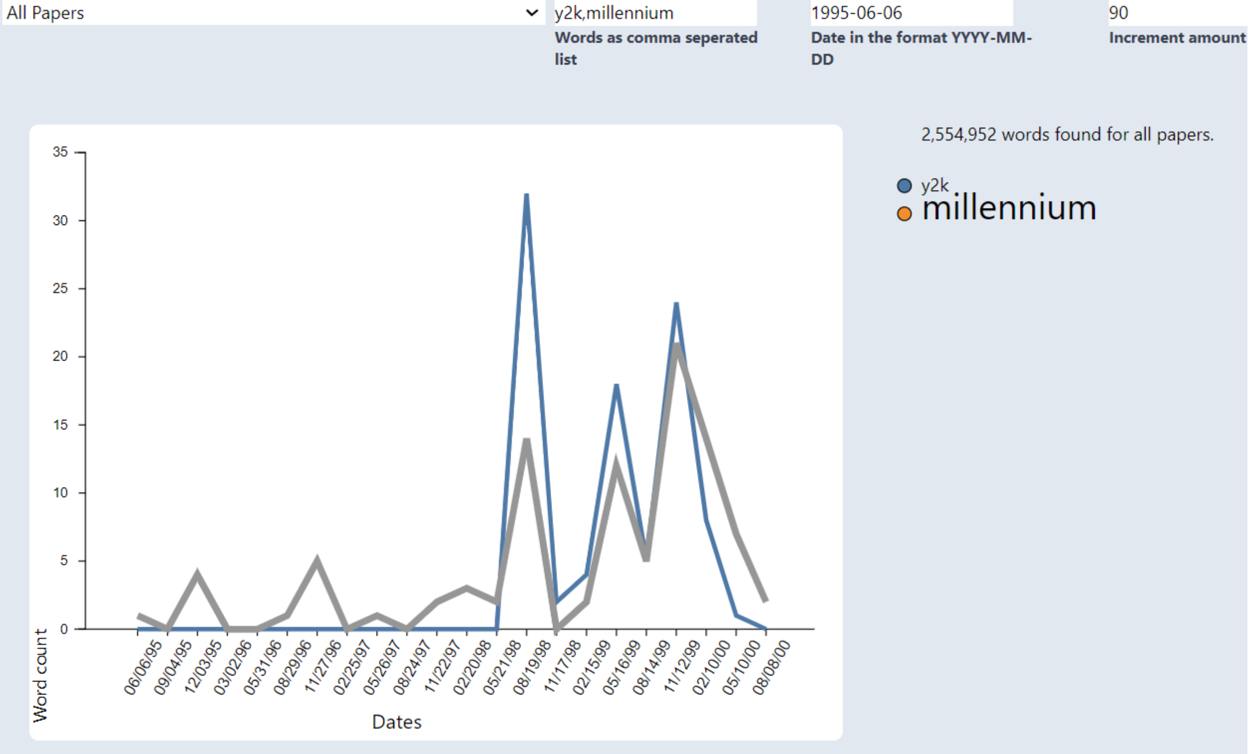
For each user query, we show the total number of words scanned from the entire collection, and also the amount of words scanned from the specific newspaper they chose. This is to show the user the magnitude of data that they are working with, and to understand how much their chosen publication contributes to the over word count of all newspapers found in their given date range. Below is an example



2,554,952 words found for all papers.
6,972 words found for Salt Lake Community College Student Newspapers

Mouse hover and lines

With many lines in a line a chart, it is easy to lose sight of certain lines, we have added mouse hover functionality that increases line size and makes it gray. This hover functionality also works with the key, and makes the key item text bigger. Below is a picture



All Papers

y2k,millennium
Words as comma seperated list

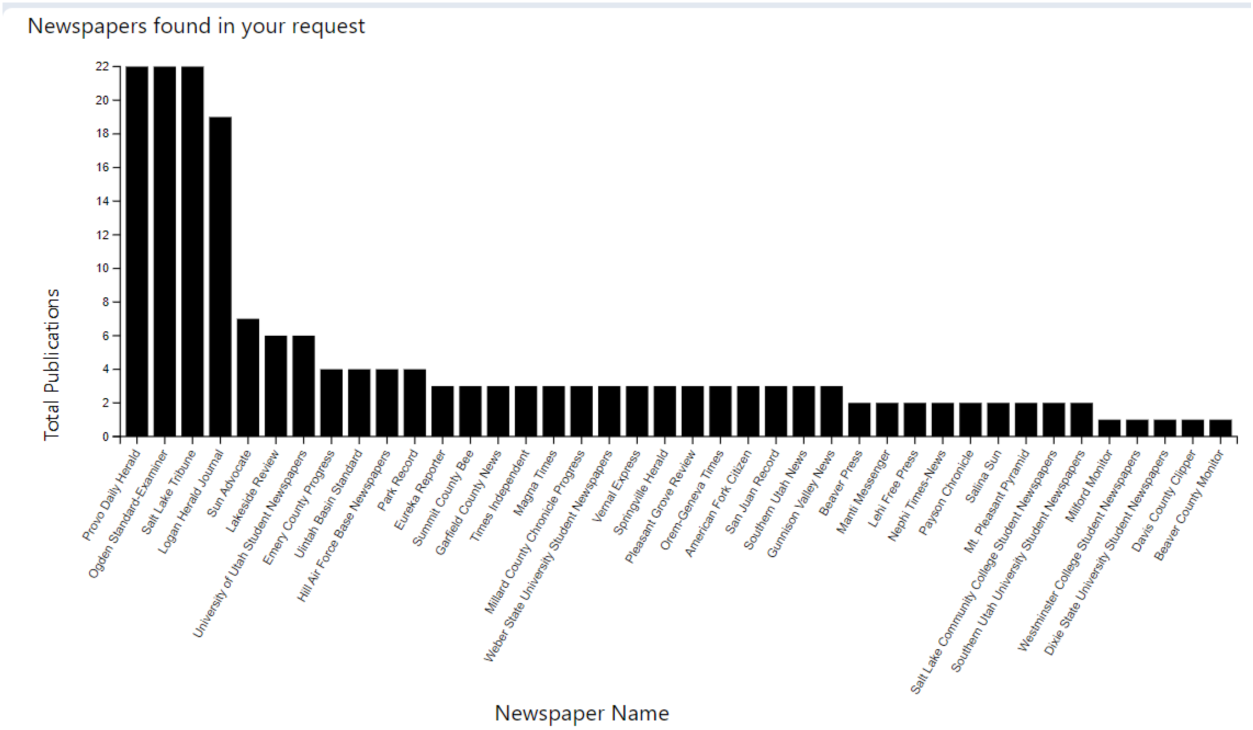
1995-06-06
Date in the format YYYY-MM-DD

90
Increment amount

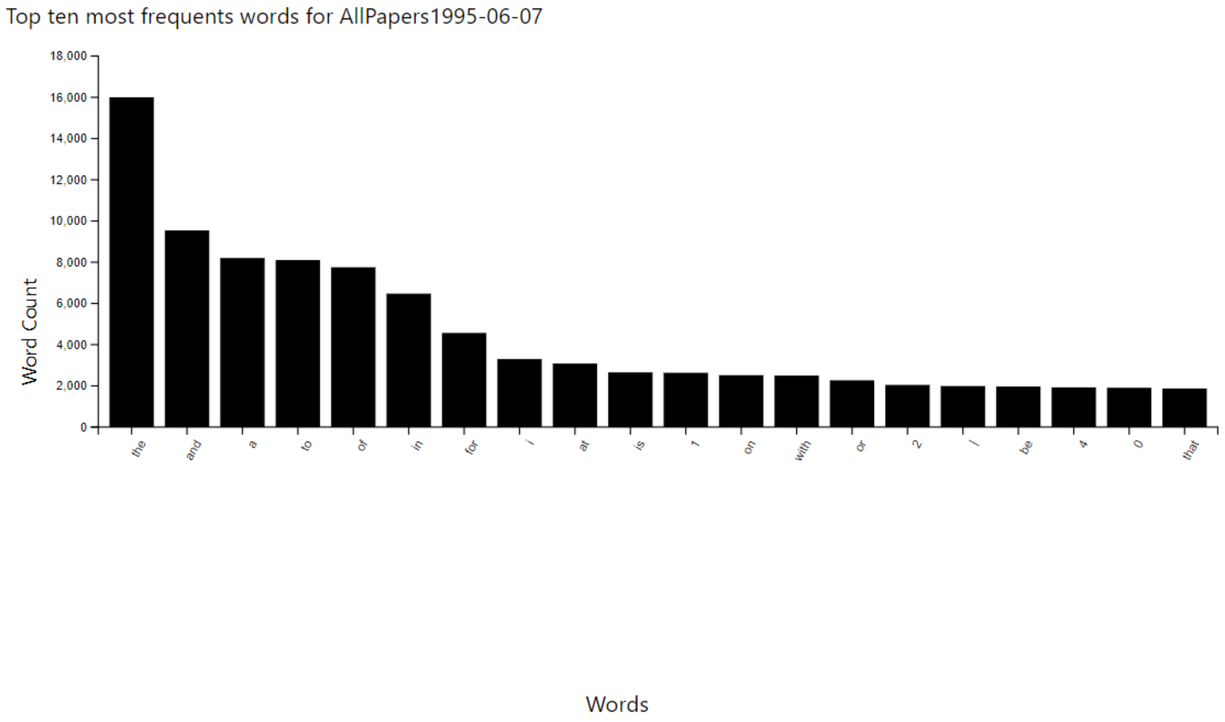
2,554,952 words found for all papers.

Supplemental Visualizations

In addition to the line chart, two types of bar charts are generated per user request. The first is the publications found in the date range the user provided. This allows the user to see what other papers were being published in their date range as well as the amount published, this also encourages exploration of other newspapers. The second bar chart is a word frequency chart that displays the top ten most used words of the paper the user requested. A bar chart is made for every date that the user requested. Below are some examples



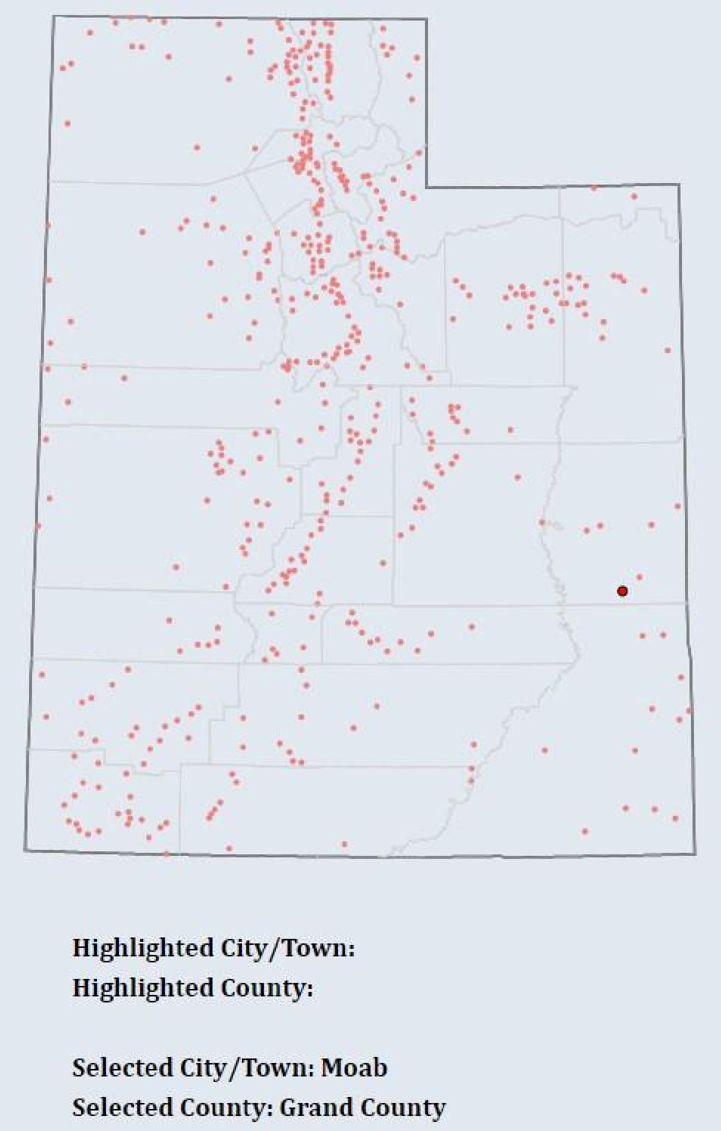
(Newspaper publications found in a user request)



Word frequency chart, the stop word filter was removed when generating this chart.)

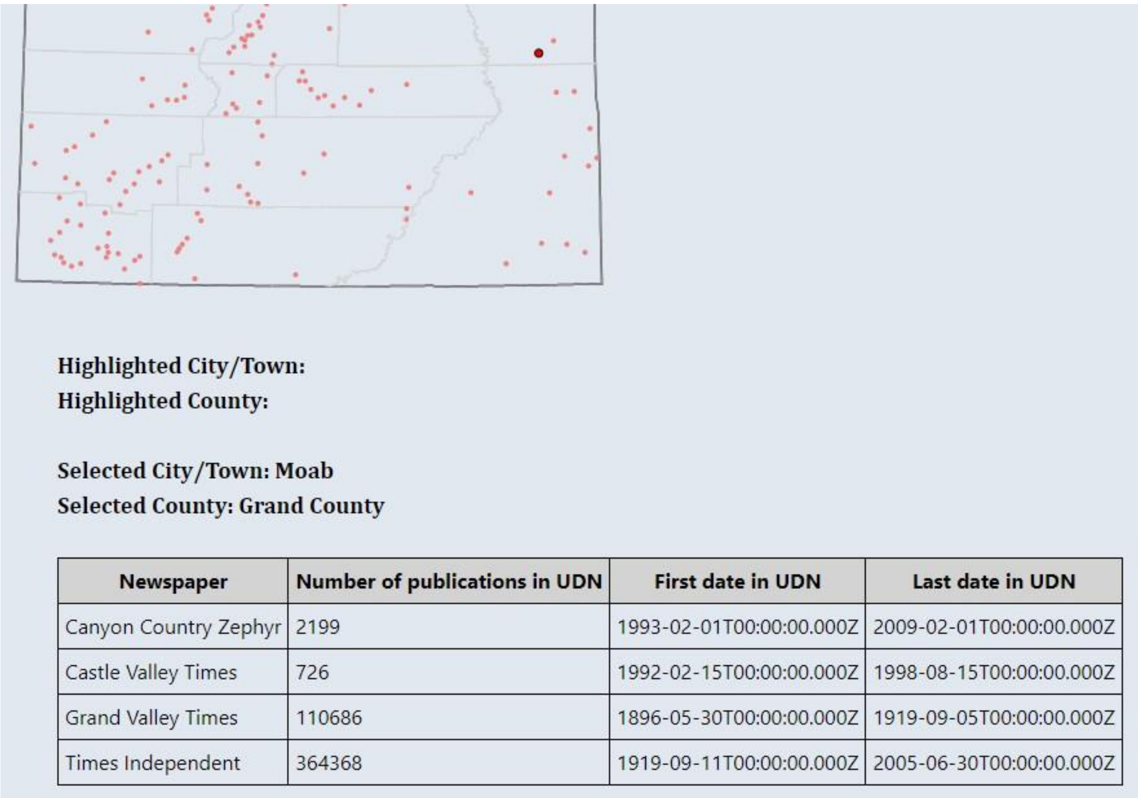
Map of Utah

The map of Utah provides users with a geographic perspective of newspapers in Utah. By clicking on a city, they can see a list of all newspapers for that city within the archive as well as how many publications it contains and over what date range. This allows users to explore in detail how different cities and regions of Utah are represented by newspapers within the archive.



Highlighted City/Town:
Highlighted County:

Selected City/Town: Moab
Selected County: Grand County



Highlighted City/Town:
Highlighted County:

Selected City/Town: Moab
Selected County: Grand County

Newspaper	Number of publications in UDN	First date in UDN	Last date in UDN
Canyon Country Zephyr	2199	1993-02-01T00:00:00.000Z	2009-02-01T00:00:00.000Z
Castle Valley Times	726	1992-02-15T00:00:00.000Z	1998-08-15T00:00:00.000Z
Grand Valley Times	110686	1896-05-30T00:00:00.000Z	1919-09-05T00:00:00.000Z
Times Independent	364368	1919-09-11T00:00:00.000Z	2005-06-30T00:00:00.000Z

A table is filled out for the selected city. This table displays what publications exist in the UDN archive, the number of publications it has, as well as the first and last dates contained in the archive.

Timeline of Newspapers

By implementing a timeline, we have allowed users to see at a glance over what date range each newspaper is present in the archive. This allows users to more deeply understand the value of the data they are looking at and how it may be used. For instance, if users are seeking information about the 2020 Magna earthquake, they would be able to see that the Sale Lake Tribune would not be represented in the data since it is only represented in the archive up to 31 December 2004.

If users desire to see more detailed information about the date ranges represented in the archive, they can see the exact dates displayed below the timeline when they hover over the line representing the newspaper.

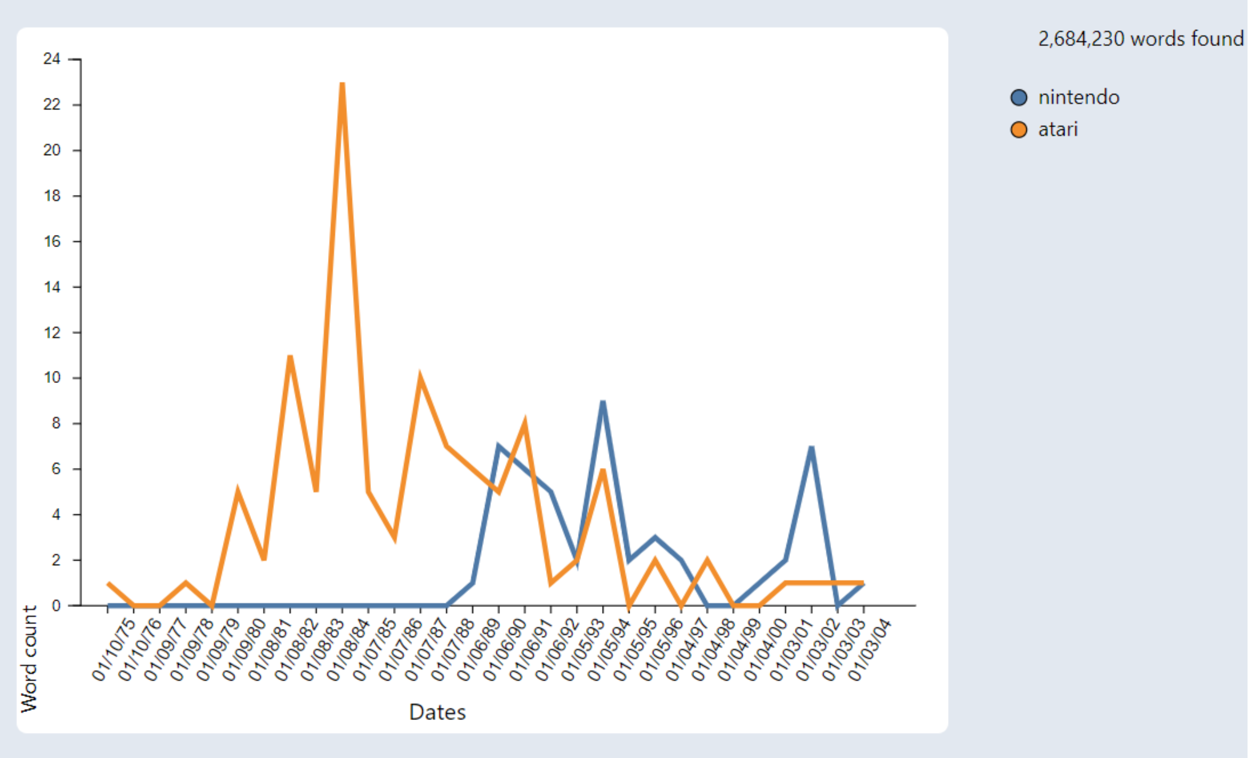
To aid in usability, we also chose to implement the timeline with a vertical scrolling visualization as it was otherwise too large, as it represents over 300 newspapers.

Evaluation:

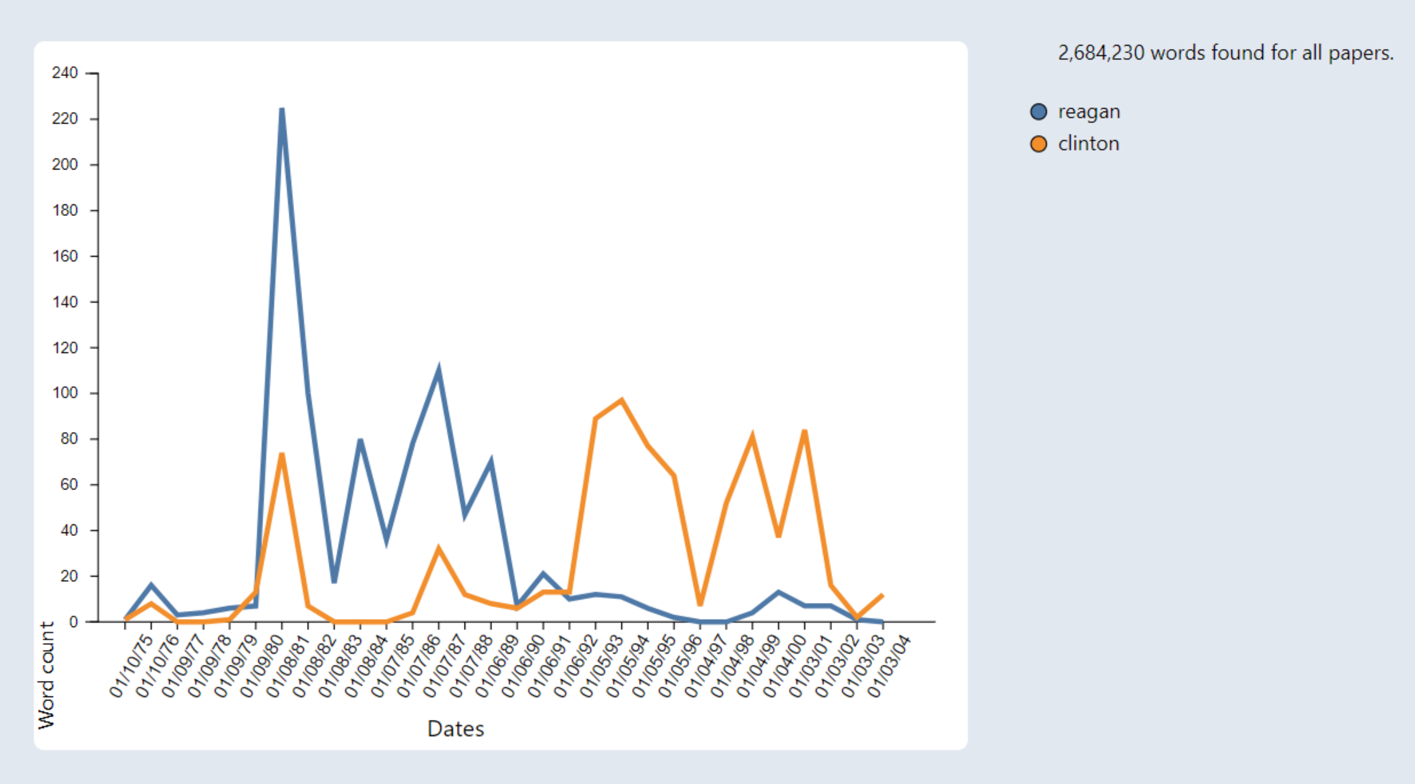
What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

The line chart visualization is particularly effective in understanding when issues, people, technologies, and policies become important in mainstream culture and when they lose relevance. There is an assumption here that words mentioned in newspapers and the frequency they are mentioned, indicate their relevance and popularity. Here are some interesting examples of the questions they answer.

When did Atari and Nintendo start gaining media relevance in Utah?

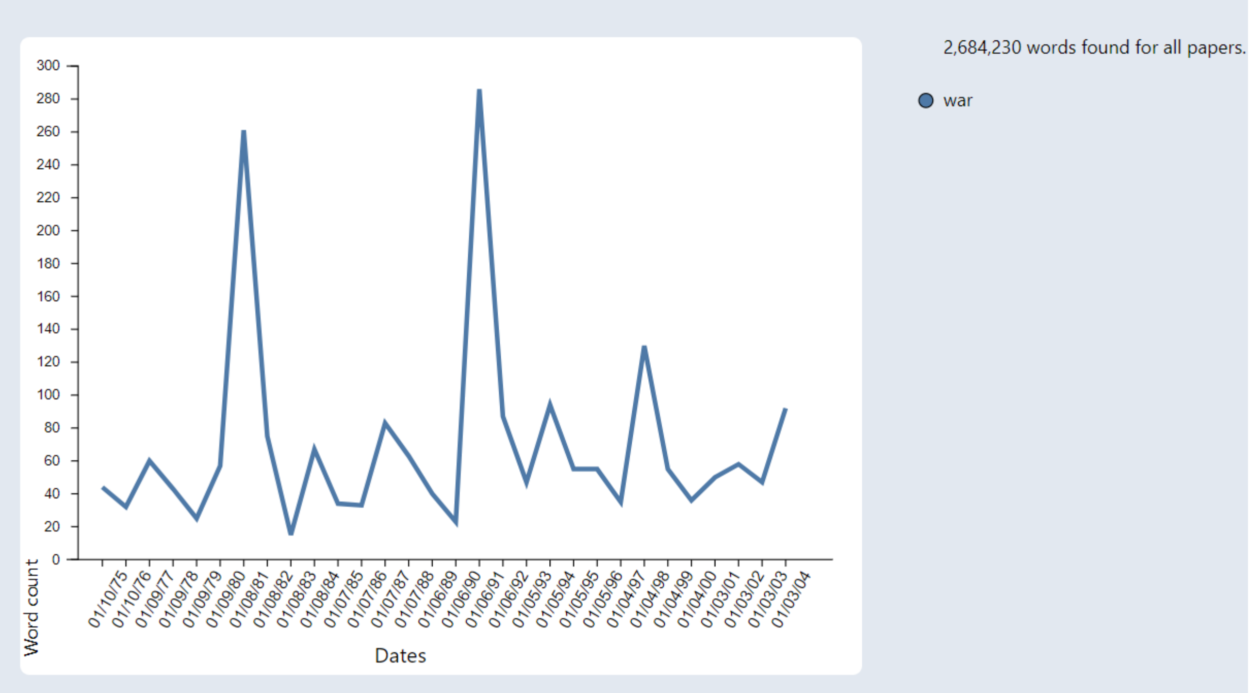


At what times were Reagan and Clinton most mentioned?



2,684,230 words found for all papers.

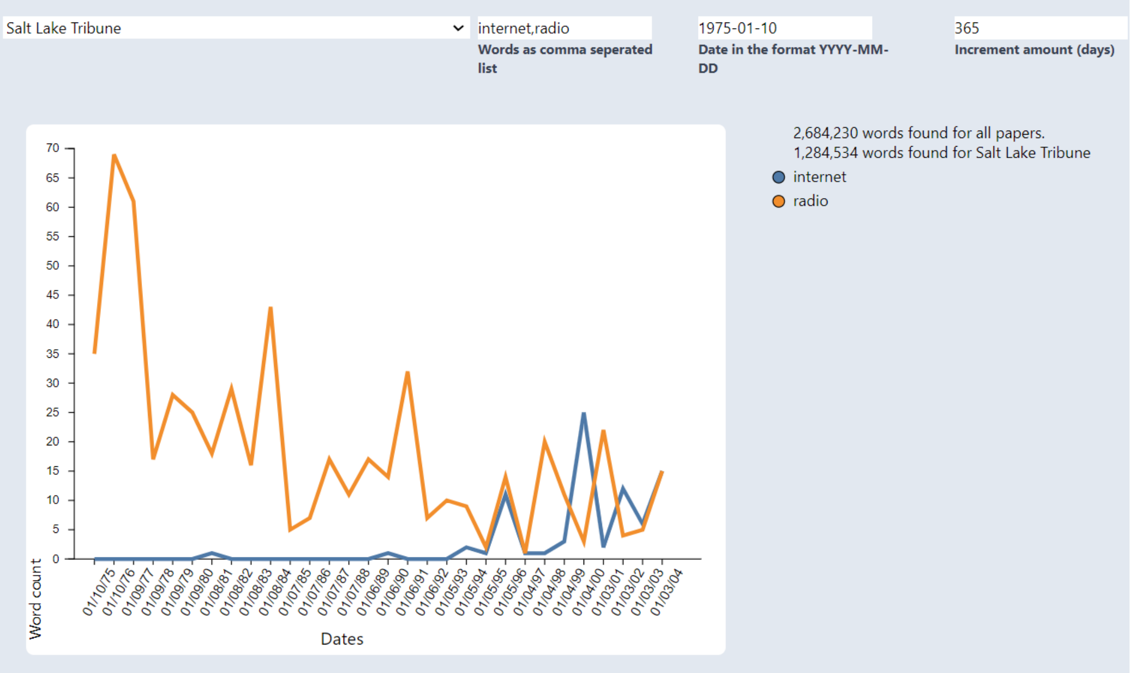
If I don't know the history of wars what are good years to study to understand the conflicts that were going on?



2,684,230 words found for all papers.

(note that 1990 is the start of the Gulf War).

How does radio relevance compare with the internet over time?



Salt Lake Tribune

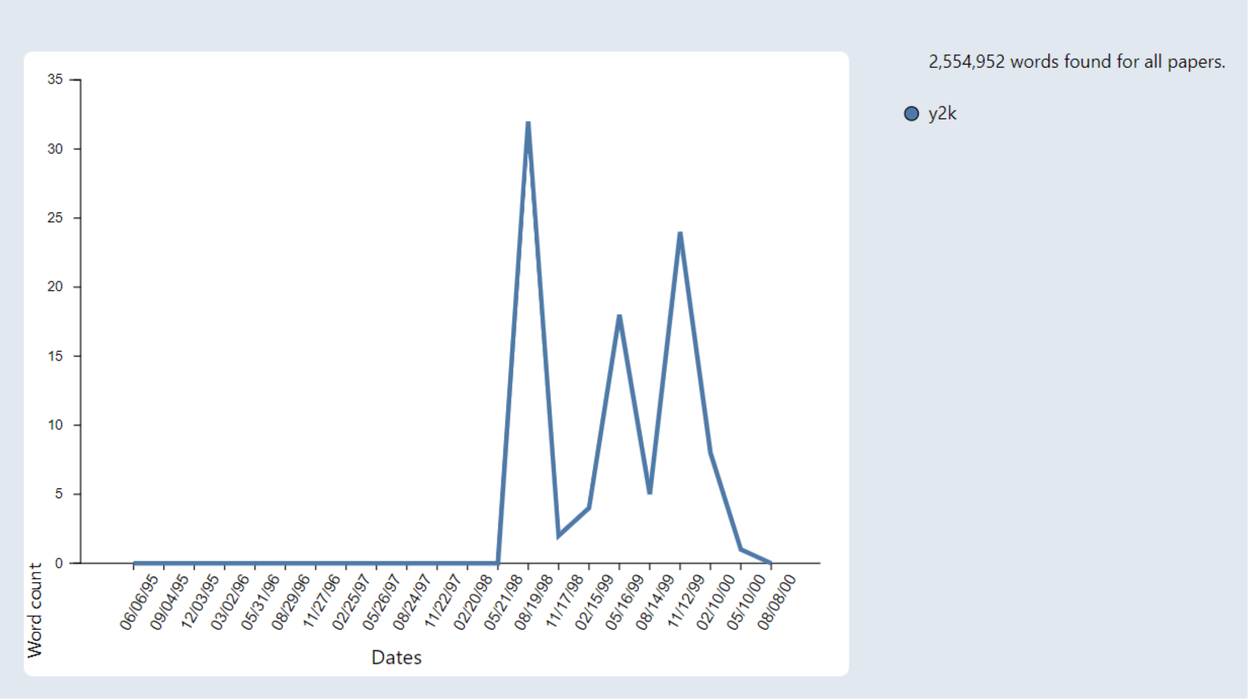
internet,radio
Words as comma seperated list

1975-01-10
Date in the format YYYY-MM-DD

365
Increment amount (days)

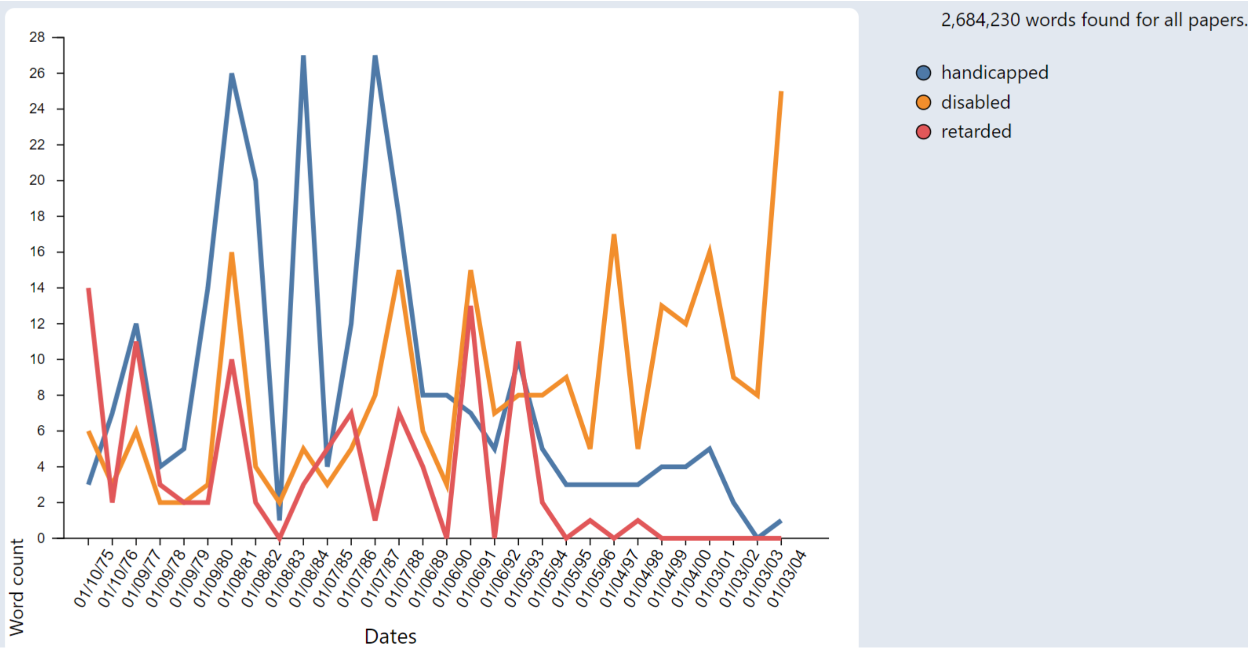
2,684,230 words found for all papers.
1,284,534 words found for Salt Lake Tribune

When did Y2K panic start?



2,554,952 words found for all papers.

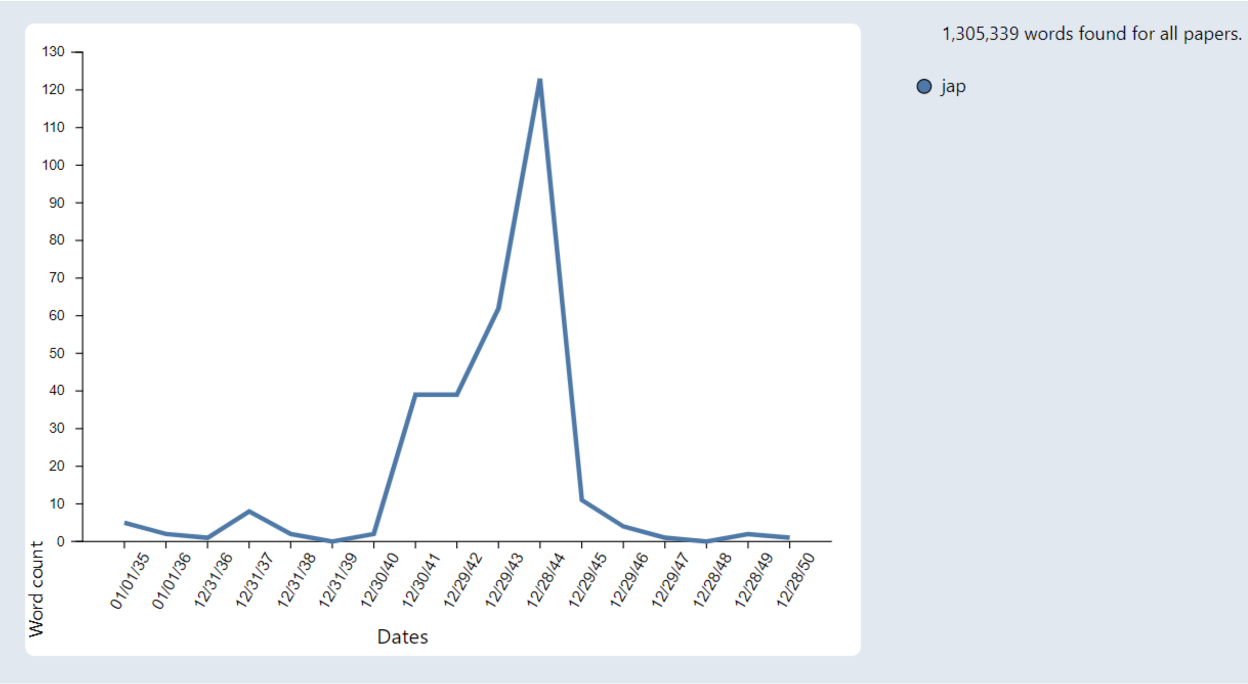
ADA impact on word choice



2,684,230 words found for all papers.

(The ADA was passed in the early 90s)

Racist slang usage and war world 2, when was it at its peak?

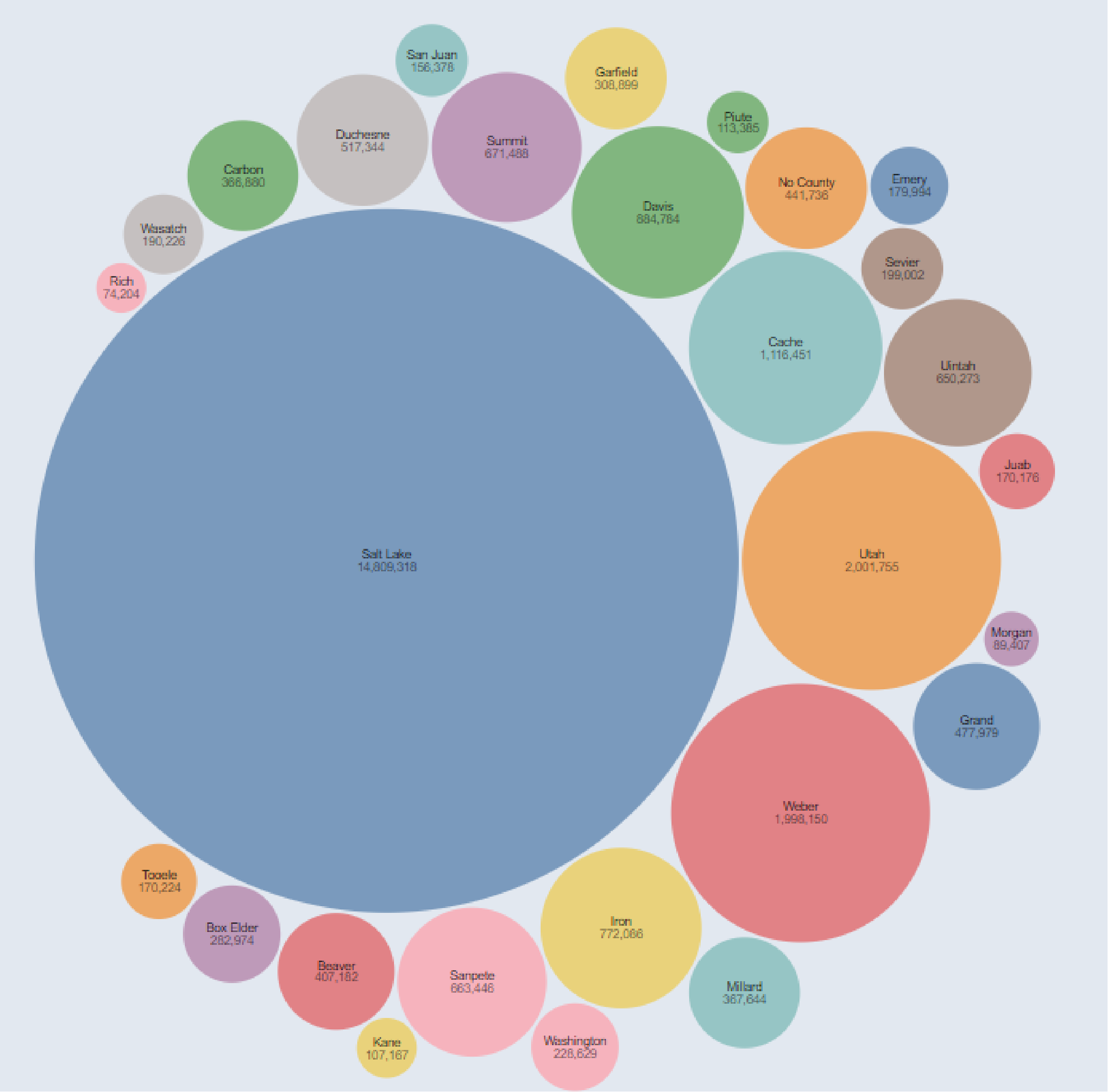


1,305,339 words found for all papers.

A use case of these charts is for people to convey when certain things were popular, however, the line charts do not necessarily explain why. The line chart could further be improved by allowing users to add a “tag” to certain parts of the lines, for instance in the Atari Nintendo example, the user could add a tag to the peak of the Atari line stating this is when Atari sold their new flagship console.

Bubble chart for counties:

Evaluating the collection itself and the distribution of its data is also interesting. We added a bubble chart to visualize which counties in Utah had the most newspaper publications in the collection. The bubble chart allows for quick comparisons between bubbles, though the downside is that for counties close in publication count like Utah and Weber seeing the difference in area is difficult. Looking at the chart below we see that Salt Lake County is by and far the most represented county in the collection, and that rich county is the least. This is not surprising, people live in cities after all. However, it is interesting that the county St. George is in(Washington) has a lower publication count than Millard or Grand County, despite both of these counties having lower populations.

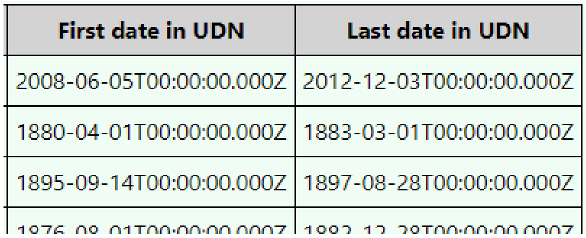


The above visualization could be improved by adding interactivity. If a user clicks on the button it would be nice to see the papers that belong to that county and their respective publication count.

Map & Table

The map allowed us to explore the data on a per-city basis. When selecting a city, a table is filled out displaying information about all newspapers within the UDN archive. This allows us to see the number of newspapers in the archive for specific cities through sheer volume. For instance, it is striking to look at the list of newspapers from Salt Lake City compared to anywhere else in Utah. Yet, it can be difficult to find the city you are looking for on the map. When a user hovers over a city, it is indicated in text just above the map. Yet, they have to know approximately where the city is to begin with in order to find it. If we were to implement a way to zoom in on the map and have city labels appear, it may dramatically help users' ability to find the locations they are looking for.

Furthermore, when there is a very large number of newspapers from a city, the table can become very large. For most cities, this is not an issue. However, with regard to Salt Lake City in particular, the table can become too large. Scrolling would help with this. We could also improve the color choices to increase the experience of colorblind users. Additionally, the start and end dates of newspapers in the archive are not formatted in a user-friendly way.



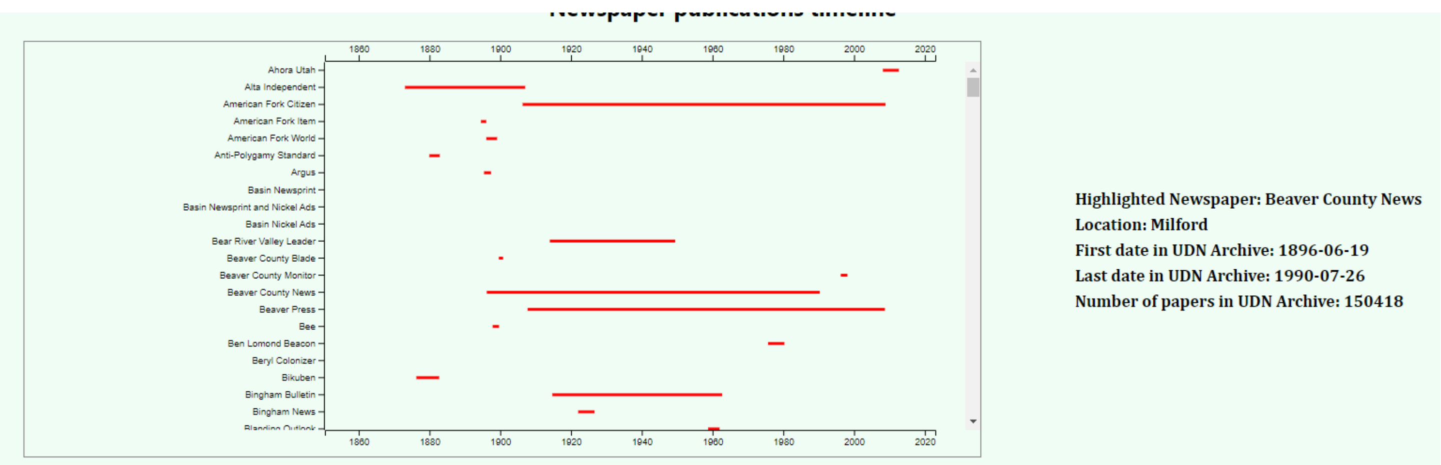
First date in UDN	Last date in UDN
2008-06-05T00:00:00.000Z	2012-12-03T00:00:00.000Z
1880-04-01T00:00:00.000Z	1883-03-01T00:00:00.000Z
1895-09-14T00:00:00.000Z	1897-08-28T00:00:00.000Z
1876-08-01T00:00:00.000Z	1883-12-28T00:00:00.000Z

We decided to leave this limitation be as it is able to be read with some effort, and attempts to correct this issue were less of a priority than other parts of the visualization.

Timeline

The timeline allows users to view the duration of a newspaper's presence in the UDN archive at a glance. This lets them see just how useful different papers are as well as roughly how long newspapers have tended to exist in Utah. The primary uses of this visualization are allowing users to see the usefulness of the archive and to gain some historical perspective on Utah's newspapers. To increase the user-friendliness of the timeline, we implemented scrolling. Otherwise, it was simply too long to not overwhelm users. To see specific data, the user can hover over the line for the newspaper they are curious about. This results in the specific dates and the location of the paper being displayed as text next to the visualization.

The timeline could be improved by highlighting the line that users hover over. Additionally, displaying the newspaper information if they hover over the name of the paper would increase intuitive interactions. Colors could be improved to increase the experience of colorblind users.



Highlighted Newspaper: Beaver County News
Location: Milford
First date in UDN Archive: 1896-06-19
Last date in UDN Archive: 1990-07-26
Number of papers in UDN Archive: 150418

An example of highlighting a newspaper. The mouse was placed over the Beaver County News timeline bar.