CS 6630 Project Proposal

Basic Info.

The project title, your names, e-mail addresses, UIDs, a link to the project repository.

Project title: Exploring Utah's History Through Newspaper Visualizations

Names, Emails, & UIDs:

Jens Phanich, u0916721@umail.utah.edu, u0916721
John Bovard, u1278172@utah.edu, u1278172

Project Repository: https://github.com/dataviscourse2023/final-project-udn

Background and Motivation.

Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

Background: Our project will involve the analysis of a digitilized collection of Utah newspapers. This collection covers various newspapers from across the state many of which are not in publication anymore. This collection of newspapers spans from the mid 1800's up until present day. There are about 30 million items in this collection. This data offers a rich and diverse insight into Utah's history and also has many attributes that lend themselves well to visualization such as dates and geographic data.

Jens: My motivation for this research is that I interviewed with a job for the Marriot Library that supplies this data and they did not hire me, despite having worked there for three years and showing I have the necessary skills and qualifications, my motivation is revenge and to prove to the people that didnt hire me that made a mistake. I want to make such a good visualization of this data that people outside of the Marriot Library IT department use it, and then I pull it. I want them to feel regret.

John: I think this looks like an interesting data set to work with and visualizing it may yield interesting insights. I think working with such a large data set looks challenging and inherently offers learning opportunities that other areas may not. I am interested in history, culture, and language, and I am curious to see what trends the data reveals. I think that the data could be used alongside other resources to visualize interesting points in Utah's history and potentially uncover unexpected results.

Project Objectives.

Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

Question 1

How has Utah responded to certain major events? For instance, this data contains newspaper article titles, news paper articles titles can tell us a lot about the events going on at the time and the perception of said events. Also, many events happen concurrently. How are certain events prioitized attention over others? How does a paper in rural Utah report some issues over others? For instance, did Utah report on the Rwandan genocide at all? Did some areas report this issue more than others?

Benefits: Answering question 1 will provide users the ability to see what events were prioritized in Utah and where in Utah those events were reported. This could potentially be used to see how different communities in Utah may have been impacted by the events.

Question 2

When were newspapers formed and when did they stop existing? How long have various newspapers existed in Utah? What is the frequency of publication for various newspapers in Utah?

Benefits: Question 2 can benefit users by allowing them to visualize how long different newspapers have existed or did exist. This could potentially be indicative of things such as how many newspapers an area was able to financially support or how successful different regions were for newspapers. If combined with location data, it could also help users to see how many different editorial perspectives a location has had historically.

Question 3

What events happened in or affected Utah on a certain day? How did reporting of certain events change from geographic region to geographic region? For instance, a user could put in their birthday and visualize the things that happened on that day. They could also infer how important these events were in different regions of Utah if they were statewide or specific to one part of the state. For instance during the earthquake in Salt Lake in 2020, it is expected that this event would dominate content for Salt Lake Metro newspapers, but would this also hold true for a St. George-based newspaper?

Benefits: By visualizing newspaper headline keywords based on geographic location we can allow users to see if different regions of the state reported on different events or the same event. This can provide insight into the breadth of the impact that an event had within Utah.

What would we like to learn and accomplish.

We want to visualize this impactful data in a good way. We want to learn the various ways to visualize newspaper article titles and understand their advantages and disadvantages of. Additionally, we want to understand how to work with massive quantities of data in an efficient way and implementing scope effectively, there are approximately 30 million items in this data collection.

Data.

From where and how are you collecting your data? If appropriate, provide a link to your data sources.

We are collecting our data from the public Marriot Library Utah Digital Newspaper API. Here is the link to the Swagger page.
https://api.lib.utah.edu/docs/udn-v1.html

The data schema of a newspaper is shown below.

```
id:                 integer *
                    example: 17380735
                    primary key field
     paper:         string
                    example: Salt Lake Telegram
                    Newspaper name
     type:          string
                    example: issue
                    The document type
                    Enum:
                    Array [ 9 ]
     date:          string
                    example: 1948-02-04T00:00:00Z
                    Date the issue was published
     year:          string
                    example: 1948
                    Year the issue was published, in format YYYY.
     month:         string *
                    example: February
                    Full name of the month the issue was published.
     day:           string
                    example: 04
                    Day the issue was published, in format DD.
     thumbnail:     string
                    example: https://newspapers.lib.utah.edu/udn_thumbs/79/4a/794a01ee1cf6663c93fe189275fad07926d24330.jpg
                    Link to the thumbnail image
```

file: string

example: https://newspapers.lib.utah.edu/udn_files/e6/e8/e6e84caede46e7299101d171f49ebf7ffa441139.pdf

Link to the PDF file

page: string

example: 2

Page number from the particular issue

articleTitle: string

example: Ceylon Celebrates End of 300 Years under British Rule

Title of the document

ocr: string

example: Celebrates Ceylon End of Years Under B British tish Rule Rul COLOMBO Ceylon Feb b. b 4 oi U UP After of foreign domination domi n nation Uon Ceylon W the free ree e the first British crown crOn r j c col colony l. l 17 ony to attain full tull self self-gov self r T Temple mple bells belli belli heralded the advent ad ad- vent of at independence p a at mi midnight Later ter salvos of at guns bursting firecrackers fire fire-crackers and arid the shriek of at- sir of-sir sirens ens battered the silence of f the night is as U the e ep p people 0 of the island began eg ege celebrations e r tons that will last two I IThe weeks The day was observed as a day of thanksgiving with religious ceremonies in Buddhist temples tempes invoking blessing on Lanka the Lanka the name by which th the island land was waa known to the ol old Hindu poets and andst st still

Text representation of the scanned article, extracted using OCR

parent: integer

example: 17380679

Parent node id of the document object. If the document is an issue, this value is 0.

version: integer

example: 1586807012077863000

Version number of the document object.

Interesting and useful attributes of this data:

An attribute of note is that we can work with dates, as that allows us to use quantitative visualization techniques. For instance, we can see the publication rate of a paper with respect to time. We can also use the **Article Title** as that provides simple strings of text that encapsulate the content matter of a newspaper article. The **OCR** attribute while potentially descriptive and rich with content can also have false text transcription contained within, using this data requires clean-up (see below). Though might need to be used if no Article Title is present. Lastly, the **paper** attribute encodes with it geographical information, for instance, the Salt Lake Tribune is a newspaper from Salt Lake City. The majority of newspapers in this data set contain the name or region of the city that they are reporting from. This gives us a unique way to compare and contrast geographical regions of Utah

Data Processing.

Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?

There may be a substantial amount of data cleanup, depending on what we find when we begin using the data. Newspapers were scanned into the dataset using OCR (optical character recognition) technology, and these results may be error-prone and therefore require a good amount of cleaning.

The count of keywords found in newspaper article titles.

Data clustering based on keywords in different regions of the state as well as overall trends in Utah. This clustered data can be based on years months or any time interval that the user wishes to see. For instance, the user can have a data cluster of keywords for 1/06/2020 to 3/04/2020.

The API doesn't provide location data for newspapers, so this will need to be found for each newspaper (there are 387 newspapers). Many newspapers have a city name in their name, and for the purposes of this project we feel it is safe to assume that if a city name is in the newspaper's name, that is where it is published. This assumption will heavily reduce the amount of work needed to correlate a newspaper to a location. In addition, we will need to do an initial scrape of all the newspaper names and tie a geographical location to them as well as what town in Utah they are located in if this is not specified, we are thinking latitude and longitude since that is easy to find online. We will have a list implemented as a HashMap<String,Array<String>> that has this data for us. Generating this HashMap will be a one-time occurrence and it will be static. Below is an example of an entry. The reason we are storing latitude and longitude is because this allows us to impose a visualization on a digital space.
Key: "Salt Lake Tribune" Value: ["Salt Lake City", "40.7608° N, 111.8910° W"]

Technical Implementation of Data Processing.

Since the response from the API comes in a JSON format (as most APIs do), the tools needed to process the response come with vanilla JS, such as JSON.parse. We don't envision using any other tools outside of vanilla js to process and store this data.

In some cases we may have to resort to using the OCR if the Article Title is not present for data clustering, as mentioned above the OCR data is not always accurate or transcribed correctly, still even with this the OCR data has some correctness to it. If we do use the OCR data we plan on matching every word found with it and seeing if it is in a dictionary, if it is not we do not use that word for data processing, while far from ideal we envision that we can still visualize this data meaningfully.

Visualization Design.

How will you display your data? Provide some general ideas that you have for the visualization design. Develop ***three alternative prototype designs for your visualization****. Create* ***one final design that incorporates the best of your three designs****. Describe your designs and justify your choices of visual encodings. We recommend you use the Five Design Sheet Methodology.*

We will display our data using a combination of word bubbles, heatmaps, and timelines.

To represent the frequency of keywords in headlines for the entire state or a selected region, we will use a modified word cloud where keywords are placed inside of a circle with a size correlated to the frequency of that word in headlines across a user-selected date range.

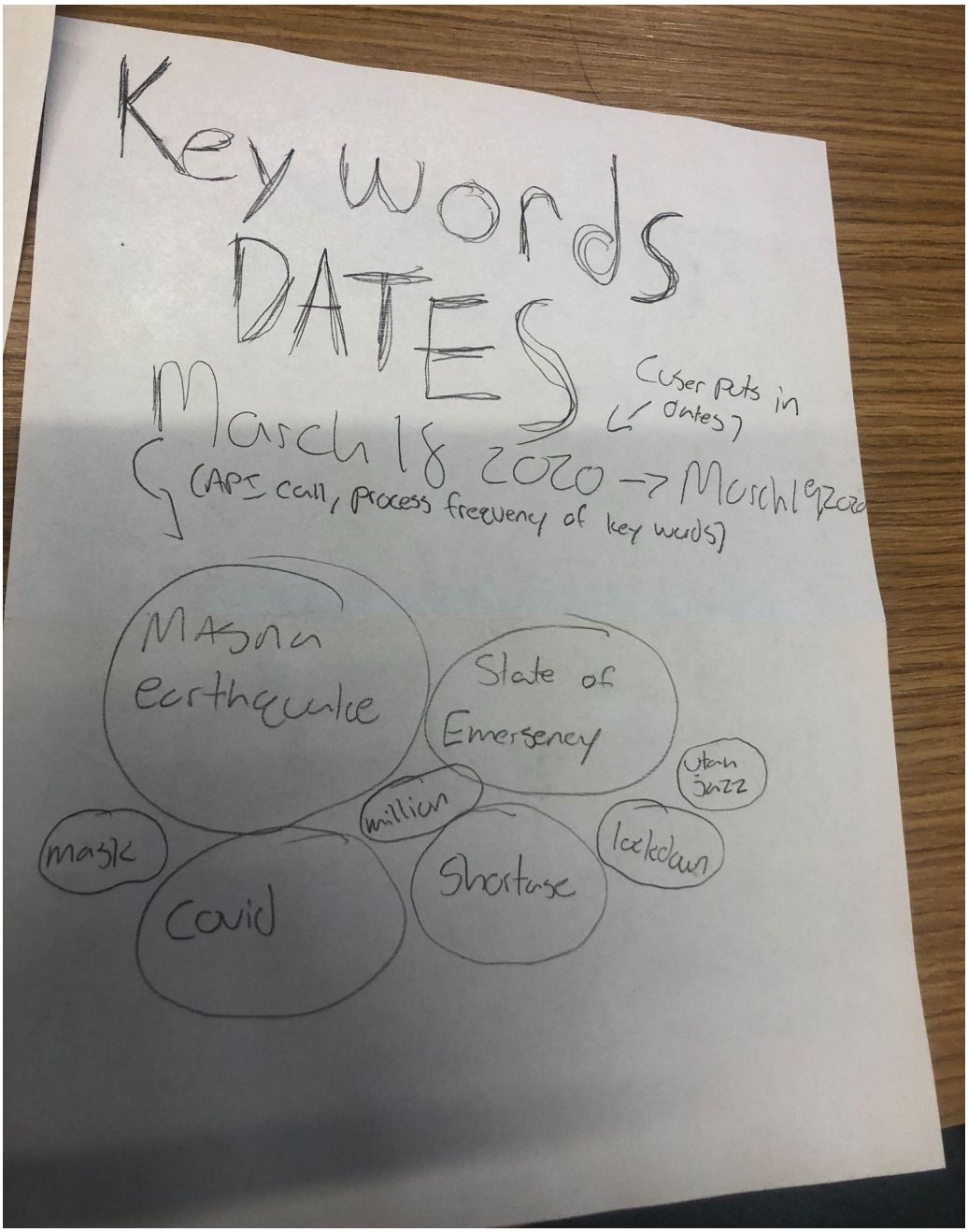
We also will visualize the most common keyword for different regions of Utah over a user-selected date range with the use of a map of Utah. This will be done by placing the most commonly occurring keyword in a bubble with a size relative to how frequently it was used in a bubble over that region on the map.

We will create a heatmap of Utah for user-selected keywords across a user selected date range. This may be animated over time to display how the use of that keyword changed. For example, the keyword “Pearl Harbor” would yield a heatmap gif that would be relatively quiet until 1941, in which the state of Utah would be red (or some other attribute to denote frequency/heat). Then as time moves on the redness of the map decreases. If a user would like to specify two words, there could be a two-color heatmap where if, for example, one word is more frequent, that area of the map is red, but if the other word is more frequent, it is blue.

Lastly, we will display a timeline of newspapers and their publication dates in Utah. This will have different newspapers listed on the y-axis with the x-axis representing time. Each newspaper will have a line depicting the time from their first publication found in the API to their last publication found in the API as a proxy for how long that newspaper has existed or did exist. We are also considering an additional element for this visualization where the user can select a newspaper from this visualization and a secondary chart will appear depicting the number of articles containing a user-selected keyword that paper has published over time. For example, if a user selects the Salt Lake Tribune from April 15, 2015 - April 14, 2016 with the keyword “Jazz”, the x-axis will represent time and the y-axis will represent the number of articles with “Jazz” in the title.

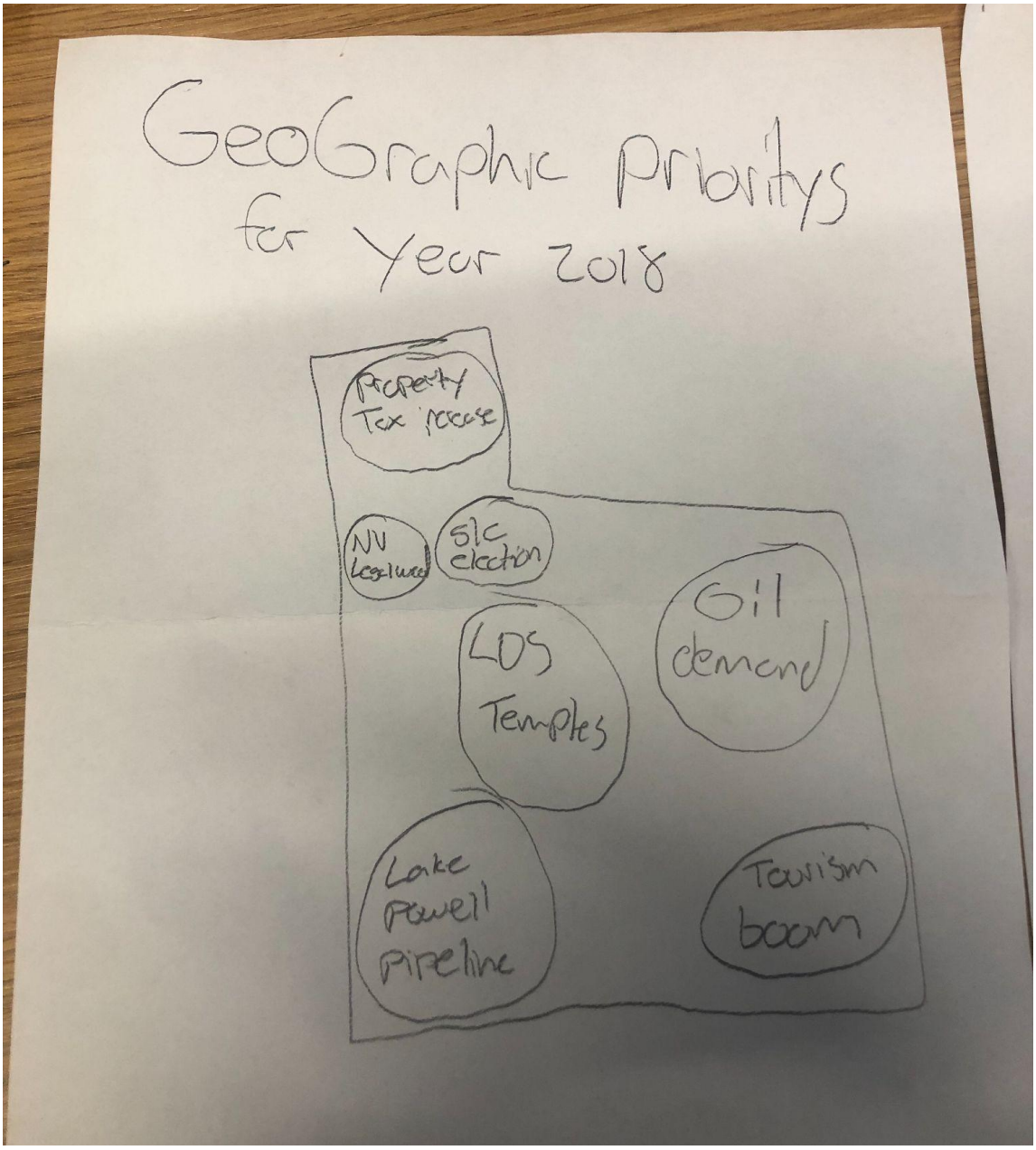
Three alternative prototype designs for your visualization.

Visualization Prototype 1



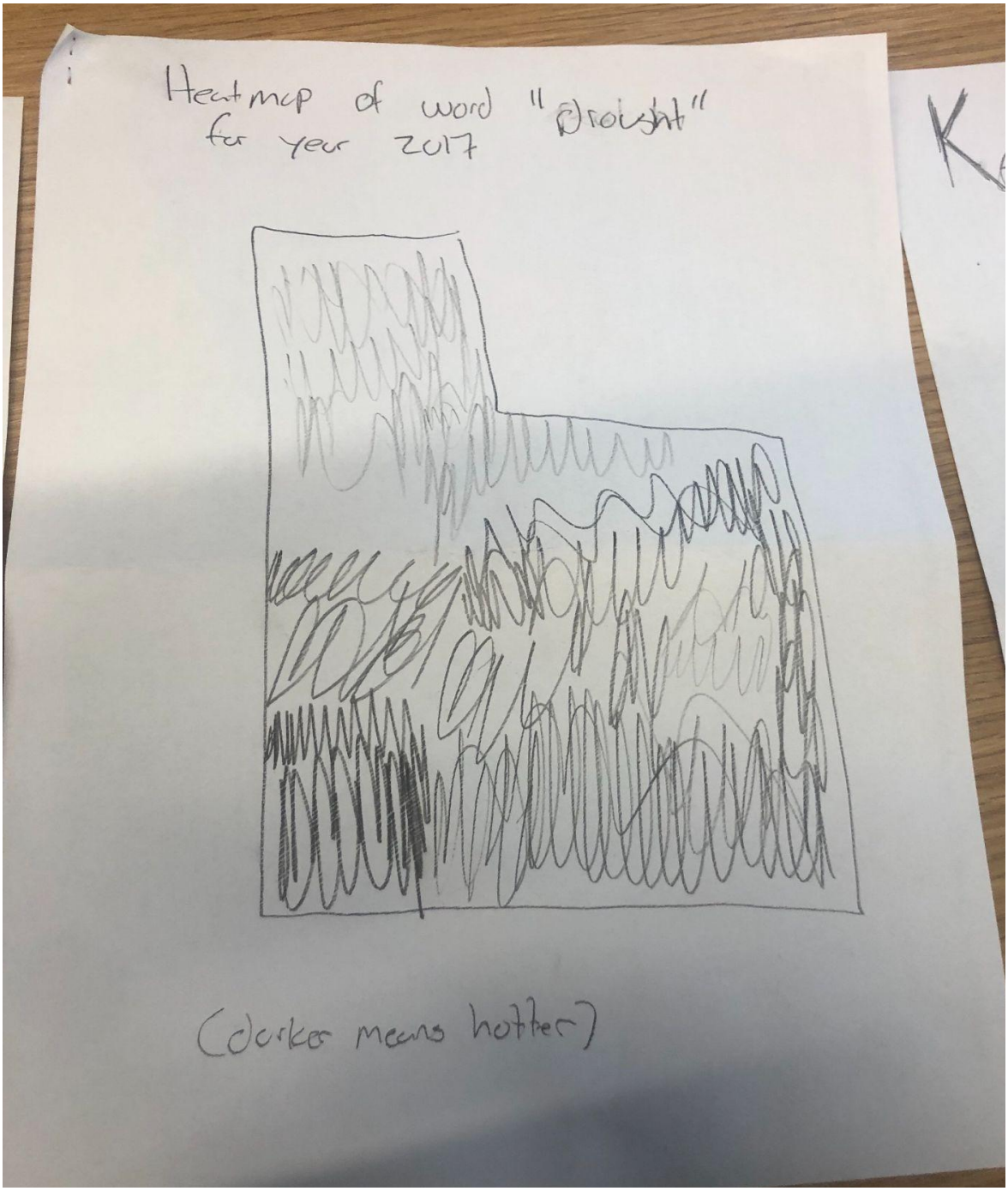
The design above shows a simple word bubble chart, the size of the bubble is correlated with the frequency of the occurrence of the word. The bubble chart is made for a certain period of time that the user can provide. In this example, we see that the Magna earthquake was the most reported on event as it is the largest bubble.

Visualization Prototype 2



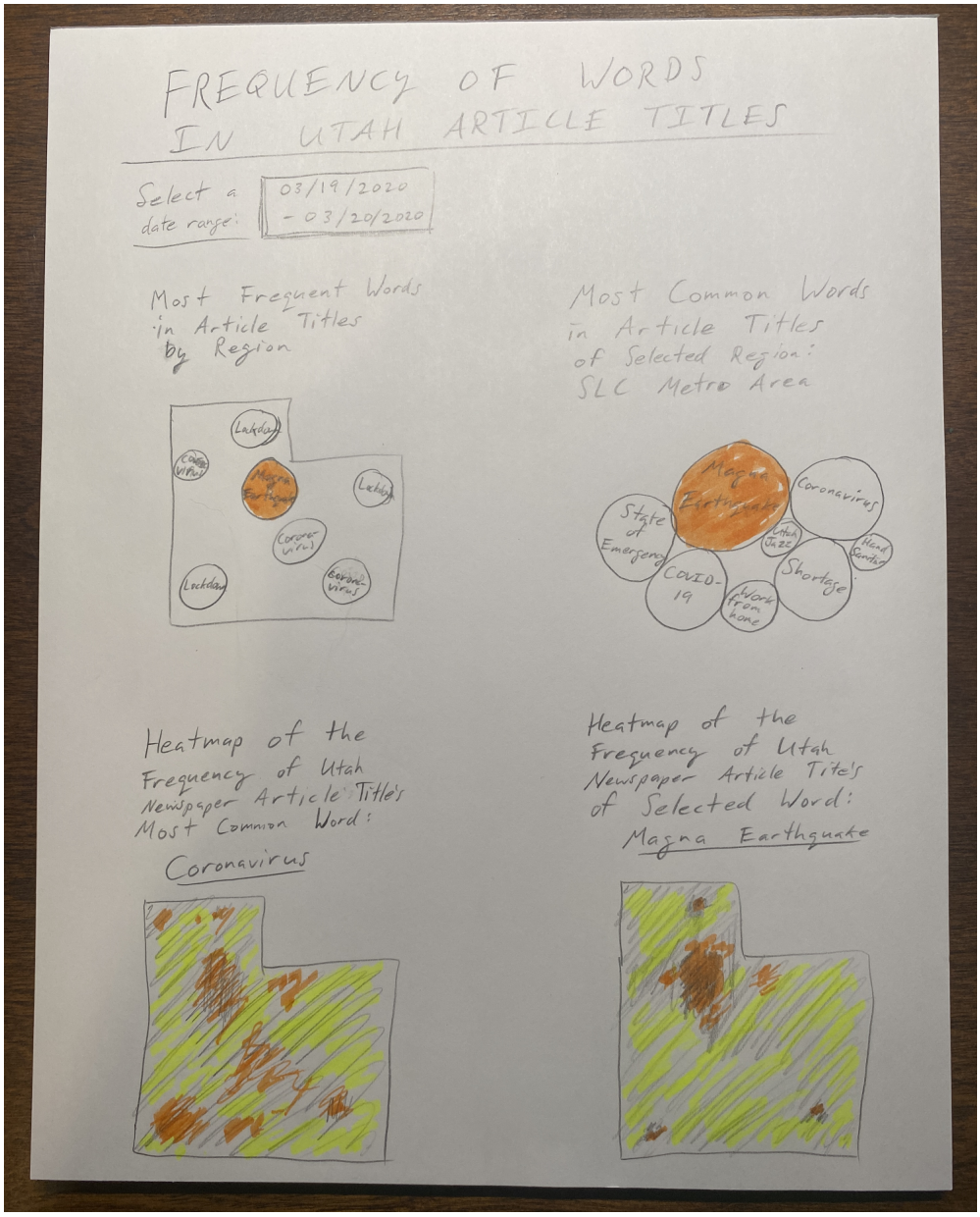
This prototype shows the biggest word bubbles associated with a certain geographical region in Utah for the year of 2018. The idea behind this prototype is to visualize each key trend or issue for each area of Utah. Notice how the Lake Powell pipeline bubble is present in the St. George area but not present in the Salt Lake City area. The purpose of this visualization is to highlight similarities and differences between regions of Utah using newspaper article titles.

Visualization Prototype 3



The visualization above shows a heatmap of a single key word for Utah in a year. The keyword in this case is drought. Notice how the southern portion of Utah is darker (hotter) than the northern region, this indicates that the keyword “drought” was included in more headlines in southern Utah. The purpose of this visualization is to see how different geographical regions of Utah and their media report and prioritize certain events.

One final design that incorporates the best of your three designs.



This visualization incorporates the best features of our prototypes into a more complete design. The user can see what the most common words in newspaper article titles are based on the region, and then select that region to see the most common words for that region. Below this, they are able to see two heatmaps: one for the most common word in the state, and one for a word they select from the region of their choice. This visualization lets the user compare how frequently words appear based on where in Utah a newspaper is. For instance, in the prototype, the words “Magna Earthquake” were mostly reported near to the SLC metro area, with some mentions in other population centers, whereas “coronavirus” was more evenly distributed.

Must-Have Features.

List the features without which you would consider your project to be a failure.

- Incorporating a map of Utah
- Comparing the frequency of keywords in headlines based on geographic location
- Showing frequency of words with respect to time for a certain paper or all papers as a whole
- Comparing and contrasting different papers and what they reported on at a certain time

Optional Features.

List the features which you consider to be nice to have, but not critical.

- Incorporating another data source like the US census, gives context to the population and demographics of a certain geographical region we are visualizing.
- Using OCR data would be nice but its inconsistency and lack of accuracy makes it hard to use. Therefore using an AI tool to clean up the OCR and generating some data from the outputted text would be nice, although this would come with its own issues(such as the accuracy of the AI tool, using a third-party API, like openAI, or even making or own AI language model)

Project Schedule.

Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

- Done by:
    - 9/22:
        - Set up data storage, and begin scraping API nightly. Scraping API data is only intended to be used as a backup if the provided API goes down as storing 30 million newspaper documents is unfeasible. However, a subset of that data could still be used to show the work we put in.
        - Mock-up visualization designs
    - 9/29:
        - Generate list/table that maps a paper to its geographical location(Lattitude and longitude)
        - Estimate the size of the data that needs to be scrapped. If necessary, set limits on goals for how much data we'll store
            - Some ideas for important data sets:
                - Time surrounding WWII
                - Time surrounding the fall of the Berlin Wall
                - (Look into important local events for time storing)
                - March 2020
                    - COVID pandemic announced and Magna earthquake in one week
    - 10/6:
        - Create a website with basic functionality
        - Design UI
    - 10/13: (none, **fall break**)
    - 10/20:
        - Create map
        - Create a prototype of at least one visualization from a small dataset
    - 10/27:
        - Tidy up prototype and try to get both visualizations working on with some basic data, not worrying about user interactions
        - Get ready for presenting for the Project Milestone the following week
    - 11/3: **(Project Milestone)**
        - Begin working on user interactions
    - 11/10:
        - Finish user interactivitions
        - Ask peers/friends for feedback on our design
    - 11/17:
        - Tidy up any loose ends (buffer time in case we need it)
    - 11/24: **(Thanksgiving week)**

        - Tidy up any loose ends (buffer time in case we need it)
    - 11/31:
        - *****Turn in finished project.*****