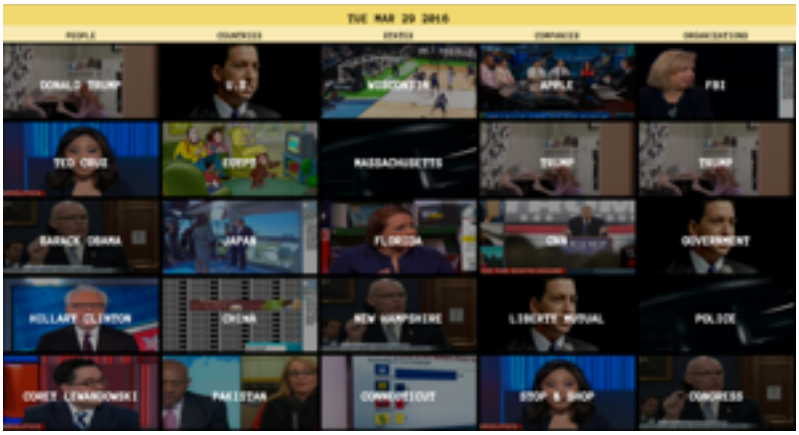
Jasmin Rubinovitz
Information Design Final Project
4/4/2016

1. **What is your concept and why are you interested in this data?**
The data I'm going to use is data scrapped from news videos and their closed captions, maybe together with data scraped from written news sources.
News-videos data:
http://super-glue-dashboard.herokuapp.com/

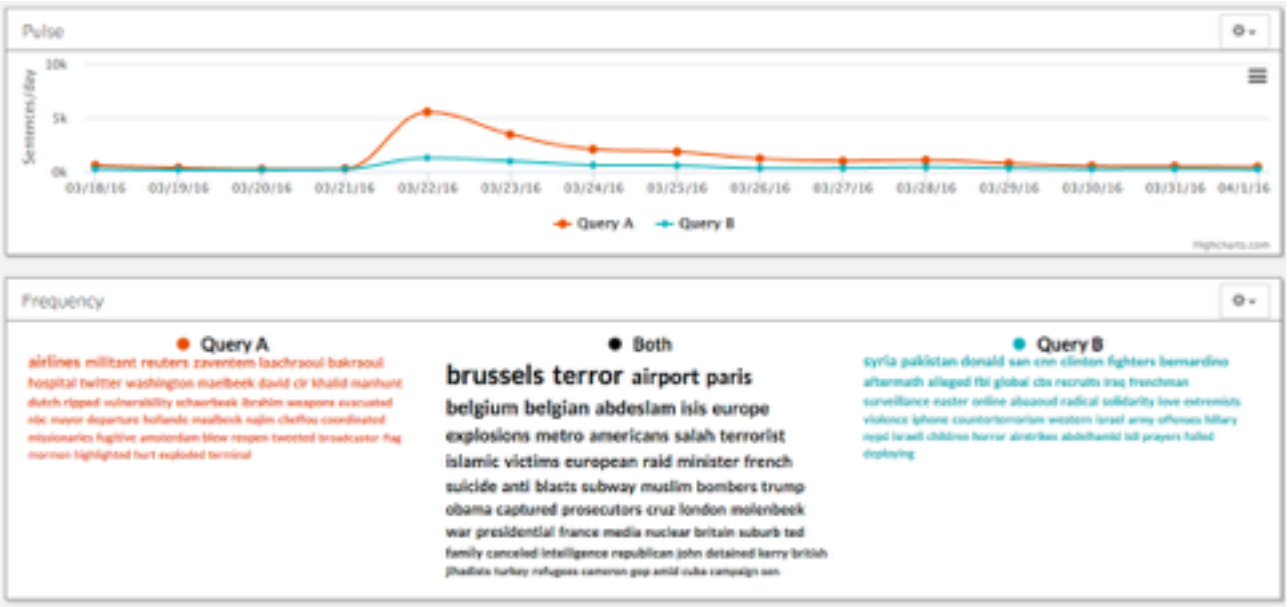


A visualisation of that data - http://wallofnow.um-dokku.media.mit.edu/
(better an a big screen, right-left arrows move between the dates, click 'c' to see the cursor). This matrix shows the top entities currently in the news under each subject (People, Countries, States, Companies, Organisations).

It does not tell the viewer the stories behind these entities, or the relations between them.

Written News data:
http://mediacloud.org/api/

image of the media-meter dashboard a tool using the media cloud api- comparing “Brussels” and “Terror”



My concept is visualising the news, what stories can we find and tell from the news data? how can we present them in an more manageable way?
How do you navigate between different stories and subjects? how do you show a connection.

I'm interested in this data because, as part of my research, I'm working on the code to collect this data (the videos data). I'm very interested in the different interaction people can have with the news, and what is going on in our world.

2. What questions do you have about your data?
I want to learn if we can find interesting patterns in the top news, what connections exist between the different data and how do they look in the data (peeks in mentioned / subjects mentioned together/ ?)
For example, Apple vs. FBI, there are a lot of news about apple and a lot of news about FBI, almost every day. At some point in time, the subjects start to appear together. in the same stories, and in the same sentences. what is the story behind that day? how can we better tell it?
Another example is Brussels and Terror (interestingly, Both terms appeared a lot together after the Paris attack as well) I'm sure we can find many more examples.

3. What audience(s) do you want to communicate your answer to?
My audience are people who are interested in the news.

4. What are 2-3 hypothetical answers you might find as you explore?
Many entities that appear in the top news are related.
We can find other terms related to the same story quiet easily and then recognise the story itself.
Some connections are surprising, or temporary (Trump saying something about Apple vs. FBI is only a small temporary connection of "Trump" to the story)

5. What types of context do you see this final form living in?
I currently see this final as an interactive mobile/desktop app (not yet sure if both or not) but maybe this will change.

Technical questions:

1. What is your data?
My data are video parts, with date and time of broadcasting, channel, closed captions text, and NLP analysis on closed caption which include entities and keywords that were found in the closed captions text.
I might also use more data, scraped written news from main U.S media sources.

2. What is the file size?
Both data sources are very big, but have an API.

3. What is the file format?
The news videos data is stored in a mongoDB database, and also has a partial http API. since I mange this database, I can add more functionality to the API and just use http request.
I'm still debating whether to work with Node.js and a mongoDB client to query the data, or add functionality to the API and use http requests.

4. What is the file shape?
MongoDB documents (which are BSON objects, which are just like JSON objects)
Each video contains many fields: Id, media url, subtitles url, thumbnails, times of scene changes, closed caption divided by time, more..

5. What will you look for?
Popular keywords from the latest news, popular keywords mentioned in the same story, interesting dates for events.

6. What is in the data that can help you find what you're looking for? Is it implicit or explicit?
I have all the keywords and dates in the data. It's not explicit but it can be found using simple queries.