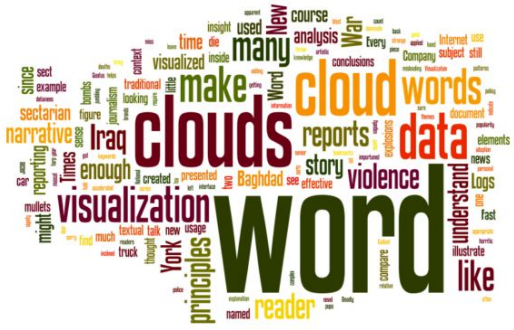
General purpose:
The goal of our project is to analyze how different media sources prioritize and cover different news topics on their websites. In particular, the expected output is interactive and two-fold. First, we will use word clouds to visualize the high-frequency words appearing on the main page of different websites and across all websites. The user will be able to select news sources and examine these most pertinent words in each source individually and aggregated. The second output is to identify which countries are most frequently identified by these news sources cumulatively. We will attempt to use Python GIS to visualize this data and to scrape data over multiple time periods (e.g. the news today vs. news tomorrow).

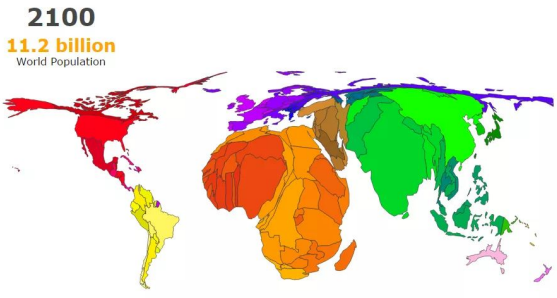
Raw data sources:
We will use three to five news websites to scrape raw text data from news articles. These could be, for example, CNN, BBC, South China Morning Post, and Russia Today. In particular, we will scrape the titles and short texts of news articles from the main pages of such websites. One complication might be that websites change their html from time to time. One challenge here will be to find out how to "blanket scrape" all the pertinent text on a given website (e.g. searching within all divs, h2, h3, p etc. for text instead of depending on a specific page's HTML structure). Another challenge will be scraping and retaining information for e.g. the past 3 days' news websites.

Data processing:
After scraping text data, we aim to process it with SQL and/or Pandas. For example, we will find the most pertinent words around a main page by setting a cutoff for relative frequency of the word in the webpage and filtering for this "relevance index." We will also use SQL to handle user requests (e.g. I want the top 5 most relevant words across all news sources, or in one news source).

Visualization:
As stated in the general purpose, we will visualize relevant words for both outputs by creating word clouds for each news source. We have verified that this visualization method is feasible. There are several resources like this one for us to self-learn new python packages and implement word clouds.



We will visualize which countries are talked about most in the news using Python GIS. The number of times a country name appears in the news source will be



proportional to the size of that country on a map (see a cartogram).

Interactive component:
The interactive component of our project entails a user-selected filter for news sources, most mentioned words, least mentioned words etc. and generating word clouds/cartograms to visualize the data based on word input. A tentative interactive feature would be the user dragging an axis of time and the cartograms of countries mentioned changes along with time.

We will use Django to build a website to host these interactive features.

Work to Complete:
- Design algorithm to scrape headlines and titles from different news sites
- Create visualization and metric for which headlines are displayed for each site
- Build a website that allows for user interaction and input
    - User is able to input desired keywords or news site
    - Must be able to scrape data multiple times per day without oversight
    - Store data over one day and replaces old data
    - Compare news stories across specific site
    - Make word cloud visualizations for each search

Timeline:
- Week 5:
    - Scrape data from different sources, designing code specific to each individual site
- Week 6:
    - Complete scraping code, begin designing metric for determining important portions of articles
- Week 7:
    - Integrate data for important words with visualization technique and create word clouds
- Week 8:
    - Begin building website, determine how to scrape in real-time given a proper interval
- Week 9:
    - Allow website to build interactive visualizations after scraping
- Week 10:
    - Finishing touches and present project