CS316 Project Final Report

Anh Trinh, Jerry Chia-Rui Chang, Srikar Pyda, Wendy Lu

December 15, 2016

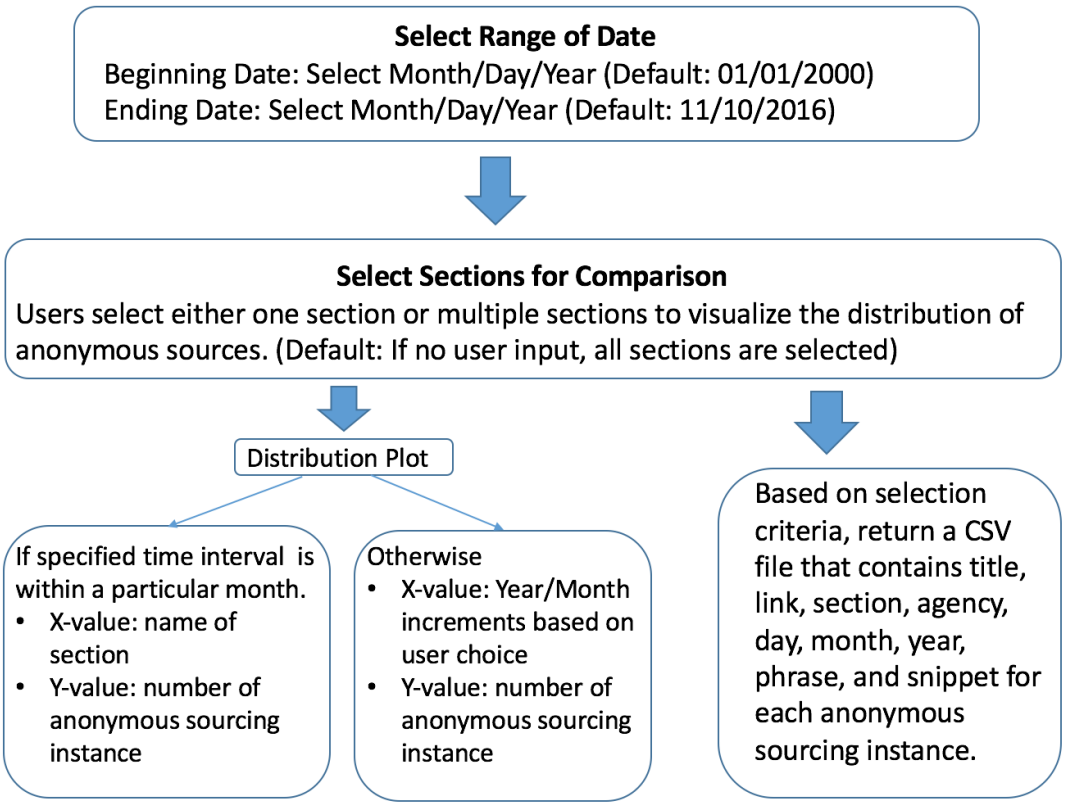
1 App Description

1.1 App Motivation and Function

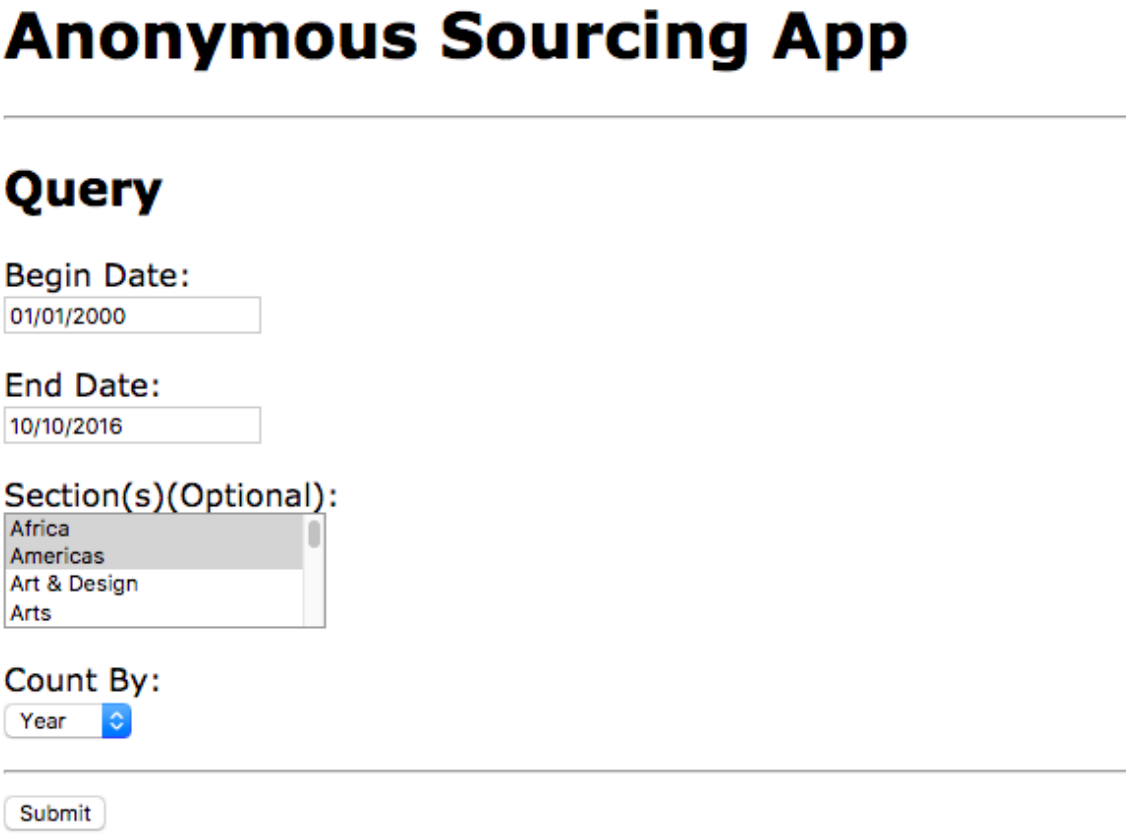
As described in previous reports, our app allows users to track the change in anonymous source usage in New York Times from 2000 to 2016. This idea comes from our belief that observing the use of anonymous sources over time can provide useful insights about potential bias and false information in modern journalistic practices.

The application users will be able to see the number of anonymous source count in month or year increments (depending on the scale specified by users in the initial user interface) by news section (U.S., Global Business, Sports, International Opinions, etc). In the initial interface, users can select the time period (by specifying starting date and ending date) and news sections they are interested in (multiple selection supported, default to all sections selected if no user input). On query submission, the application visualizes the distribution of anonymous source count with a bar graph if the specified dates are within a particular month, otherwise a line plot is returned. In addition, a CSV file of anonymous sourcing details returned from query is available to view and download.

The flow chart below summarizes the app functions and set up. More set up details with Python/Flask are in the README.



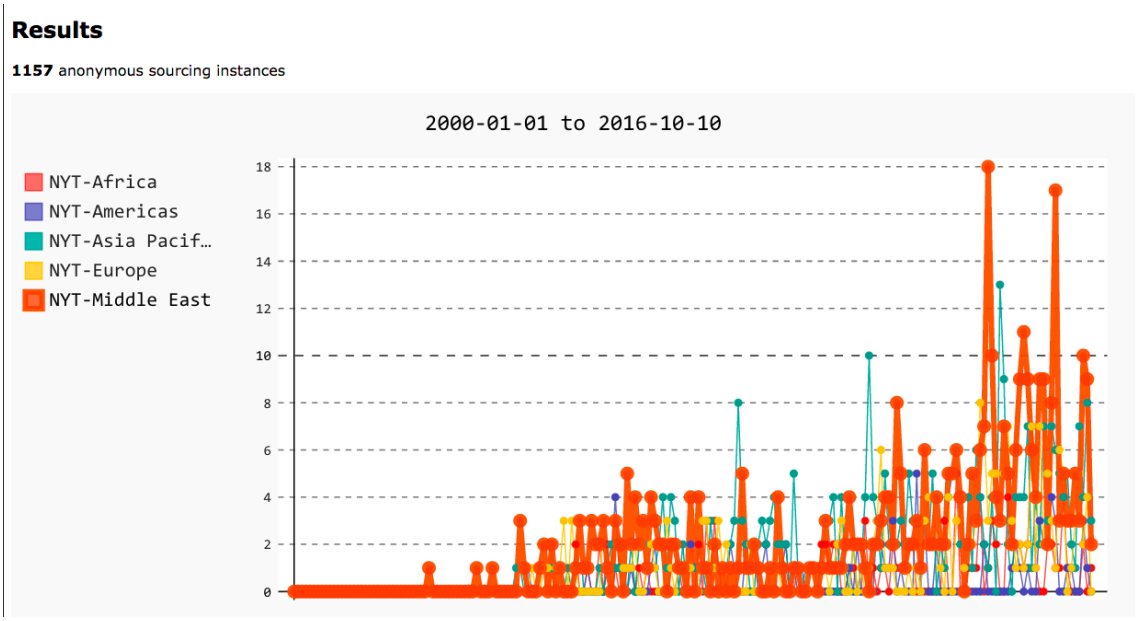
Screenshot below describes the user interface. (Here we used the default values for Begin and End Date and selected several sections that correspond to different geographical regions)

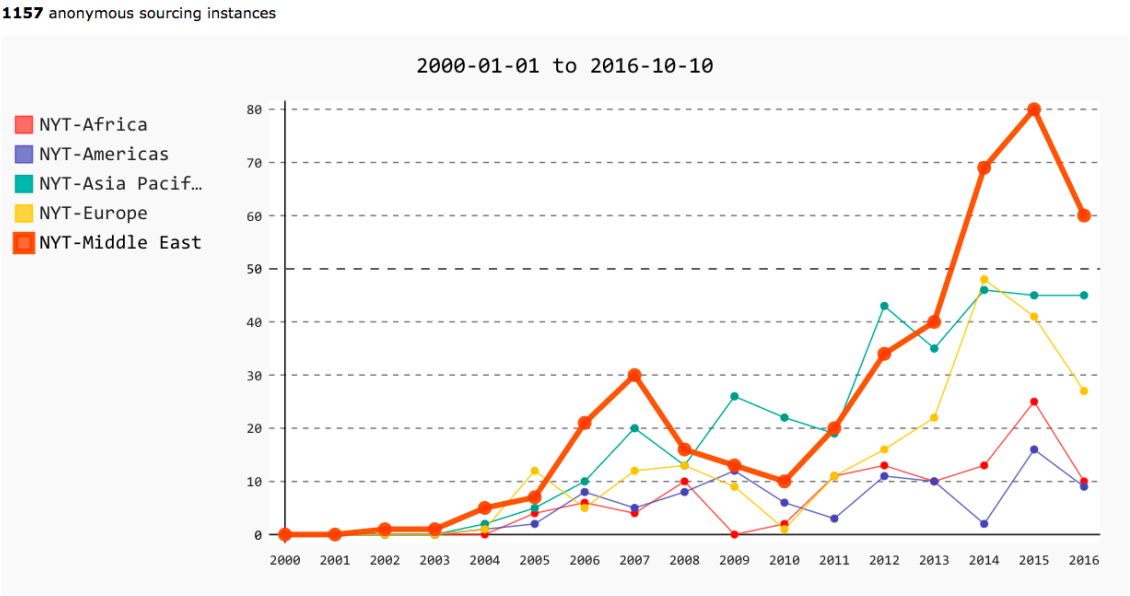


1.2 Data Analysis

As Prof. Bill Adair predicted, there does not seem to be an obvious trend in how the number of anonymous sourcing differs throughout one particular year, regardless of whether it is an election year or not. Since there were several major incidents with anonymous sourcing at major news outlets over the past few decades and those incidents resulted in increasingly stringent rules against the use of anonymous sources, the interesting story might lie within a longitudinal study after the year 2000. It seems that for several sections (especially international news sections as shown below) there has been a rapid increase of anonymous source use in recent years (2016 data is still incomplete) following a drop around 2008. Therefore, we are speculating that there might be a major incident with fake anonymous sources happened around the year 2008.

The graphs below are examples of how the number of anonymous sourcing changed for the sections corresponding to different geographical areas in New York Times (first peak near 2007 with a drop that followed, and a rapid increase in recent years.) The first graph is scaled by month and the second one is the same results scaled by year. As we can see, there are a lot more anonymous sources being used for the Middle East and Asia Pacific sections, while the Americas section has relatively low usage throughout the 16 years. This corresponds to our expectation since sources are more inclined to remain anonymous in politically unstable regions. Additionally, many news articles regarding American politics or economics are published under the "U.S" or "Economics" sections, while news articles about other regions are mainly listed on their respective geographical section.





1.3 Survey of Related Work

From our previous research and discussion with Prof. Adair, we learned that there was very limited data regarding the changes of anonymous source usage. We mainly referenced the Anonymous Source Tracker developed by Mark Schaver in our project.

More specifically, Mark Schaver's Anonymous Source Tracker application tracks when news organization utilize anonymous sources within an article. Looking for particular key phrases, Schaver identifies articles citing unnamed sources along with its date of publication and news organization. All of this information will be important in our implementation of anonymous source data-visualization within our application.

Although Schaver is able to identify anonymous sourcing based on certain key-phrases, his data spans only three months. He did not visualize the data retrieved (he simply listed the articles and the original quote by publishing time). Going beyond his application, our project analyzes anonymous sourcing over a much longer period time and provides meaningful visualizations.

2 In-depth Discussion of System

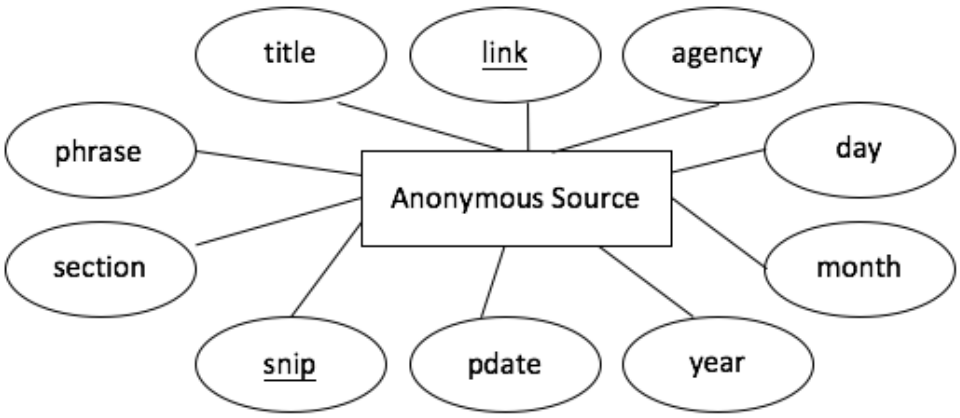
2.1 Database tables and Description

Bigtable(title, link, agency, day, month, year, pdate, snip, section, phrase)

- Title: Title of the article
- Link: link to the article
- Agency: The agency article belongs too (e.g. NYT, Washington Post)
- pdate: Date the article is published
- snip: Combination of all anonymous phrases and the date being published
- section: The section in which the article belongs to (e.g. politics, finance, sports)
- phrase: Anonymous phrase in the article

2.2 E/R Diagram

The E/R diagram for our database design is as follows:



2.3 Assumptions

We made several assumptions in our analysis and app setup mainly due to missing data and our imperfect method of data pulling and data cleaning. We pulled data (5907 rows) by filtering a list of keywords used in anonymous sources provided by Prof. Adair, namely retrieving and parsing the data by customized search through the New York Times (for those keywords) using Google Custom Search API. We cleaned the data by categorizing missing data (there are 159 out of 5907 articles whose snip and original link does not account for the date and its news section) to the "uncategorized" news section, and since they are missing dates, they are not accounted for in the number count.

- We assumed that the keywords provided to us are indicative of actual instances of anonymous sourcing (to validate anonymity requires actually reading and assessing the paragraph containing those key words in the context of the article).
- We assumed that Google Custom Search API did not miss any articles containing the keywords we searched for from 2000 until 2016

- We assumed that the general trend of all anonymous sources would follow the trend of the anonymous sources we pulled (we only pulled part of all the anonymous sources).
- We assumed that the missing data (159 out of 5907) would not be highly disruptive to the trend we found.

2.4 Design-choices

- We write Python program that connects to and retrieve data from Google Custom Search API
- We chooses Flask-Python framework to create our application. Web pages are rendered using WTForms and Jinja templates. Mappings to our database are processed with SQLAlchemy. For the graphical visualization, we use pygal visualization Python package.

3 Detailed Description of New Approaches or Algorithms

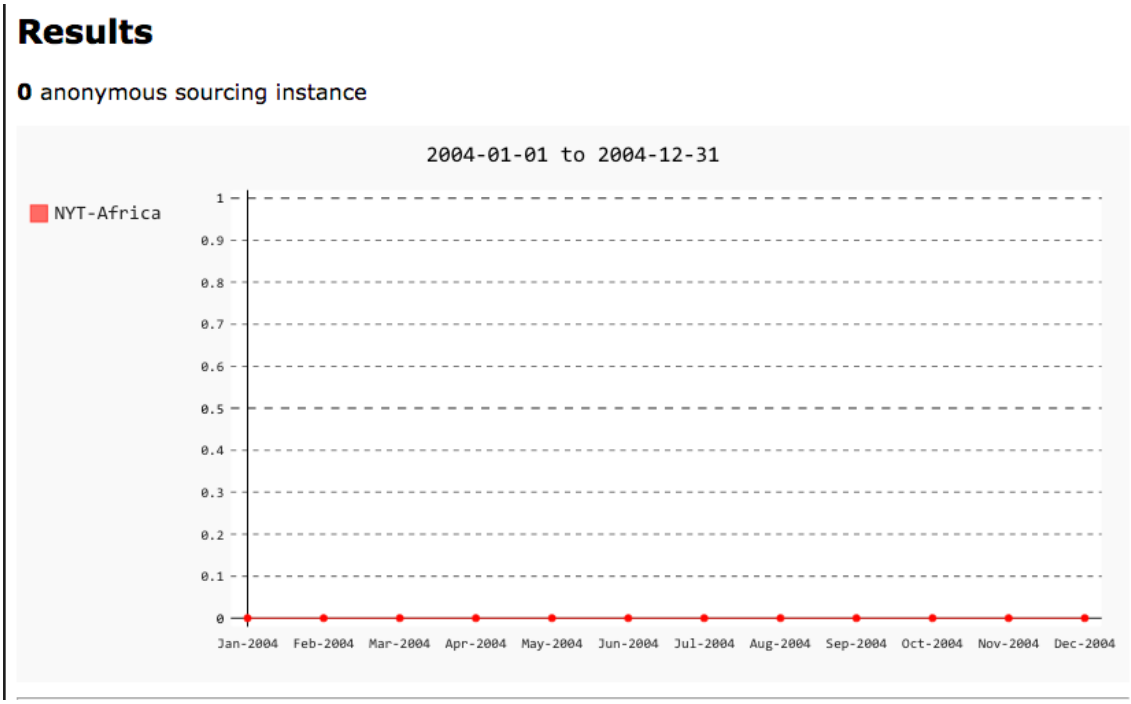
Not Applicable.

4 Evaluation of System

4.1 Compare Online Platform Results with SQL Results

Our system works relatively well: when we provide input through online form submission, the results we get are consistent with the results we get from retrieving the same data from our database through SQL queries.

As an example, we retrieved the number of anonymous sources in the "Africa" section during 2004-12-01 to 2004-12-31, and got count 0 in both cases (as shown below).



```
anons=# SELECT COUNT(*) FROM anon WHERE pdate <= '2004-12-31' AND pdate >= '20
04-01-01' AND section = 'Africa'
;
 count
-------
     0
(1 row)
```

4.2 Open Issue

We want to allow users to be able to download a csv file with the query output after they submit the web form. However, we cannot achieve this since this requires us to write a file in the VM, yet we do not have the permission.

5 Open Discussions for Future Work

1. Currently, our application only analyzes data from NYT. In the future, we hope to incorporate more media agencies (e.g. Fox news) for comparison.
2. We hope to develop machine learning algorithm that will automatically identify anonymous sourcing instead of one that depends on fixed number of keywords.
3. Strengthen our data through users' rating input (allows users to validate if the query results are actual anonymous sourcing instances)
4. Improve the atheistic layout of the application

References

- Project Group Github source code https://github.com/anhntrinh/anonymousSourcing
- Schaver's Anonymous sourcing Tracker source code http://schaver.com/anonymous/
- Algorithm used https://github.com/markschaver/anonymous
- This final report can be accessed online at https://www.overleaf.com/read/wjftwzxhnpyw