Identifying and Visualizing Historic News Trends

Georgia Institute of Technology

Sachin T Sany, Rajdeep Singh, Akshay JD, Aditya Vadhavkar, Pradyumna Mukunda

Summary

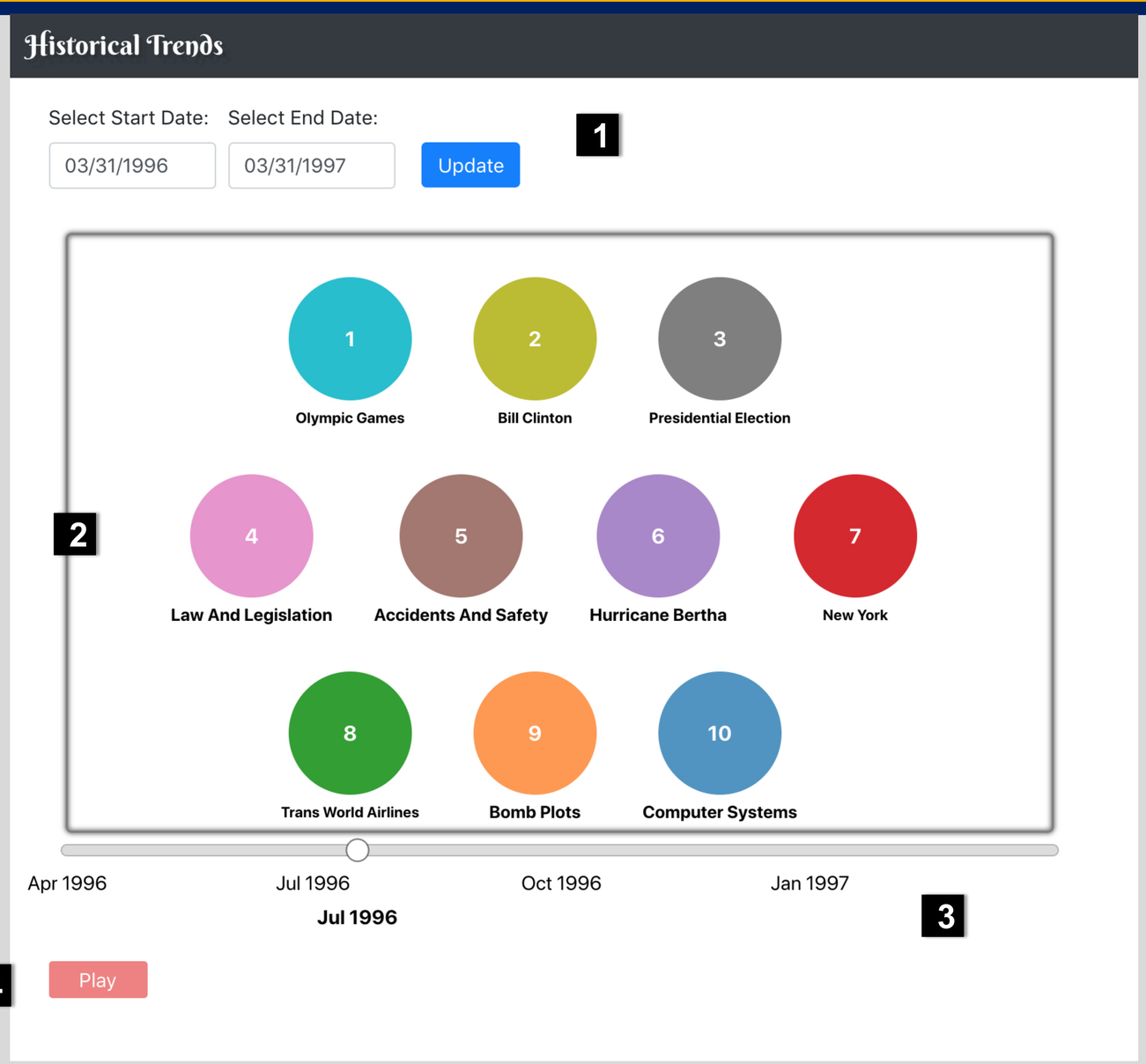
The goal of our project is to identify and visualize historical news trends for the past 100 years within a time period selected by the user. Anyone interested in learning about historical events would also be interested in a one stop destination for visualizing historical news trends. Unfortunately, popular news services like Google News Trends lists the top news stories only for a limited time period. We intend to bridge in this limitation by using the New York Times news archives containing articles dating back to the early 20th century.

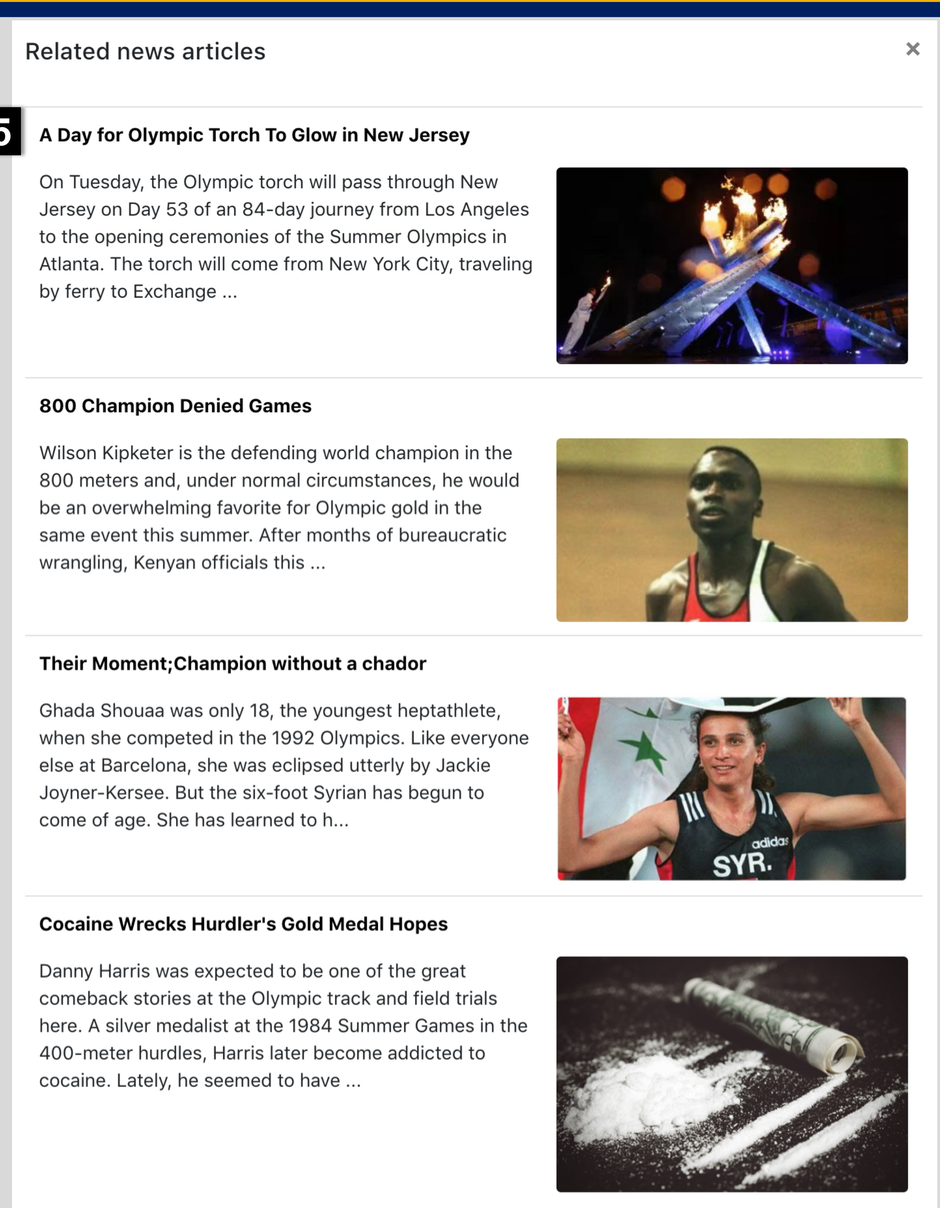
Visualize trending topics

Our application is a one stop destination for visualizing news articles related to a trending topic. The user can select the time period over which he/she wants to explore the news trends. When the user selects the time frame and starts the visualization, a transition of the trending topics is displayed from the start period to the end period in steps of one month each. The user can pause the visualization at any point of time and adjust the slider manually to skip to a different time frame.

Identify Related News

The user can pause the visualization and click on a topic to view all the new events related to the selected topic in the specified time period. We cap the number of related news articles displayed to 10. The user can browse through brief snippets on the articles including the headlines and images and can view more details on the news topic by clicking on the hyperlinks that takes them to the New York Times news archives. The user can switch back to the time series of trending topics by closing the modal and continue the visualization.





The user has opted to view the news trends in the time range from March 1996 to March 1997. 1) User can select the start and end dates for viewing the trending news topics. 2) The trending topics get displayed on the canvas, each with a different color. Each topic is ranked from 1 to 10 in the order of their popularity in the given time period. 3) The slider gets updated in steps of 1 month as the time moves forward. 4) Hitting the play button starts the video displaying the trending topics in the selected time range. 5) When a bubble representing the topic is clicked, a modal pops up displaying the news articles related to the topic. The modal above shows news articles related to 1996 Olympic Games.

Algorithm

We used Named Entity Recognition for grouping together entities having multiple words, to facilitate weight assignment to entities. We then ran TF-IDF Based Clustering for identifying relevant keywords in the document, taking into consideration only those topics with a TFIDF score of above 10. We aggregated the popular keywords for each year and month. A batch job transformed the output from the Apache Spark instance into key value pairs and stored it in a Redis instance. This information was retrieved by the Node.js based web application framework for rendering the visualization.

Evaluation

Top 20 keywords for some selected months in the period of 2014-2017 were compared with event data from the Wikipedia page entries for those years. (e.g. “2017 in the United States”). The accuracy was calculated as the percentage of keywords in our top 20 results that have a matching event on the Wikipedia page. We got an accuracy of 65-80% in our evaluation.

Factors limiting the accuracy of our system:

- Articles related to politics dominate New York Times, so smaller categories like entertainment and sports take a backseat.
- New York Times publishes more articles local to New York City which is not reflected in Wikipedia
- Repeated appearance of generic keywords like “Books and Literature”