CS 410 PROJECT REPORT

News Article Categorization

Team Members: Himay Jesal Desai, Bharat Thatavarti, Aditi Satish Mhapsekar

Overview:

Our project, News Explorer, is a system that categorizes news articles and helps visualize them based on the entities location, organization and person. In addition it also lets one classify articles based on the category they may correspond to - criminal articles, political articles etc. We have observed that such a kind of tool does not exist - at least not in its entirety. Looking for similar systems revealed that not a lot of similar work has been carried out - neither in terms of categorizing news articles extensively nor developing a user friendly visualization tool. The tool would be beneficial for newspaper owners and journalists to analyze as to how many articles were published in a particular category. The idea is also to allow the tool to assist in more sophisticated analysis - for instance determining the count of a particular crime by location. It is a given that anyone interested in a particular category of news articles stands to benefit as well - (s)he can browse through news articles that interests them. In addition, there also are sites like EMM News Explorer, but as pointed out previously they don’t do a very good job of categorizing the news articles and also do not have an easy to use/appealing UI. The project primarily had three challenges - the first was to categorize news articles successfully. The second was to deal with the enormous amount of data and to set up an infrastructure that supports this - the response time of the web portal needs to be minimal, failing which irrespective of how successful the classification algorithm is, the application will not be a success. The third was to develop a visually appealing UI that captures the functionality expected by the end user.

Motivation:

Text categorization or classification, is a way of assigning documents to one or more predefined categories. This helps the users to look for information faster by searching only in the categories they want to, rather than searching the entire information space. The importance of classifying text becomes even more apparent, when the information is too big in terms of volume. Google directory is one such web classification system. But these systems are generally taken care of by human experts and they do not scale up well due to tremendous growth of web pages on the internet. To make this system automatic, classification methods based on machine learning have been introduced. In these techniques, classifiers are built (or trained) with a set of training documents. The trained classifiers are then used to assign documents to their suitable categories. In our project, we have adopted a similar approach. Amongst the vast information available on the web, we chose the domain of news because we observed that the current news websites do not provide efficient search functionality based on specific categories and do not support any kind of visualization to analyze or interpret statistics and trends. The fact that news data is published and referenced on a frequent basis makes the problem even more relevant. Currently, Yahoo News and Google News provide some form of categorization, but do not support much visualization due to which journalists and naive users cannot analyze the news data and obtain significant insights. This motivated us to build a system keeping two types of users in mind, the first user is the news reader who is interested in browsing news articles based on category and the other is the stakeholder or analyst who is interested in analyzing the statistics to identify past and present patterns in news data.

Problem:

The problem we are trying to address is the gap between news data and its users – the readers and the analysts. To do so, we had to zero in on the shortcomings in current news portals and address them subsequently. We identified three areas of significance (which according to us have not been addressed adequately by the existing portals) – first, identifying the category that a news article corresponds to. Typically, a user would want to browse through a particular domain (category of articles) and it is therefore imperative that this functionality be exposed; The second essential requirement is to visualise news article data in a manner that facilitates extraction of useful information (for analysts etc). The last and perhaps the most important one was to provide an easy to use UI. We decided to improve on these areas in our system. We were provided with raw news data set for the years 2005 - 2013 from students at the University of California, Los Angeles. Below is a description of the milestones :

Milestone 1: Get data ready and make it parseable.
Milestone 2: Parse the whole huge data set from 2005-2013 and store it in database.
Milestone 3: Categorize the dataset in terms of various entities.
Milestone 4: Build a prototype and experiment it.
Milestone 5: Build a visualization module on top of it.
Milestone 6: Make the visualization interactive and add search functionality.
Milestone 7: Document the whole project.

Solution:

The solution, a news portal called News Explorer, was formulated such that it addressed each of the milestones. First, we parsed the whole data set with our own custom parser and stored meaningful information in the form of meta-data in a MySQL database. As the volume of the data increases, the idea is to move to a distributed datastore for fast read and writes - at this point, for news data spanning over 8 years, MySQL proved reasonably fast, providing response times in the order of milliseconds. Next, we used classification algorithms - kNN, Naive Bayes and stochastic gradient descent to classify news articles and tag each article with its category. Each of them proved pretty successful in classifying documents successfully. Finally, we built a visualization module on top of the backend setup, which can be used to analyze and obtain insights regarding news data pertaining to specific domains. For instance, we have incorporated a visualization, where if you select a person's name, you can view all articles related to the person and a chart showing the popularity of these articles, depending on the number of clicks/views by the users. Similarly, as mentioned previously this has been incorporated for other entities like organization, location and type of article. Moreover, one can also select a combination of these entities and view the corresponding data/visualization as well. Another interesting visualization module is the one that displays the count of articles on the world map, where users can analyze as to how many articles correspond to different sections of the world, based on the selection of the entities. For instance, one can view the map in such a way that only the distribution of criminal articles is displayed and can make decisions as to which country has the most crimes and if that trend has changed over time. In addition, we provide the word cloud feature, where one can view word clouds based on the entities provided. For future work, we plan to generate word clouds for various time lines (an idea influenced from a paper by Microsoft Research) where different world clouds are created for a range of year/month selection and then generate a line graph or scatter graph to show trends over the years, which can be useful for some applications.

System Architecture:

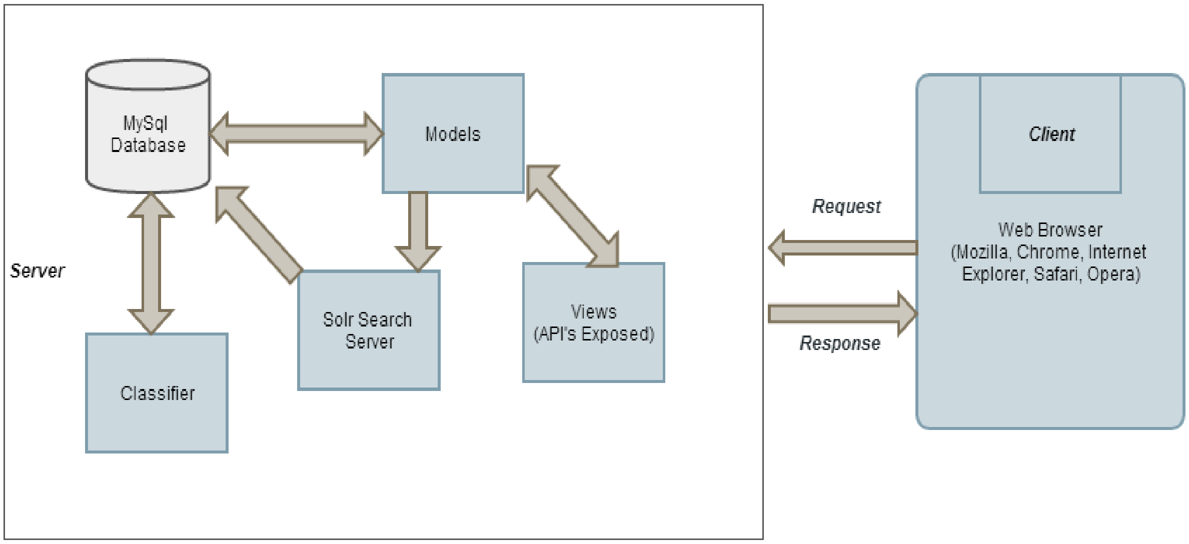


Figure: High level Design of News Article Categorization

The high level design of the system as shown in the figure above is based on the client-server architecture.

- The client is a web browser which makes HTTP requests to the server and gets the desired response. It is basically the presentation layer of the system and corresponds to the user interface.

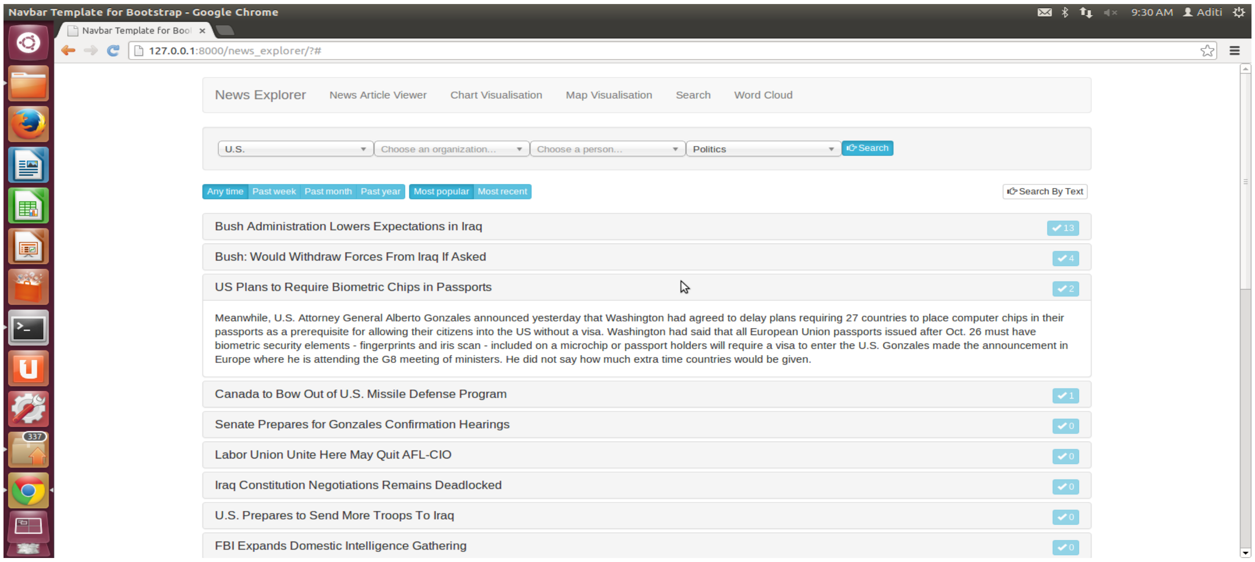
- The server consists of five main components: MySQL database, models, views, classifier and Solr search server. The last two components implement the important functionality of classification and search.
  - The MySQL database stores the entities in the parsed news data in a normalized manner - it also captures the relationships between entities.
  - The models form the data access layer which describe how the data is characterized, how to go about accessing the data, the properties it captures and the behavior that it exhibits. Essentially every model corresponds to a table in the relational database.
  - The views contain the logic to access the models and return a response depending on the client request/user input. Thus, the business logic of what data to display and how to display it is defined in the views.
  - The classifier is initially trained using a hand tagged training data set of some news articles. Once trained, it runs in the background and classifies news articles into different categories such as politics, natural disasters, etc. based on their content and records the category in the database. We have used a stochastic gradient descent classifier (after experimenting with several classification algorithms). To capture the similarity between the articles, the TF-IDF similarity measure has been used.
  - The final component is the Solr search server from the Apache Lucene project. It supports a REST-like API which enables fast and scalable full-text search on the news articles.

Working of the System:

This section provides a walk-through of the system prototype along with the corresponding screenshots. There are mainly five components:

- News Article Viewer
- Chart Visualization
- Map Visualization
- Search
- Word Cloud

News Article Viewer



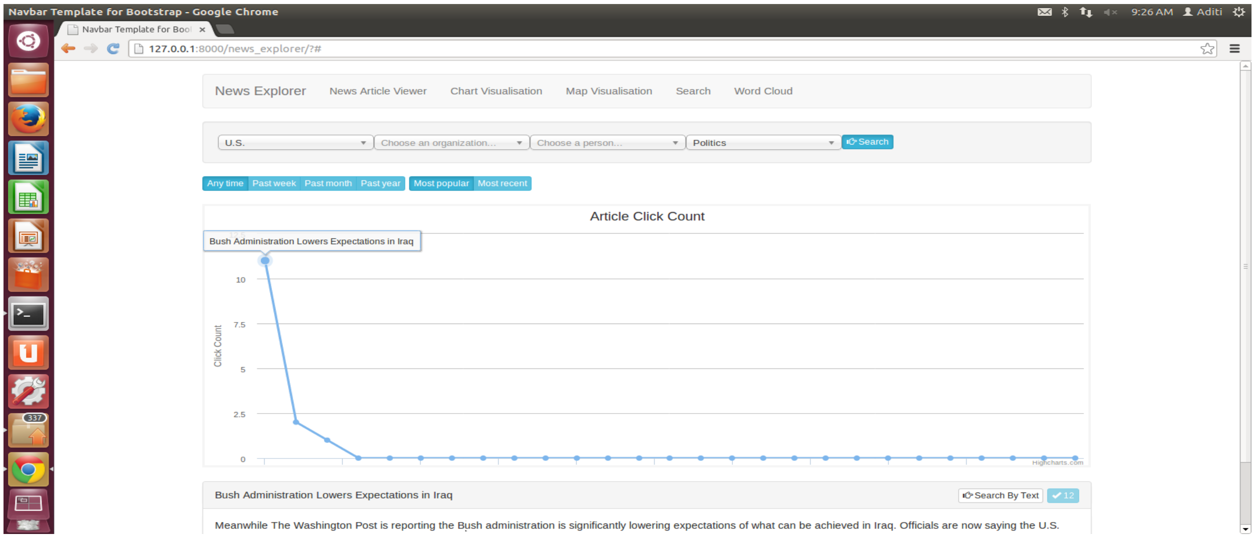
Bush Administration Lowers Expectations in Iraq	✓13
Bush: Would Withdraw Forces From Iraq If Asked	✓4
US Plans to Require Biometric Chips in Passports	✓2

Meanwhile, U.S. Attorney General Alberto Gonzales announced yesterday that Washington had agreed to delay plans requiring 27 countries to place computer chips in their passports as a prerequisite for allowing their citizens into the US without a visa. Washington had said that all European Union passports issued after Oct. 26 must have biometric security elements - fingerprints and iris scan - included on a microchip or passport holders will require a visa to enter the U.S. Gonzales made the announcement in Europe where he is attending the G8 meeting of ministers. He did not say how much extra time countries would be given.

Canada to Bow Out of U.S. Missile Defense Program	✓1
Senate Prepares for Gonzales Confirmation Hearings	✓0
Labor Union Unite Here May Quit AFL-CIO	✓0
Iraq Constitution Negotiations Remains Deadlocked	✓0
U.S. Prepares to Send More Troops To Iraq	✓0
FBI Expands Domestic Intelligence Gathering	✓0

The news article viewer lets a user choose a location, organization, person and/or category based on which the news articles are filtered and displayed. Thus a user can simply browse specific news data according to his/her interest. In addition, search tools that enable viewing articles within a time period (past week, past month, past year, any time) as well as in a particular order (most recent, most popular) are supported. The popularity of articles is measured using the number of times an article has been read i.e. the number of clicks an article has recorded which is also displayed on the page. An interesting feature is 'search by text' in which a user can select any text on the page and choose to search articles related to the desired selection enabling an intuitive navigational search flow.

Chart Visualization

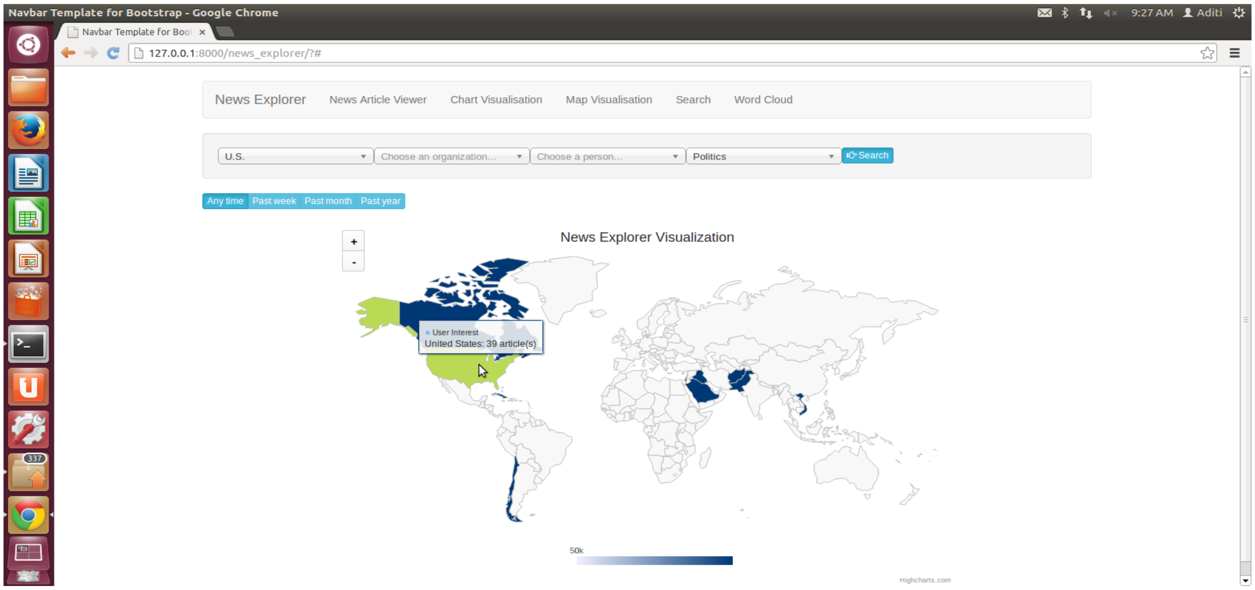


Bush Administration Lowers Expectations in Iraq

Meanwhile The Washington Post is reporting the Bush administration is significantly lowering expectations of what can be achieved in Iraq. Officials are now saying the U.S.

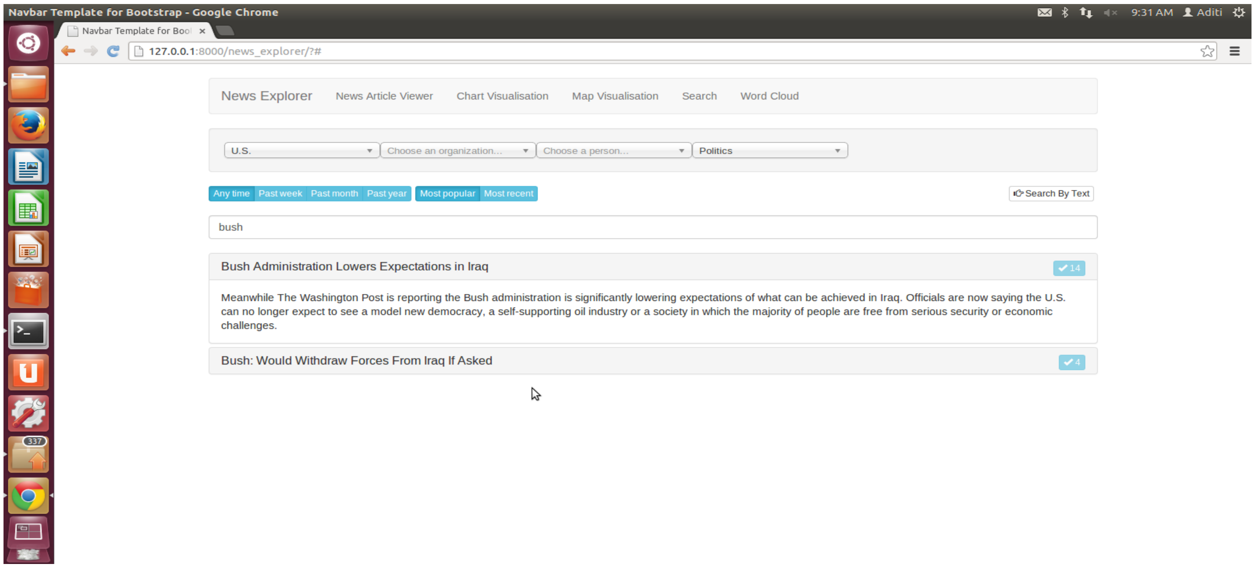
The chart visualization is a line chart that plots popularity i.e. number of clicks with respect to the articles according to the user selection (the same filtering criteria are adhered to). Every point on the chart represents an article which can be clicked to be viewed. This visualization is helpful to analysts as well as common users who want to identify popular articles or in general the public trends. The chart is rendered using High Charts API.

Map Visualization



The map visualization is a world map on which the total number of articles by country is plotted. The same filtering criteria and tools can be applied in this case to get corresponding article counts. This feature enables analysts and common users to get an idea about the article distribution as per countries around the world, based on various selections. The map is rendered using High Charts API.

Search



bush

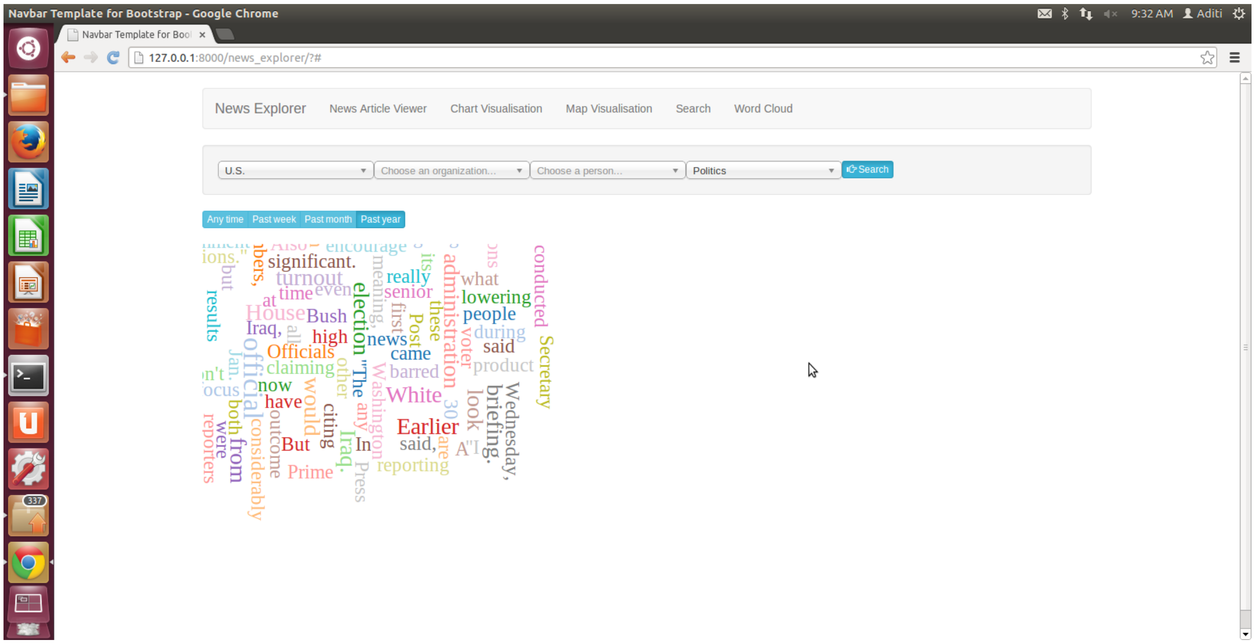
Bush Administration Lowers Expectations in Iraq ✓ 14

Meanwhile The Washington Post is reporting the Bush administration is significantly lowering expectations of what can be achieved in Iraq. Officials are now saying the U.S. can no longer expect to see a model new democracy, a self-supporting oil industry or a society in which the majority of people are free from serious security or economic challenges.

Bush: Would Withdraw Forces From Iraq If Asked ✓ 4

The search feature allows a user to search for any free form text and returns the results using the Apache Solr search server which extends the Apache Lucene library. A remarkable aspect of the search functionality is that it incorporates the same filtering criteria as before and thus very specific search can be performed according to customized selections.

Word Cloud



The word cloud is an interesting visualization in which common words in news articles (again belonging to specific selections as before) are visualized such that the larger the appearance the higher is their occurrence in the particular selection. This feature can be used by analysts and common users alike to learn about trending words or in general topics that are being mentioned often in news data. The word cloud is rendered using D3.js libraries.

Lessons Learned:

The purpose of this project was to gain insights into the field information retrieval field in terms of how classification and the search functionality work. In addition, we wanted to make the product useful for a particular type of stakeholders which in our case are consumers of news – both readers and analysts. To accomplish this, we developed a system with useful visualizations and provided a support for browsing as well as searching articles based on various categories.

Through the development process, we faced some issues and the following were some of the lessons learnt in the process:

- While setting up the infrastructure, we had compatibility issues between technology components. So, whenever one tries to setup the infrastructure, verifying the compatibility between components and their versions is of significance. In addition, the most important thing is to check whether the product/library being used has full support available and is documented adequately, else it may lead to difficulties during implementation.
- Try to make system modular, as it will reduce dependency between modules and building several components can be carried out simultaneously by each of the contributors. We achieved modularity due to our infrastructure, as it was setup with this very goal in mind - that changes by one do not interfere with those of the other.
- Develop the system using some kind of version control to ensure that multiple team members can work simultaneously and efficiently on the same project without any inconvenience or loss of information. We used git to achieve the same.

Related Work:

As of today, there exist multiple systems that aim to bridge the gap between news and its consumers. Based on our research two prominent ones are:

- Google News: It is one of the most advanced news portals that aggregates headlines from various sources, groups similar stories together and considers reader's personalized interests. But apart from a simple timeline for a story, there is not much support for visualization. Also the filtering criteria are quite restrictive in that there isn't much scope for cross domain selection. On the other hand, our system provides some interesting visualizations and supports higher user control by granting the user the freedom to choose specific entities.
- EMM News Explorer: It is a popular news portal that is closer to what our system aims to accomplish. It enables news exploration based on people and countries along with a map visualization of articles. But lack of an easy to use and appealing user interface is a prominent limitation. Our system provides a more intuitive and easy-on-the-eyes user interface along with two additional entities categorization i.e. by organization and type of articles.

Thus our system tries to overcome the shortcomings of above mentioned related work to build a more complete and comprehensive news explorer for a broad range of users. Though currently at an initial stage, not as efficient or big as other popular portals, we believe that our system can go a long way in improving how news is communicated and analyzed.

Future Scope:

The project can be extended to utilize the data parsed in the backend and perform sentiment analysis. This can be used to detect biases in news articles to suit the requirements of users who want to perceive the angle in which the news is portrayed. Most temporal analysis has already been done in the current project, but to extend it further, live news articles using a web crawler can be taken as input instead of batch processing to facilitate real-time analysis of news data. Another aspect that can be included is the comparison of news article coverage based on their source, for example: the difference in BBC News and CNN when covering news in a particular domain.

Acknowledgments:

We would like to thank Professor Cheng Zhai for approving the project and giving us the opportunity to work on it. We are grateful for his support throughout the semester, teaching us various concepts of text information systems and guiding us through the milestones. In addition, we would like to thank the Django, Python and High Charts developer community for being a major help during the project development.

Conclusion:

Our project, News Explorer, is a portal to browse, search and visualize news data. Our system caters to two sets of users – a news reader and a news analyst. We have categorized the news article dataset with high accuracy into categories such as location, organization, person and type of article. We have created relevant visualizations for the same.

Appendix:

Team Contribution Details:

We divided our project into following modules:

Sr.No.	Module Name	Module Explanation	Team Member
1.	System design	Making decisions regarding the technology stack and finalizing the design of the system architecture.	Bharat Thatavarti and Himay Jesal Desai.
2.	Database design and modeling	Storing of the raw data set in a meaningful schematic form i.e. settin up the database and models associated with it.	Aditi Mhapsekar and Himay Jesal Desai
3.	Categorization with respect to type of articles	Classification of the data set into thre types of articles: politics, terrorism and natural disasters.	Bharat Thatavarti and Aditi Mhapsekar
4.	Categorization with respect to entities	Classification of the data set based o three entities: person, organization ar location.	Aditi Mhapsekar and Himay Jesal Desai
5	Parser	Reading of the raw data set in accordance to the database design and dumping data in a MySQL database.	Aditi Mhapsekar and Himay Jesal Desai

6	Basic visualization	Visualization of different aspects of the system in terms of charts and maps.	Bharat Thatavarti and Aditi Mhapsekar
7	Search functionality	Incorporation of a generic article search plus a search by selected text functionality with the help of JQuery and Solr.	Bharat Thatavarti
8	Word cloud visualization	Visualization of a word cloud based on categories selected.	Aditi Mhapsekar
9	Testing and bug fixes	Testing the system thoroughly after implementation to fix any bugs.	Bharat, Aditi and Himay
10	Documentation and infrastructure setup of the system	Documentation of the system design/infrastructure.	Himay Jesal Desai, Bharat Thatavarti and Aditi Mhapsekar.

Note: Every team member has contributed to every module because we followed a rotational policy. The purpose was to ensure that each team member is aware of the working of the whole system and can debug any issues if and when they arise.

References:

[1] O. S. Community, "https://www.djangoproject.com/," Django Software Foundation, 2014. [Online]. Availabl https://docs.djangoproject.com/en/1.6/. [Accessed 2014].

[2] M. D. Community, "http://dev.mysql.com/," Oracle Corporation, 2014. [Online]. Available: http://dev.mysql.com/doc/refman/5.0/en. [Accessed 2014].

[3] H. C. Developers, "High Charts," Highsoft AS, 2014. [Online]. Available: http://www.highcharts.com/. [Accessed 2014].

[4] O. S. Community, "BootStrap," Apache Software Foundation, 2014. [Online]. Available: http://getbootstrap.com/2.3.2/. [Accessed 2014].

[5] J. Developers, "JQuery," Wordpress: The jQuery Foundation., 2014. [Online]. Available: http://jquery.com/. [Accessed 2014].