William Paterson University EXPLORATIONS 2020

A Web-based Platform for Mining and Analyzing Social Media Data

Daniel Novikov, Cyril S. Ku (Department of Computer Science), and Jin-A Choi (Department of Communication)

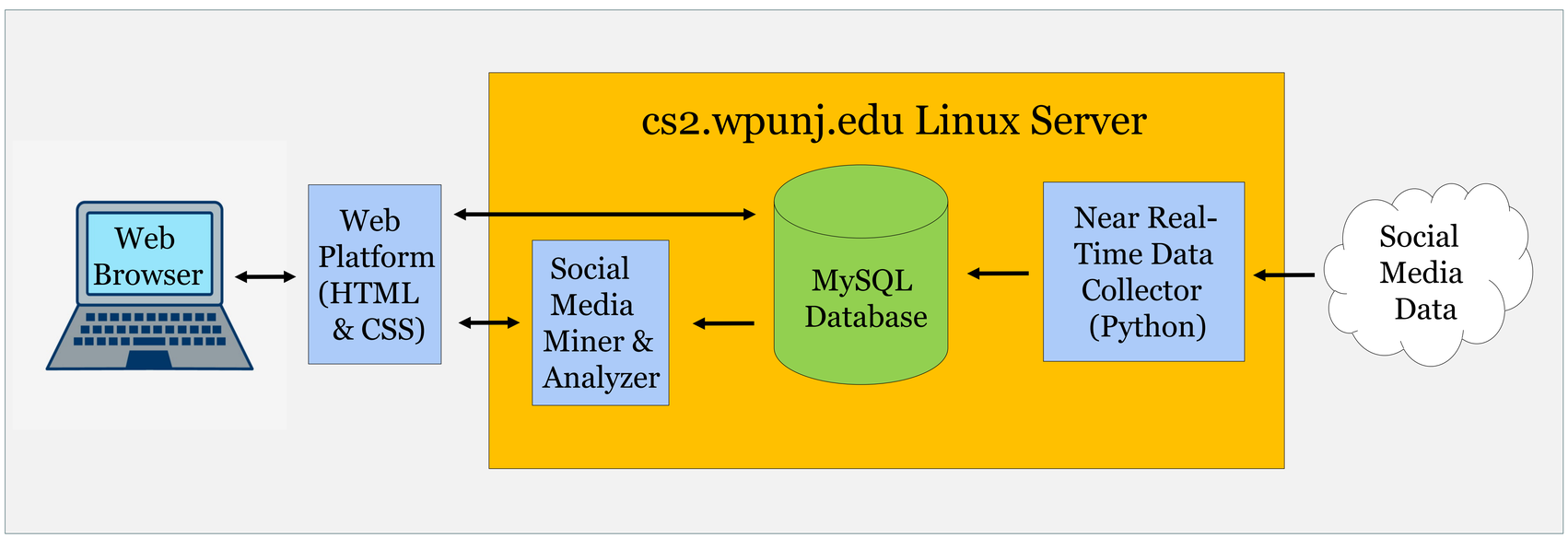
Introduction

This project is a collaboration between the Department of Computer Science and Department of Communication at William Paterson University. The purpose of the project is to build a mining and analyzing platform for communication researchers to investigate social, cultural and health issues using the large magnitude of unstructured data in social media such as Twitter, YouTube, Facebook, and other online environments.

A prototype system is being constructed using social media data from Twitter and YouTube. The system will provide data collection function (for example, keywords extraction) from Twitter tweets or YouTube transcripts and user comments. Built-in mining algorithms such as regression, classification, association, and clustering will be available. Special analyzing tools for trends, hidden relationships, sentiment analysis, as well as semantic extraction will be supported.

This poster specifies the overall architecture of this web-based platform. The front-end web/user interface uses HTML5 and CSS. It connects to a back-end relational database (MySQL) server for data storage. The interface is broadcasted via a Flask web server, which is also responsible for organizing the computation of the data-analytic requests. These requests are computed with data mining and machine learning algorithms which enable this special kind of data analysis. Examples of social, cultural, or health issues using this web-based platform are also presented.

Platform Architecture



Text Mining of Social Media Data

Social media data presents various advantages over data gathered through traditional methods. Information is created continually, and it spreads quickly. Employing text mining techniques allows researchers to gather large, real-time data from user-generated content and garner time-sensitive insights with the help of machine learning. These techniques enable a wide array of cases to be examined in science, communication, industry, and public health, including:

- The ability for businesses to predict new trends.
- The ability for health communicators to create early response strategies.

We use this platform with Twitter to gain real-time insights into the public's conversations about the COVID-19 outbreak, and with YouTube to see how the presence of certain words in a video's transcript, description, and comments affect the likelihood of brand endorsement.

References

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.), Morgan Kaufmann Publishers.
- McLaughlin, M. (2013). *MySQL Workbench: Data Modeling & Development,* Oracle Press, McGraw Hill.
- Python Software Foundation. *https://www.python.org/*

Acknowledgements

This research was supported in part by the ART (Assigned Released Time) program, Office of the Provost, and in part by the College of the Arts and Communication Center for Creative Activity & Research Summer 2019 Grant, William Paterson University.