Team 5

Joint Analysis of Hotel Review & Historical Local Economy Metrics for Causal Topics

One potential instance of the data mining loop (described in Text Data Management and Analysis book[1]) that combines text and non-text data would be a joint analysis of hotel review content and economic indicators. See Figure 1 which is a modified diagram found in the Zhai book. As human sensors record their experiences during their hotel stays, they generate the hotel review content. These hotel reviews can have associated time and location information. Applying various text mining techniques to unearth sentiments or topics provide another means of eliciting information. Economic indicators expressed with time and location context coupled with the sentiments or topics may contribute valuable insight and context about the performance of local economies.

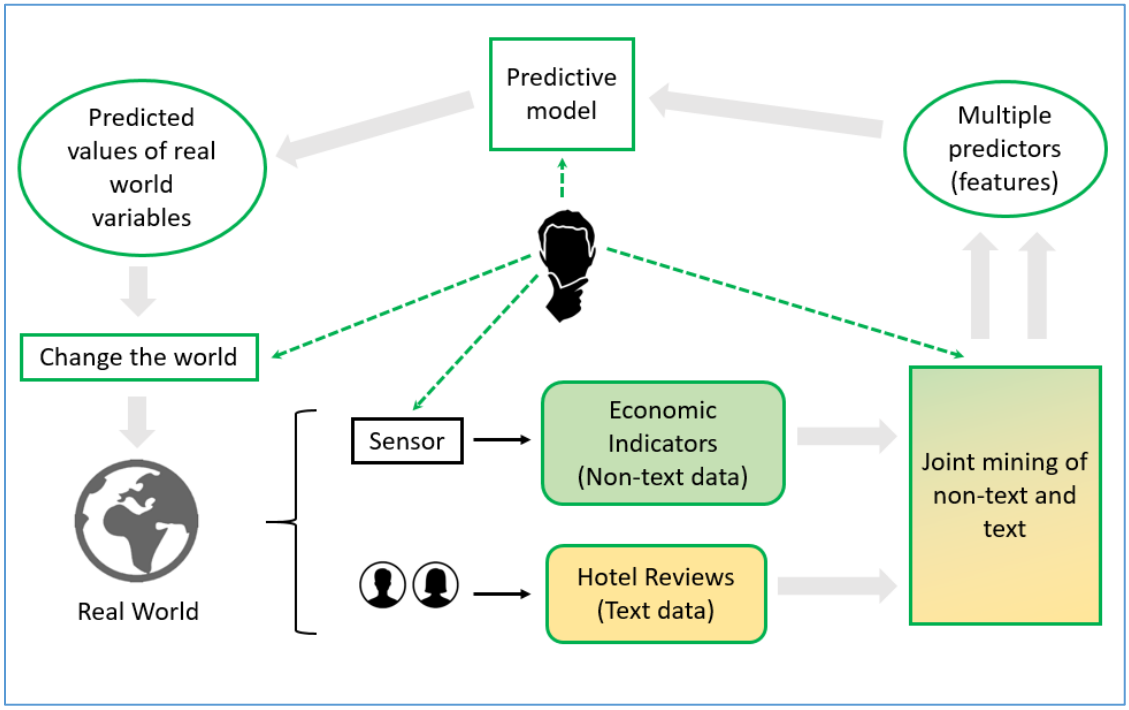

Figure 1. Diagram of the data mining loop with hotel review and economic indicators

The end goal of this project would involve building a tool to ingest hotel review text data and non-text economic indicator data to provide insights for local economies. Potential users of the tool may include:

- Government agencies who would like to gather insights about local economies
- Corporations in the hospitality industry aiming to strategically locate new businesses or franchisees
- Curious economists who want to understand consumer behavioral trends

The tool would have at least three major parts:

1. A hotel review text data ingestion agent
2. A economic indicator data agent
3. A main analytic engine to depict insight

[1] ChengXiang Zhai, Sean Massung, *Text Data Management and Analysis* (ACM Books 2016) 413-418.

The hotel review text data ingestion agent would be able to take raw hotel review text data and extract sentiments or topics. The sentiments or topics would be processed to become time-series and location specific. For this project, the hotel review text data ingestion agent would be structured to specifically extract the information from the UCI Machine Learning repository that has a reasonably sized set of hotel reviews from ten different cities (link here).[2] To make a hotel review text data ingestion agent that can work with any text data from any hotel review source would be outside of this project.

The economic indicator data agent would be able to take time period and location as inputs and generate a time series dataset for particular cities or metropolitan areas. The Open Data Network has historical metrics (like GDP, personal income, and population) for metropolitan areas (link here to example data). Building an agent that has the capability to process information beyond what is available through the Open Data Network will be outside the scope of this project.

The main analytic engine that takes the data generated by the hotel review test data ingestion agent and the economic indicator data agent would be the central portion of this tool. It would be able to accommodate other sources if other hotel review text data ingestion agents built to take in other sources than the UCI Machine Learning repository or other economic indicator data agents built to take in other sources than the Open Data Network were built. The main analytic engine would provide graphical depiction of various economic indicators like employment, gross domestic product (GDP), or personal income as a function of time and locality.

I will attempt to apply the Granger Causality Test to the various topics or sentiments in the hotel reviews. The results of the Granger Causality Test would inform the user of meaningful insights that can be drawn from the analysis.

Some possible hypotheses that this tool could explore include:

- Complaints about service may correlate with rising economic trends because rising GDP may incur labor shortages
- Increasingly positive reviews for leisure visits may indicate improving local economies
- Seasonal reviews about quality can correlate with maintenance needs

[2] Kavita Ganesan, ChengXiang Zhai, Jiawei Han. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). Beijing, China.