BizTrender: Project Summary
Group 9 : ab2083, pt357, cd817

Note : Please refer the project report for details.

BizTrender is a *search* and *visualization* solution to streamline the data of various businesses listed on Yelp and discover trends on top of it. We build an end-to-end data pipeline to ingest, transform, search and aggregate publicly available data sets from the *Yelp Dataset Challenge* on the backend, and display useful information or visualize key trends on a GUI frontend. The user driven system enables interactivity between data and analysts, assisting the latter to draw insights from the data and make accurate business strategy.

Data – 5 JSON files from the dataset : business.json, review.json, user.json, checkin.json, tip.json. The files add up to 8.6 GB of data when uncompressed. The datalake created by ingesting these files is static.

Questions – WHAT? : web application that provides a centralized platform to search & perform visual analytics on businesses listed on Yelp; WHY? : analysts or consultants who would like to study the performance of a business or group of businesses, with respect to various metrics, identify correlations that drive those performances, & infer trends to devise business strategies; HOW? : search functionality provides a starting point & defines a domain of analysis for the analyst, & visualizations to explore trends in the dataset.

Target Users – business analysts or independent consultants, business owners, real-estate consultants.

Data Processing – raw data set in JSON format; master data (business) indexed on Elasticsearch for the search functionality; other data attributes are represented as facts and dimensions stored in MongoDB database which can be pulled to join with the master data and perform any aggregation/transformation for visualization.

Software Stack – Datalake : Elasticsearch & MongoDB; Backend : JAVA Spring Boot & MVC Architecture; Frontend : Javascript, AngularJs & Highcharts; Scripting/Commands : Bash, Mongo shell, Python

Visual Representations – Use of standard UI fonts, sizes, color & texture palettes; Elements : Search box, click buttons, dropdown menus, unordered lists, navigation buttons, canvas divisions for plots; Interactive plots: 3D column chart with stacking and grouping, map (US and Canada) with latitudes/longitudes, packed bubble chart, dynamic column charts, dynamic time-series line chart, histogram, fixed placement column chart, & word cloud; Multi-colored labels and legends.

Interactivity – Hierarchical abstraction : Generic (trend summary) and specific (search) views; Bidirectional vertical sliders, Both textual and graphic representations for answered user queries; Mouse clicks, mouse hovering, mouse selection and their combinations; Menu driven plots and tabs; Mouse controlled zooming and panning; Built-in options for full-screen view and exporting plots; Different views linked through navigation using mouse clicks.

Key Insights – Largest number of businesses in the dataset are either restaurants or food-based; Religious organizations is the only category that has no business with rating 1; For any category, most of the businesses are rated between 2 & 4; Smallest number of businesses in each category are the ones that are rated the worst at 1; all businesses lie in one of the clusters within these states/provinces - Alberta, Quebec or Ontario in Canada, & Nevada, Arizona, Illinois, Wisconsin, Ohio, Pennsylvania, North Carolina or South Carolina in the USA; Higher rating does not imply more visits for any business; Religious organizations receive the highest number of visits on Sunday, which is significantly higher than the visits received on any other days; businesses receive more visits during late evening or early morning as compared to daylight hours; For poorly rated restaurants (1 to 2), the review sentiment is mostly negative with few positive scores in between; For middle-rated restaurants (2 to 3, the review sentiment is mostly centered (roughly symmetrical) about zero (neutral); For restaurants with good ratings (above 3), the review sentiment is mostly positive with few negative scores in between; For a specific business, the number negative reviews corresponds to the number of users rating the business below their average user rating.

Conclusion – Through BizTrender, we successfully demonstrated how visual analytics can be used to support the process of mining trends in a large dataset. We also used many visual metaphors, interaction mechanisms & abstractions learnt in class to build the final project.