City Genres

Columbia Innovation Lab x Priceline Data

priceline

Goals of the Project

Identify what makes two cities similar from a traveler's perspective, find data to build your case, create features from diverse website data to help you build a clustering model, deliver the "genre" for each city, along with key artifacts.

Web Scraping & Parsing:	Scrape publicly available websites to find numerical and topical insights. Use NLP where relevant.
Data Mining:	For each city, use the mined data to create a series of features that describe its unique characteristics
Data Refinement & EDA:	Identify trends and outliers. Apply descriptive statistics and/or inferential statistics. Document data description and transformations.
Build a Data Model:	Cluster the cities based on the features you create and return the cluster labels
Present Results	Summarize findings with narrative storytelling. Present limitations and assumptions of your work. Identify follow-up problems and opportunities for further analysis.

priceline

Deliverables of the Project Goals

Start with a SMART goal: specific, measurable, achievable, relevant, time-bound.

Web Scraping & Parsing:	Python code for scraping and parsing for each site, with good documentation / commenting for us to productionize, for a different city list.
Data Mining:	Documentation on data transformations and feature engineering, including hypotheses and reasoning.
Data Refinement & EDA:	Appendix slides for final slide deck with detailed analyses of trends and topics across cities and by cluster.
Build a Data Model:	Final data set should have all features, final cluster labels from model and original city,state,country as main identifier for each row.
Present Results	Final slide deck with narrative of analysis outlining each stage of project in a captivating and interesting narrative, with extra analysis charts and insights in appendix.

priceline

Data Sources

- **City List provided by Priceline Data**
- Sites to scrape (up to project, but recommendations)
  - TripAdvisor city pages for topic modeling
  - Wikipedia city pages for topic modeling
  - Google Trends by city for “popularity” score
- APIs to scrape
  - Weather data by city across different seasons in past 12 months
  - Social media sentiments about each city

priceline