

---

# Weekender : The SF Restaurant Planner

By: Aanvi Goel

## ABSTRACT

This project aims to build a Content-Based Recommender System for recommending restaurants in the San Francisco Bay Area. Place data from [Google Places API](#) is queried to build a dataset with features such as restaurant names, ratings price-levels and reviews. The user reviews are pre-processed using NLP techniques and converted into feature vectors. The reviews are further used to extract Topics based on common themes using Latent Dirichlet Allocation Model. Finally, cosine similarity is used as the evaluation metric for obtaining mathematical similarity between places and to provide top 5 recommendations.

## DESIGN

San Francisco Bay Area has more than 10k restaurant options, with over 4000 restaurants just within the city of San Francisco. With so many options available, deciding which restaurant to eat at or order from can be a nerve-wracking and tiring experience causing analysis paralysis and decision fatigue. This project uses easily-accessible Google Maps data and the power of NLP to build a recommender system that can provide personalized recommendations for restaurants based on Reviews and other factors such as price level and ratings.

## DATA

### Web scraping Google Maps using Places API:

<https://developers.google.com/maps/documentation/places/web-service>

Features are extracted from the JSON response, providing 20 Places per response. Using the Place ID, corresponding reviews are queried which outputs 5 reviews/place.

## ALGORITHM

### *Data Pre-Processing:*

- Text reviews are cleaned by removing punctuations, whitespaces using RegEx along with removing Stop Words
- WordNetLemmatizer() is used to convert text into 'root words'

---

*Models:*

- **Vectorization:**
  - **TF-IDF:** Term Frequency-Inverse Document Frequency are used to convert text reviews into feature vectors for the content-based recommender Model #1
  - **spaCy:** spaCy open library is used to process text using its pre-trained models and NLP pipelines for Recommender Model #2
- **Topic Modeling using LDA :**
  - Text reviews are used to obtain 10 topics using LDAModel from Gensim library
  - Topic probabilities are converted into Feature columns for the dataset for Recommender Model #3
- **Cosine Similarity:** Cosine similarity is used as the evaluation metrics for obtaining mathematical difference between two vectors

## TOOLS

- *Data Acquisition:* [Google Places API](#) along with Requests python package
- *Data Cleaning and Processing:* Pandas, NLTK, RegEx, spaCy
- *Data Modeling:* Scikit-learn, Gensim
- *Data Visualization:* Matplotlib, WordCloud

## COMMUNICATION

- A slidedeck and Jupyter notebook code are included along with this write-up as part of the project on [Github](#)