Project Ideas

FOR YOUR DATA SCIENCE PORTFOLIO

WWW.THEARTISTSOFDATASCIENCE.COM

Summarize Review Text

I. Problem

Given text from a set of reviews (e.g., product, movie, restaurant) investigate information extraction and summarization methods to automatically extract and summarize different aspects of the reviews, e.g., price, service, quality of food.

II. Possible Data Sources

Yelp Challenge data sets, Amazon product review data sets, Wikipedia pages with lists of attributes (food types for restaurants), etc.

III. Technical Approaches

Information extraction techniques for finding sentences with specific words, Machine learning classification methods for sentiment analysis, etc.

IV. Interaction/Visualization Goals

Form-based interface that allows a user to select a product or a movie or a business and that generates a visual summary of the reviews.

V. Extensions

Display other information about the business, for example sentiment over time

Tagging of Tweets or News Articles

I. Problem

Use Wikipedia pages and categories as training data and build a classification algorithm that can classify tweets and news articles into predefined topics. Or cluster tweets/articles into trending topics.

II. Possible Data Sources

Wikipedia pages and categories, Tweets, Kaggle or NYT news articles.

III. Technical Approaches

Machine learning classification algorithms (possibly with hierarchies, Web crawling to gather news articles in real-time.

IV. Interaction/Visualization Goals

Interface where a user provides a Website URL (newspaper or other news media site) and system downloads stories, tags them, and displays the results

V. Extensions

Also automatically extract and highlight entities (people, places, organizations).

Analyze Geolocated Tweets

I. Problem

Use geolocated tweets (with GPS) from any city of your choice over multiple months to see how sentiment in tweets varies over time and spatially – or how entity mentions (e.g., SF 49ers versus Oakland Raiders, Sacramento Kings versus Golden State Warriors) varies spatially

II. Possible Data Sources

Twitter API,OpenMaps data, census data

III. Technical Approaches

Information extraction, entity detection, spatial and temporal visualization

IV. Interaction/Visualization Goals

Interface where a user can provide entity names and get back a spatial map or time showing prevalence of tweets over space or time.

Visualization of Tweet Sentiment

I. Problem

Given a large set of tweets related to a particular query or hashtag, generate an interactive map of where the tweets are being generated (by country, by US state, by city) and spatial distribution of their sentiment - take into account population density

II. Possible Data Sources

Twitter API, map data, Census data, labeled Twitter data sets.

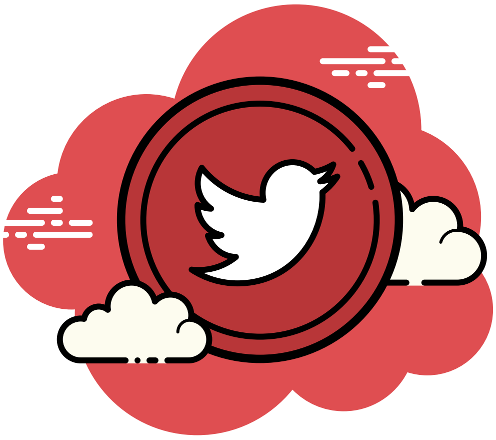
III. Technical Approaches

Understanding and analysis of census data and map data (GIS aspects), classification and/or clustering algorithms.

IV. Interaction/Visualization Goals

Allow a user to select a topic (possibly from a predefined list, with stored Tweets) and system then generates a 2d map visualization. Possibly allow more detailed “drill down” on specific locations or via further keyword or hashtagrestrictions to subset the tweets and then display the results.

Visualization of Tweet Sentiment



I. Problem

Given a hashtag (or set of related hashtags), download Tweets with this hashtag and automatically summarize the large set of tweets for a user

II. Possible Data Sources

Twitter API, labeled twitter data sets, Wikipedia categories.

III. Technical Approaches

Clustering algorithms, topic models, machine learning sentiment classification algorithms

IV. Interaction/Visualization Goals

Allow a user to select from a large set of predefined hashtags, and visually display the clustering/classification results in an understandable manner

Public Opinion Topics

I. Public Health Issues

Predict impact(s) of legalization of marijuana (you can look for data at healthpolicy.ucla.edu/), or examine aspects of the opioid epidemic. UCLA Public Health policy has some great resources. Dive in!

II. Gun deaths and gun control

Examine time distribution (magnitude, sentiment, etc) of public response in social media. Dimensions to examine might be positive/negative positions – as a function of time after an incident related to: Increasing background checks, raising age requirements, banning semi-automatic weapons, arming additional individuals (for example, teachers), etc.

Combined CDC data for total gun death rates, gun suicide rates, gun homicide rates and gun homicides caused by legal intervention each of different size.

Other Topics

I. Health Care Analytics

Performing feature selection to identify the most predictive variables in a database of cancer patients, customized based on personal traits, and using them for creating predictive models that can aid general patient medical care as well as personalized healthcare.

II. Data Science for A Sustainable Planet

Visually exploring energy usage patterns by smart homes equipped with green energy production devices as well as energy monitoring devices to optimize system-wide energy-usage, determine optimized charging, and predict future shortfalls and distribution challenges for smart grids affected by environmental characteristics such as hot summer days or long cold winters.

III. Big Data Security Analysis

Applying machine-learning approaches for classifying and categorizing Android sources and sinks to determine malicious agents or analyzing network big data for real-time intrusion detection.

Other Topics

IV. Intelligent Transportation Planning

Mining and querying data from the so-called internet of things such as GPS traffic data sets from a large metropolitan area to establish insights into taxi routes, best places to relocate bus stops to maximize flow and/or save fuel for the fleet, the identification of traffic patterns that contribute to increased traffic jams and/or accidents as the region's population undergoes growth.

V. Analytics for Business and Engineering Optimization

Development of predictive models that extract insights and enable optimization of engineering processes for avoidance for down-time under equipment failure or for monitoring of supply chain enterprise performance for agile adaption under supply change interruption due to weather or natural disasters.

Using Your Personal Data

Activity Prediction

Using data from your wearable device, try to predict your daily activity over the next week. Try to predict your caloric burn over the next week. You can also tie in historical weather data and upcoming weather forecasts to see the effect of weather on your activity level. If you track your food intake, can you try to understand the impact of or relationship between your food intake and your activity levels?

Spotify Listening Habits

What if we could analyze the music we listen to using Data Science derive insights on the types of music we listen to? You can use the Spotify API, or the Python package **SpotiPy** to pull your listening data. What sort of music styles are you often listening to on Spotify? Is the music that is featured often more acoustic? Does the style of music I listen to differ from those that are played in the top 100 and different from those featured on Spotify? Are you generic: Meaning, do you listen to music that follows a similar pattern when compared to the most popular songs? Can you use machine learning to build your own recommender system for songs?

Sports Data

Major League Baseball

Baseball is probably the world’s best documented sports. The history has cumulated records in the past hundred years of the baseball statistics. In fact, there is an entire field focusing on the statistical analysis of baseball called Sabermetrics. Baseball stats consist of numerous metrics, some of them straight-forward, some of them quite advanced. Can you use data to decipher baseball pitching strategy? How about a Bayesian analysis of pitching strategy? The MLB has an API that you can use here: https://appac.github.io/mlb-data-api-docs/

Here is a list of sites you can checkout for baseball related data science projects:

- https://blog.galvanize.com/the-data-science-behind-baseball-pitching-strategy/
- https://towardsdatascience.com/bayesball-bayesian-analysis-of-batting-average-102e0390c0e4
- http://datascience.uconn.edu/index.php/component/k2/item/194-a-predictive-model-for-runs-scored-by-mlb-players
- https://www.kdnuggets.com/2019/09/time-series-baseball.html

Sports Data

National Football League

The National Football League is America's most popular sports league, comprised of 32 franchises that compete each year to win the Super Bowl, the world's biggest annual sporting event.

Founded in 1920, the NFL developed the model for the successful modern sports league, including national and international distribution, extensive revenue sharing, competitive excellence, and strong franchises across the country.

Checkout the API here to learn more: https://api.nfl.com/

Here is a list of sites you can checkout for football related data science projects:

- https://medium.com/swlh/predicting-nfl-scores-in-python-3560ccd58cb1
- https://towardsdatascience.com/simulating-the-2020-nfl-season-100-000-times-6f82644b67af
- https://github.com/brendan-d-freeman/nfl-playcall-prediction
- https://github.com/TheDavidChen/NFL-JameisWinston
- https://www.kaggle.com/c/NFL-Punt-Analytics-Competition
- https://www.youtube.com/watch?v=zYSDlbr-8V8

Sports Data

National Football League

The National Football League is America's most popular sports league, comprised of 32 franchises that compete each year to win the Super Bowl, the world's biggest annual sporting event.

Founded in 1920, the NFL developed the model for the successful modern sports league, including national and international distribution, extensive revenue sharing, competitive excellence, and strong franchises across the country.

Checkout the API here to learn more: https://api.nfl.com/

Here is a list of sites you can checkout for football related data science projects:

- https://medium.com/swlh/predicting-nfl-scores-in-python-3560ccd58cb1
- https://towardsdatascience.com/simulating-the-2020-nfl-season-100-000-times-6f82644b67af
- https://github.com/brendan-d-freeman/nfl-playcall-prediction
- https://github.com/TheDavidChen/NFL-JameisWinston
- https://www.kaggle.com/c/NFL-Punt-Analytics-Competition
- https://www.youtube.com/watch?v=zYSDlbr-8V8

Other Random Project Ideas

I. Speech Emotion Recognition

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. librosa is a Python library for analyzing audio and music. You can create a project to build a model to recognize emotion from speech using the librosa and sklearn libraries and the RAVDESS dataset. Google "RAVEDESS dataset", it's easy to find!

II. Traffic Sign Recognition

There are several different types of traffic signs like speed limits, no entry, traffic signals, turn left or right, children crossing, no passing of heavy vehicles, etc. Traffic signs classification is the process of identifying which class a traffic sign belongs to. You can build a deep neural network model that can classify traffic signs present in the image into different categories. You can use this dataset from Kaggle:
https://www.kaggle.com/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign