**Yelp Review Critiques**

Deepa Warrier Nikita Sathish

**Abstract**

We have taken up the data from yelp data challenge and since yelp does not provide / categorize the business reviews into positive and negative, through this project we will achieve the same. Using the Business data and review data provided in <http://www.yelp.com/dataset_challenge> we will perform text mining and categorize the review for the specified business name as positive or negative.

A GUI will be created through which the user can enter the business name and the business name will be sent to the model which in turn will fetch the reviews from the review data based on the business id determined by the business name which is common in both business and review data. The reviews will then be categorized as positive and negative based on the positive/negative words used in the reviews. Various visualizations are portrayed for the dataset given.

**Background**

Yelp is a multi-national corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as online reservation service [SeatMe](http://en.wikipedia.org/wiki/SeatMe" \o "SeatMe) and food delivery service [Eat24](http://en.wikipedia.org/w/index.php?title=Eat24&action=edit&redlink=1). The company also trains

small businesses to respond to reviews responsibly, hosts social events for reviewers, and provides data about businesses, such as health inspection scores. The Yelp user community is primarily active in major metropolitan regions.[[9]](http://en.wikipedia.org/wiki/Yelp#cite_note-one-9) The site has pages devoted to individual locations, such as restaurants or schools, where Yelp users can submit a review on their products or services[[48]](http://en.wikipedia.org/wiki/Yelp" \l "cite_note-registerregister-48) using a one to five star [rating system](http://en.wikipedia.org/wiki/Rating_scale).[[9]](http://en.wikipedia.org/wiki/Yelp#cite_note-one-9) Businesses can also update contact information, hours and other basic listing information or add special deals.[[9]](http://en.wikipedia.org/wiki/Yelp#cite_note-one-9)[[14]](http://en.wikipedia.org/wiki/Yelp#cite_note-nyttttt-14) In addition to writing reviews, users can react to reviews, plan events or discuss their personal lives.[[4]](http://en.wikipedia.org/wiki/Yelp#cite_note-oneone-4) According to Sterling Market Intelligence, Yelp is "one of the most important sites on the Internet."[[21]](http://en.wikipedia.org/wiki/Yelp#cite_note-NYSE-21) It has 132 million monthly visitors and 57 million reviews.[[49]](http://en.wikipedia.org/wiki/Yelp#cite_note-49) As of 2010, Yelp had an established user-base in 33 cities. Yelp added the ability for business owners to respond to reviews in 2008.[[9]](http://en.wikipedia.org/wiki/Yelp#cite_note-one-9)[[63]](http://en.wikipedia.org/wiki/Yelp#cite_note-nytloi-63) Businesses can respond privately by messaging the reviewer or publicly on their profile page. In some cases, Yelp users that had a bad experience have updated their reviews more favorably due to the businesses' efforts to make it right. In other cases disputes between reviewers and business owners have led to harassment and physical altercations.[[9]](http://en.wikipedia.org/wiki/Yelp#cite_note-one-9) The system has led to criticisms that business owners can bribe reviewers with free food or discounts to increase their rating, though Yelp users say this rarely occurs.[[51]](http://en.wikipedia.org/wiki/Yelp#cite_note-twotwotwo-51) A business owner can "claim" a profile, which allows them to respond to reviews and see traffic reports.[[9]](http://en.wikipedia.org/wiki/Yelp#cite_note-one-9) Businesses can also offer discounts to Yelp users that visit often using the "check in" feature.[[54]](http://en.wikipedia.org/wiki/Yelp#cite_note-lsoppio-54)[[55]](http://en.wikipedia.org/wiki/Yelp#cite_note-lsoppio2-55) In 2014, Yelp released an app for business owners to respond to reviews and manage their profiles from a mobile device.[[64]](http://en.wikipedia.org/wiki/Yelp#cite_note-64) Business owners can also flag reviews to be removed, if they violate Yelp's content guidelines.[[65]](http://en.wikipedia.org/wiki/Yelp#cite_note-65)

**Yelp Data Challenge**

Yelp is providing their data and reviews to users to predict, analyze and visualize Yelp data through Yelp data challenge. The dataset is a single gzip-compressed file, composed of one json-object per line. Every object contains a 'type' field, which tells you whether it is a business, a user, or a review.

**Business Objects**

Business objects contain basic information about local businesses. The 'business\_id' field can be used with the Yelp API to fetch even more information for visualizations, but note that you'll still need to comply with the API TOS. The fields are as follows:

{

'type': 'business',

'business\_id': (a unique identifier for this business),

'name': (the full business name),

'neighborhoods': (a list of neighborhood names, might be empty),

'full\_address': (localized address),

'city': (city),

'state': (state),

'latitude': (latitude),

'longitude': (longitude),

'stars': (star rating, rounded to half-stars),

'review\_count': (review count),

'photo\_url': (photo url),

'categories': [(localized category names)]

'open': (is the business still open for business?),

'schools': (nearby universities),

'url': (yelp url)

}

**Review Objects**

Review objects contain the review text, the star rating, and information on votes Yelp users have cast on the review. Use user\_id to associate this review with others by the same user. Use business\_id to associate this review with others of the same business.

{

'type': 'review',

'business\_id': (the identifier of the reviewed business),

'user\_id': (the identifier of the authoring user),

'stars': (star rating, integer 1-5),

'text': (review text),

'date': (date, formatted like '2011-04-19'),

'votes': {

'useful': (count of useful votes),

'funny': (count of funny votes),

'cool': (count of cool votes)

}

}

**User Objects**

User objects contain aggregate information about a single user across all of Yelp (including businesses and reviews not in this dataset).

{

'type': 'user',

'user\_id': (unique user identifier),

'name': (first name, last initial, like 'Matt J.'),

'review\_count': (review count),

'average\_stars': (floating point average, like 4.31),

'votes': {

'useful': (count of useful votes across all reviews),

'funny': (count of funny votes across all reviews),

'cool': (count of cool votes across all reviews)

}

}

 Yelp challenges to use this data in an innovative way and break ground in research.

**Cultural Trends:**By adding a diverse set of cities, we want participants to compare and contrast what makes a particular city different. For example, are people in international cities less concerned about driving in to a business, indicated by their lack of mention about parking? What cuisines are Yelpers raving about in these different countries? Do Americans tend to eat out late compared to the Germans and English? In which countries are Yelpers sticklers for service quality? In international cities such as Montreal, are French speakers reviewing places differently than English speakers?

**Location Mining & Urban Planning:**How much of a business' success is really just location, location, location? Do you see reviewers' behavior change when they travel?

**Seasonal Trends:**What about seasonal effects: Are HVAC contractors being reviewed just at onset of winter, and manicure salons at onset of summer? Are there more reviews for sports bars on major game days and if so, could you predict that?

**Infer Categories:**Do you see any non-intuitive correlations between business categories e.g., how many karaoke bars also offer Korean food, and vice versa? What businesses deserve their own subcategory (i.e., Szechuan or Hunan versus just "Chinese restaurants"), and can you learn this from the review text?

**Natural Language Processing (NLP):**How well can you guess a review's rating from its text alone? What are the most common positive and negative words used in our reviews? Are Yelpers a sarcastic bunch? And what kinds of correlations do you see between tips and reviews: could you extract tips from reviews?

**Predict Attributes:**Can you correlate our reviews with business attributes and predict whether a business is good for kids? Has Wi-Fi? Has Parking?

We have chosen the Natural Language Processing (NLP) to predict the reviews as negative or positive for restaurants. This helps the users to choose the restaurant they want to go. We have also shown various visualizations for the data provided.

**Datasets**

The challenge dataset has 1.6M reviews and 500K tips given by 366K users for 61K businesses and 481K business attributes, e.g., hours, parking availability, ambience. It also has Social network of 366K users for a total of 2.9M social edges and aggregated check-ins over time for each of the 61K businesses. In this project we have used the business file and the review file.

**Data Preprocessing**

Data pre-processing is an important step in the [datamining](http://en.wikipedia.org/wiki/Data_mining" \o "Data mining) process. The phrase ["garbage in, garbage out"](http://en.wikipedia.org/wiki/GIGO) is particularly applicable to data mining and [machine learning](http://en.wikipedia.org/wiki/Machine_learning)projects. Data-gathering methods are often loosely controlled, resulting in [out-of-range](http://en.wikipedia.org/w/index.php?title=Range_error&action=edit&redlink=1) values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), [missing values](http://en.wikipedia.org/wiki/Missing_values), etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and [quality of data](http://en.wikipedia.org/wiki/Data_quality) is first and foremost before running an analysis.[[1]](http://en.wikipedia.org/wiki/Data_pre-processing#cite_note-1)

If there is much irrelevant and redundant information present or noisy and unreliable data, then [knowledge discovery](http://en.wikipedia.org/wiki/Knowledge_discovery) during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

The business and review file is cleaned from all stop words, punctuations and words that begin with digits. NLTK is used to preprocess the data.

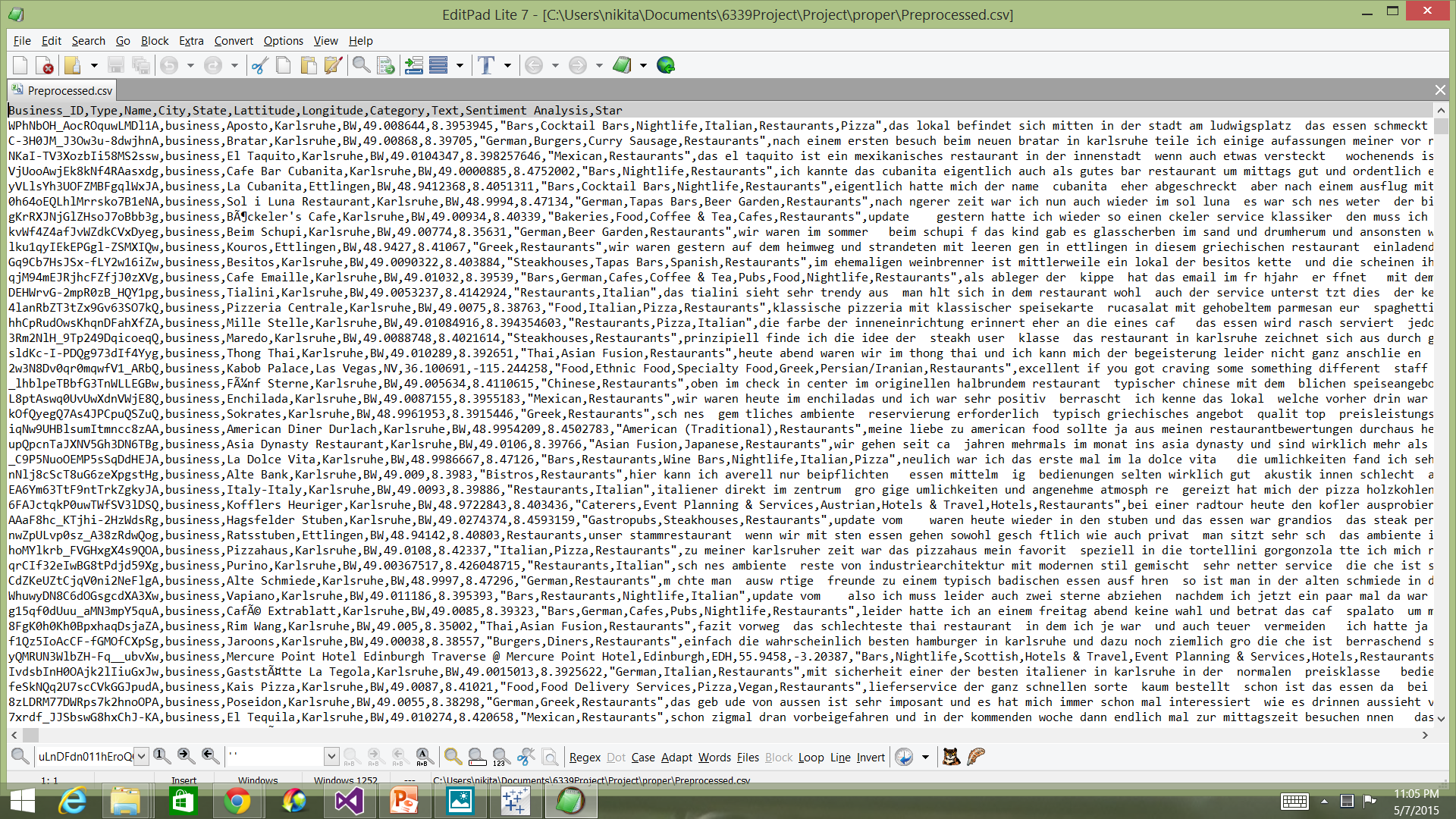


Fig1: Preprocessed data

**Classifiers**

The preprocessed Data is fed into classifiers. The data run through classifiers gives the sentimental value for the reviews. We have run the data in the following classifiers : logistic regression, SVM (Linear) and SVM (RBF) and extra tree classifier.

**Logistic Regression**

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) [continuous](http://en.wikipedia.org/wiki/Level_of_measurement#Interval_scale), by estimating probabilities. Thus, it treats the same set of problems as does [probit regression](http://en.wikipedia.org/wiki/Probit_regression) using similar techniques; the first assumes a [logistic function](http://en.wikipedia.org/wiki/Logistic_function) and the second a standard [normal distribution](http://en.wikipedia.org/wiki/Normal_distribution) function.

**SVM**

support vector machines are [supervised learning](http://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](http://en.wikipedia.org/wiki/Algorithm) that analyze data and recognize patterns, used for [classification](http://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](http://en.wikipedia.org/wiki/Regression_analysis). Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-[probabilistic](http://en.wikipedia.org/wiki/Probabilistic_classification) [binary](http://en.wikipedia.org/wiki/Binary_classifier) [linear classifier](http://en.wikipedia.org/wiki/Linear_classifier).

**Extra tree classifier**

ExtraTreeClassifier is an extremely randomized version of DecisionTreeClassifier meant to be used internally as part of the ExtraTreesClassifier ensemble. If your main goal is maximizing prediction accuracy you should almost always use an ensemble of decision trees such as ExtraTreesClassifier instead of training individual decision trees.

**Output file**

After preprocessed file is run through the classifiers, a CSV file is generated which has all the business IDs, names, stars, reviews and the sentimental values. The sentimental value is either 0 or 2. If the value is 0, then the review is tagged to be negative and if the value is 2, it is tagged to be positive.

**GUI**

A web page is created for the users to enter the business name and get the sentiment for that business as positive or negative. The web page is created using ASP .Net. Once the user enters the business name, the web service fetches the sentimental value for it by linking with the business ID in the output CSV file. It then renders the business as positive or negative to the user on the web page.



Fig 2: Yelp critique GUI



Fig 3:GUI output

**Visualiztion:**

Data visualization or data visualisation is viewed by many disciplines as a modern equivalent of visual communication. It is not owned by any one field, but rather finds interpretation across many (e.g. it is viewed as a modern branch of [descriptive statistics](http://en.wikipedia.org/wiki/Descriptive_statistics) by some, but also as a grounded theory development tool by others). It involves the creation and study of the [visual](http://en.wikipedia.org/wiki/Visual_system) representation of [data](http://en.wikipedia.org/wiki/Data), meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of [information](http://en.wikipedia.org/wiki/Information)".[[1]](http://en.wikipedia.org/wiki/Data_visualization#cite_note-MF08-1)

A primary goal of data visualization is to communicate information clearly and efficiently to users via the [information graphics](http://en.wikipedia.org/wiki/Infographic) selected, such as tables and charts. Effective visualization helps users in analyzing and reasoning about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look-up a specific measure of a variable, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Data visualization is both an art and a science. The rate at which data is generated has increased, driven by an increasingly information-based economy. Data created by internet activity and an expanding number of sensors in the environment, such as satellites and traffic cameras, are referred to as ["Big Data"](http://en.wikipedia.org/wiki/Big_data). Processing, analyzing and communicating this data present a variety of ethical and analytical challenges for data visualization. The field of [data science](http://en.wikipedia.org/wiki/Data_science) and practitioners called [data scientists](http://en.wikipedia.org/wiki/Data_scientists) have emerged to help address this challenge.

In this project, we have used the Tableau visualization tool. Our first visualization shows the state in US which has mst number of restaurants. Second visualization shows the restaurants in Arizona state and the correlation between sentimental value and the star rating. Third visualization shows the count of restaurants in US and the last visualization shows the location of restaurants in Arizona state and their star ratings.

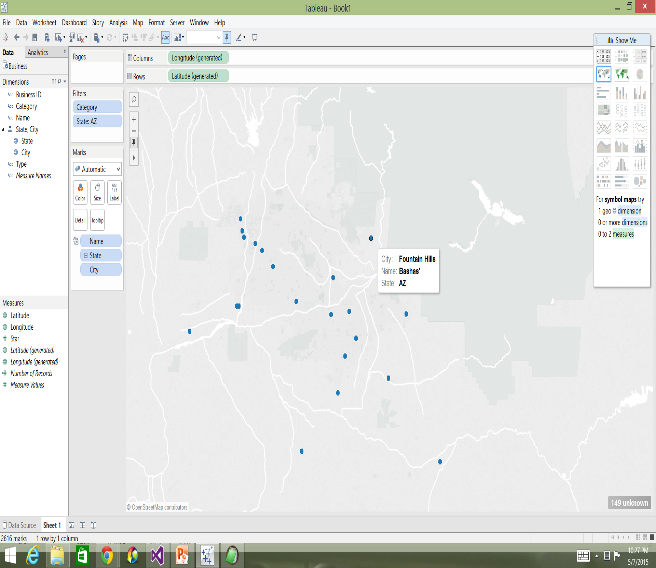


Fig 4: visualization 1

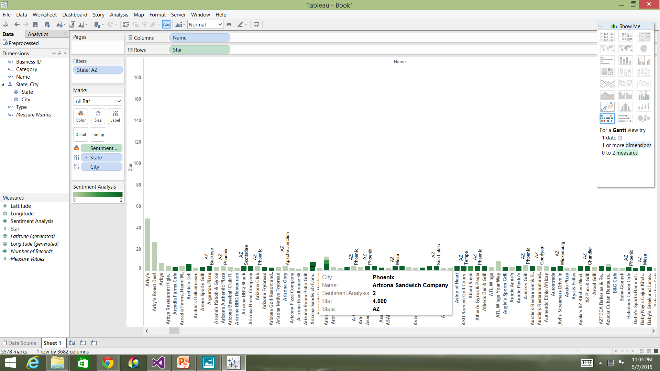


Fig 5: visualization 2

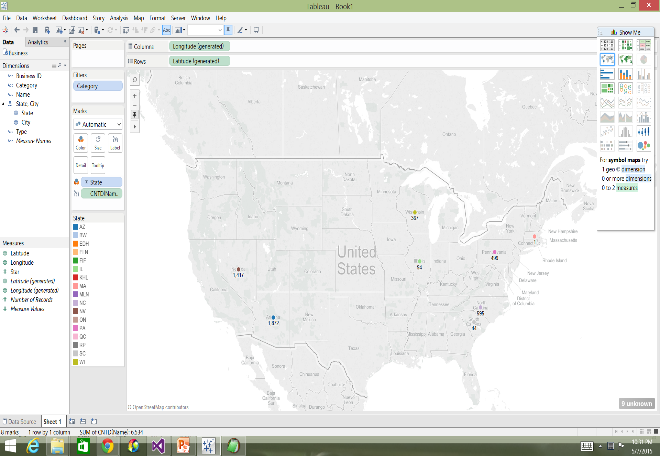


Fig 6: visualization 3

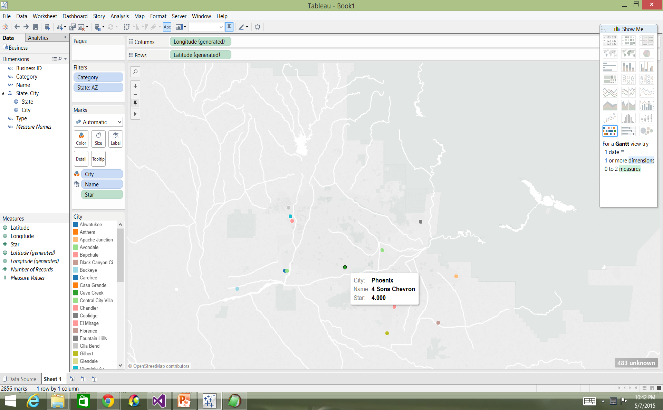


Fig 7: visualization 4

**Conclusion**

This project provides the overall sentiment for a business in the Yelp data set. The interactive web page helps the users to choose the business based on the review sentiment. Visualizations give a high level view of the data provided. Further work can be made on visualization in order to help the users make better decisions on the business they want to opt for.

**References**

<http://en.wikipedia.org/wiki/Yelp>

<http://www.yelp.com/dataset_challenge>

<http://en.wikipedia.org/wiki/Data_pre-processing>

<http://en.wikipedia.org/wiki/Logistic_regression>

<http://en.wikipedia.org/wiki/Support_vector_machine>

<http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html>

<http://en.wikipedia.org/wiki/Data_visualization>