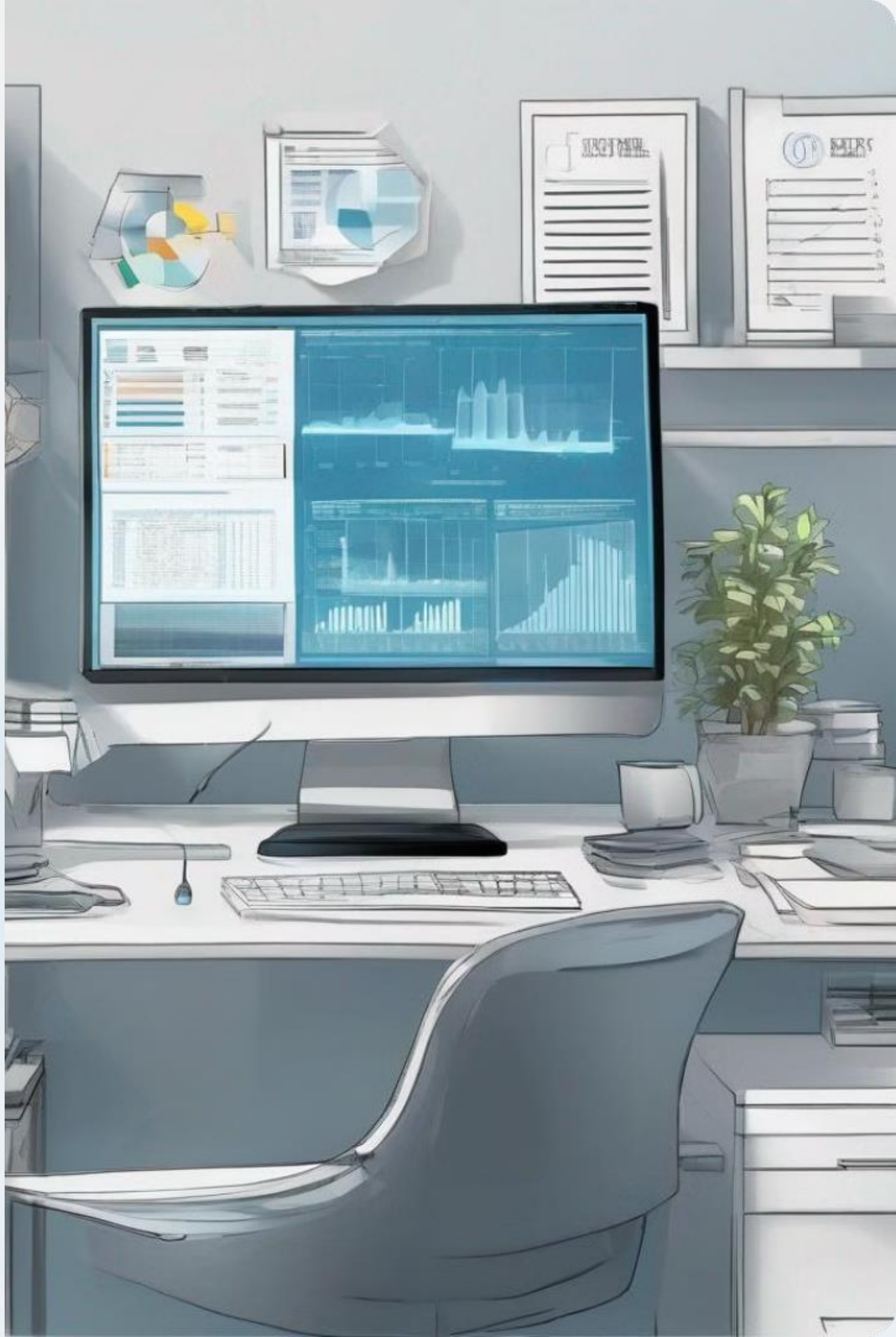
Web Mining: An Introduction

DATA MINING COURSE (SPRING 2024)

Dr. Keivan Borna

by Fatemeh Barati

Web mining is the process of extracting valuable information and insights from the vast amounts of data available on the World Wide Web. It encompasses a wide range of techniques and tools to analyze web content, structure, and user behavior, ultimately leading to better decision-making and strategic planning.



Types of Web Mining

Web Content Mining

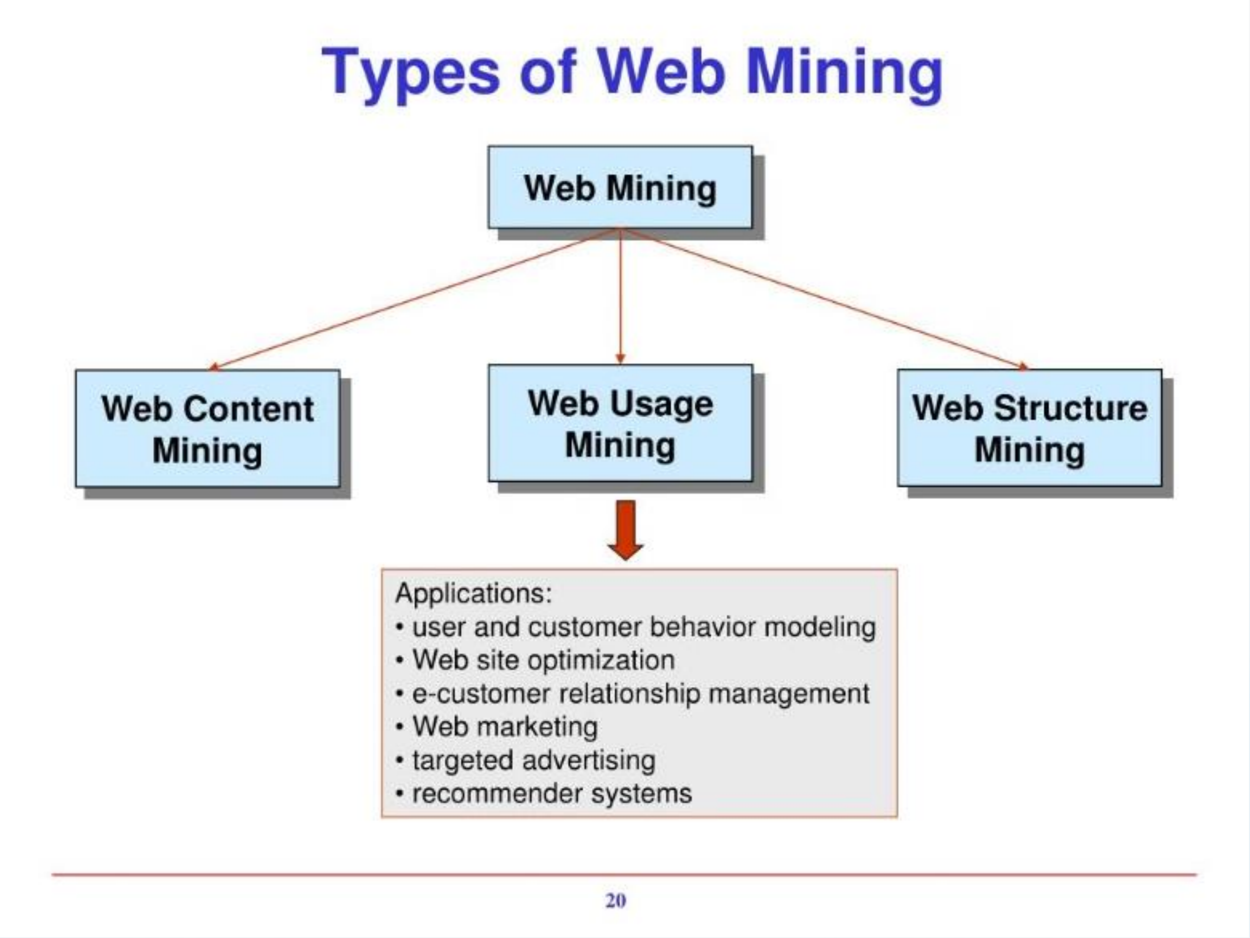
Focuses on extracting information from the actual content of content of web pages, such as text, images, and multimedia. multimedia.

Web Structure Mining

Analyzes the structure of websites and the relationships between web pages, such as hyperlinks and site hierarchies.

Web Usage Mining

Examines the patterns and trends in user behavior, including navigation, click-through data, and search queries.



Types of Web Mining

Applications:
- user and customer behavior modeling
- Web site optimization
- e-customer relationship management
- Web marketing
- targeted advertising
- recommender systems

20

Web Content Mining

1

Text Extraction

Identifying and extracting relevant text from web pages, such as article content, product descriptions, and user reviews.

2

Multimedia Analysis

Analyzing and understanding images, videos, and and other multimedia content on websites.

3

Information Retrieval

Developing algorithms to search, index, and rank rank web content based on relevance and user user intent.

4

Topic Modeling

Discovering and organizing web content into meaningful topics and categories.

Web Structure Mining

1

Link Analysis

Examining the structure and relationships between web web pages based on their hyperlinks.

2

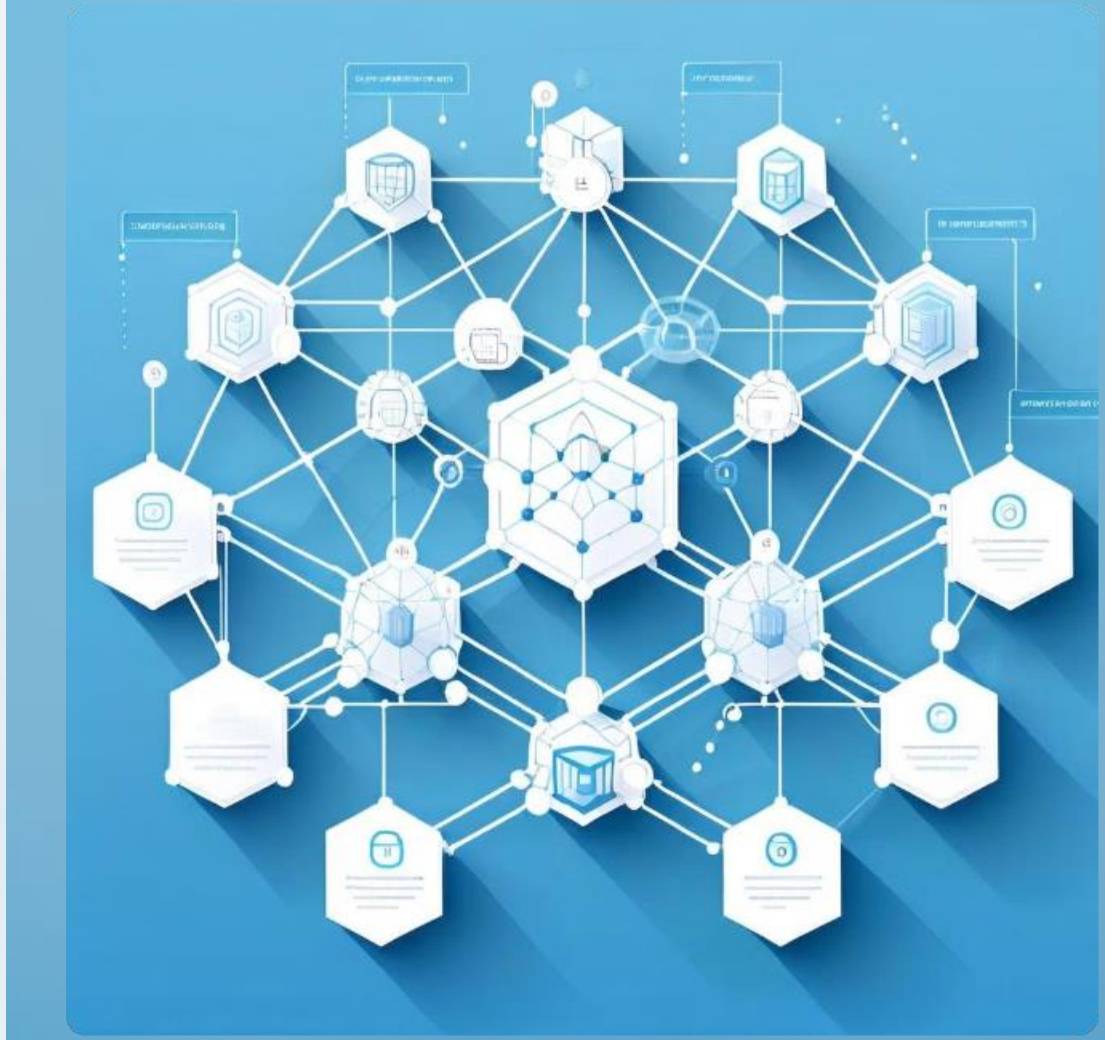
Site Taxonomy

Organizing and categorizing websites and web pages based pages based on their structural characteristics.

3

Community Detection

Identifying groups of related web pages or websites that that share common features or interests.



Web Usage Mining

User Behavior Analysis

Studying user interactions, such as page views, clicks, and navigation patterns, to understand understand their interests and preferences.

Personalization and Recommendations

Developing personalized recommendations and content based on individual user behavior and preferences.

Session and Clickstream Analysis

Analyzing user sessions and clickstream data to identify trends, patterns, and anomalies in user anomalies in user behavior.

Predictive Modeling

Using machine learning algorithms to predict user actions, such as purchases, churn, or content or content engagement.

Algorithms for Web Mining

PageRank

A graph-based algorithm used by by search engines to rank web pages pages based on their importance importance and authority.

Frequent Pattern Mining

Algorithms that identify frequently frequently occurring patterns and and associations in web data, such such as user behavior and content content co-occurrence.

Clustering Algorithms

Grouping web pages or users into into clusters based on their similarities, often used for segmentation and personalization. personalization.

Web Crawling and Scraping

1

Web Crawling

Automatically traversing the web web and following hyperlinks to to discover and index web pages. pages.

2

Data Extraction

Extracting and parsing relevant data data from web pages, such as structured content, unstructured unstructured text, and multimedia. multimedia.

3

Storage and Processing

Storing the extracted data in a structured format and performing performing further analysis and and processing.

Web Mining with Python

Web Scraping

Using libraries like BeautifulSoup and Scrapy to extract extract data from websites.

Data Analysis

Analyzing the extracted data using libraries like NumPy, NumPy, Pandas, and Matplotlib.

Machine Learning

Applying machine learning algorithms from libraries like libraries like scikit-learn and TensorFlow for web mining mining tasks.

Visualization

Creating interactive visualizations and dashboards using dashboards using libraries like Plotly and Matplotlib. Matplotlib.

Challenges and Limitations

1

Data Quality and Reliability

Web data can be noisy, inconsistent, and may contain errors, requiring careful data cleaning and preprocessing.

2

Privacy and Ethics

Web mining techniques must be used responsibly and in compliance with privacy regulations and ethical guidelines. guidelines.

3

Scalability and Performance

Handling the massive scale and volume of web data can be computationally intensive, requiring efficient algorithms and algorithms and infrastructure.

4

Dynamic and Evolving Web

The constantly changing nature of the web requires web mining systems to adapt and update their models and techniques and techniques continuously.

Applications and Future Trends

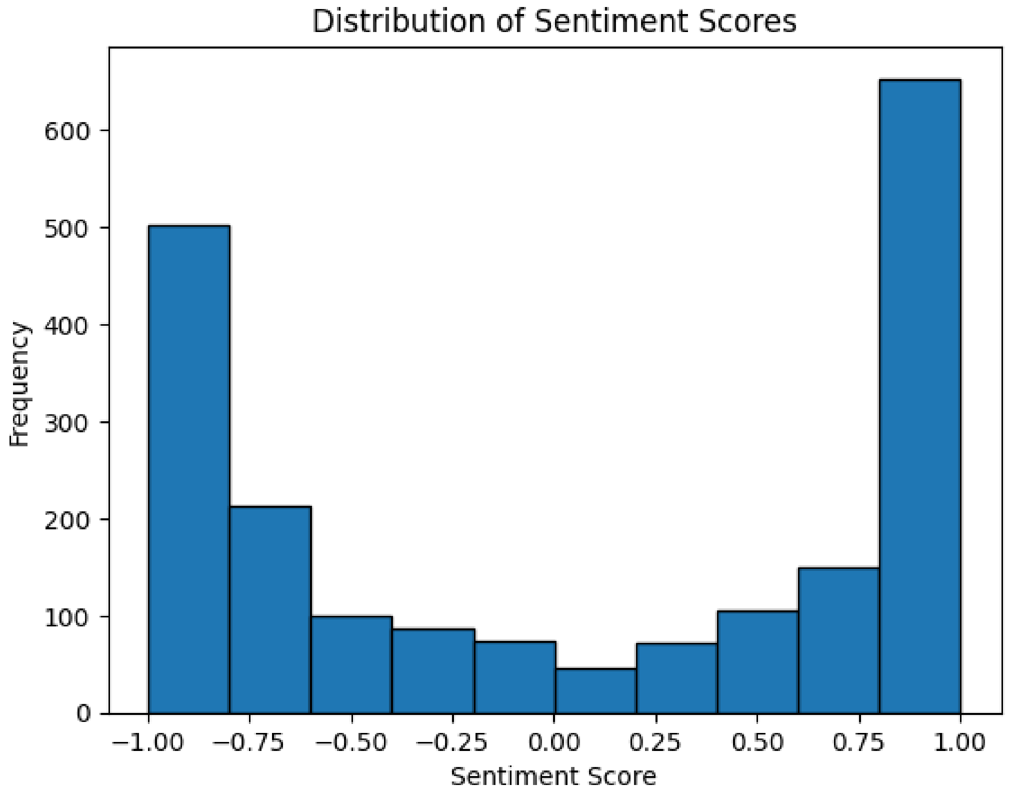
Search Engine Optimization	Understand user behavior and optimize website content content and structure to improve search engine rankings. rankings.
Personalized Recommendations	Leverage web usage data to provide personalized product, content, and service recommendations.
Predictive Analytics	Use web mining techniques to predict user actions, trends, trends, and market changes.
Sentiment Analysis	Analyze user-generated content, such as reviews and social and social media posts, to understand public sentiment. sentiment.

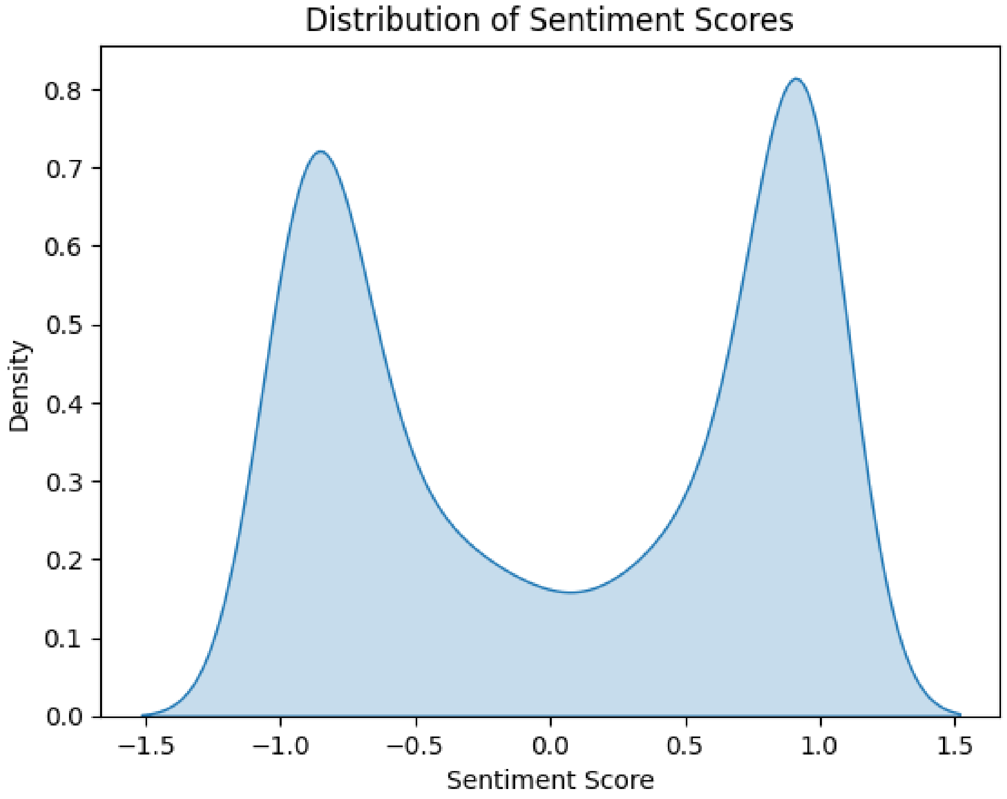
As the web continues to evolve, web mining techniques will become increasingly important for businesses and organizations to gain valuable insights, improve decision-making, and stay competitive in the digital landscape.

Scrape data from the web and analyze data

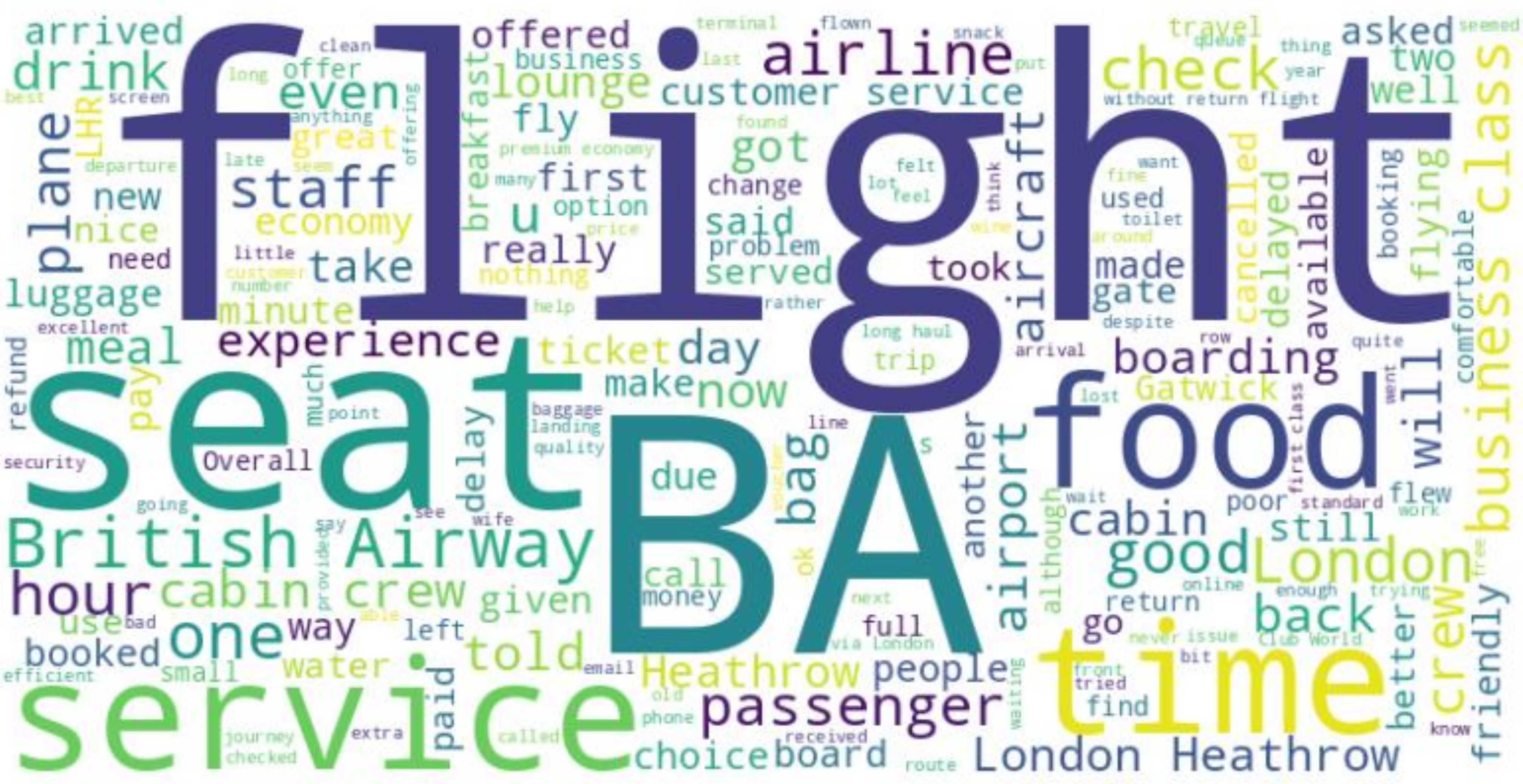
I scraped review data from the British Airways web and analyzed it. Here is a summary of the results:

- By Sentiment Analysis, I realized that 1024 reviews were positive, 8 were neutral and 968 were negative, out of 2000 reviews.





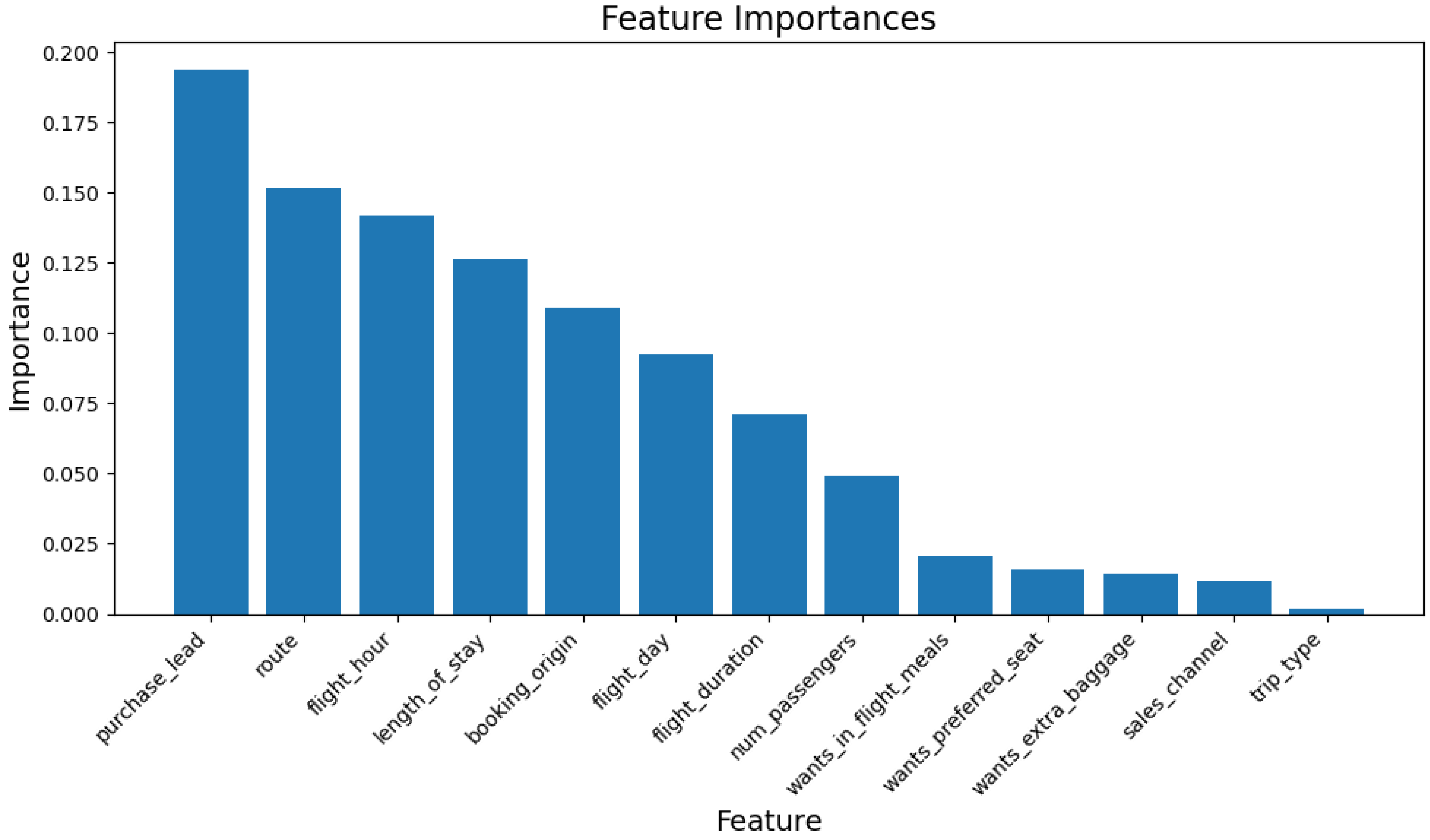
- Then, I visualized a WordCloud that represents the frequency or the importance of each word in the data.



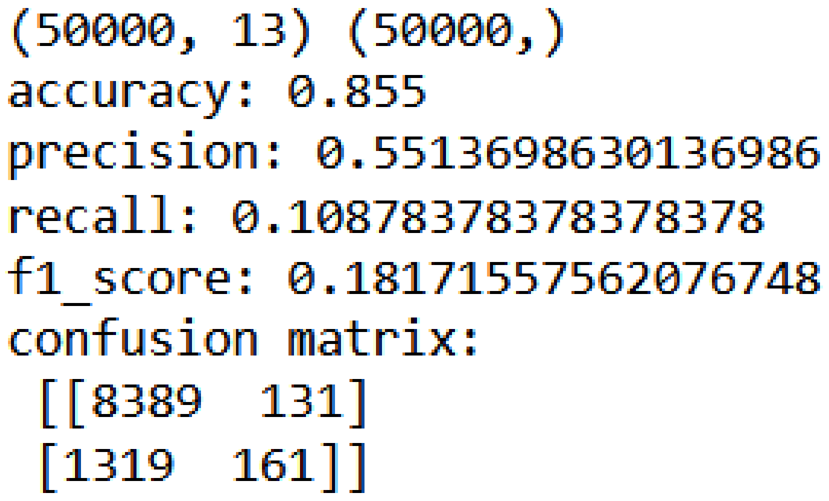
Explore and prepare the dataset, Train a machine learning model, and Evaluate model

First, I cleaned and prepared the data, then, I trained a RandomForest model and evauated it using different metrics. Here is a summary of the results:

- I sorted the DataFrame by importance in descending order.

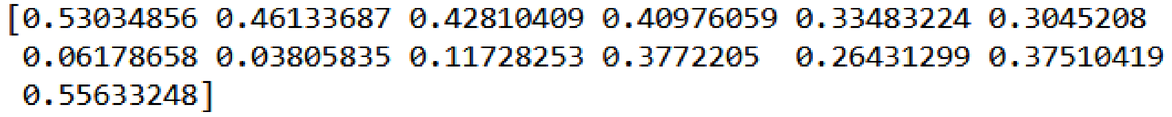


- With RandomForest model, I got these results:



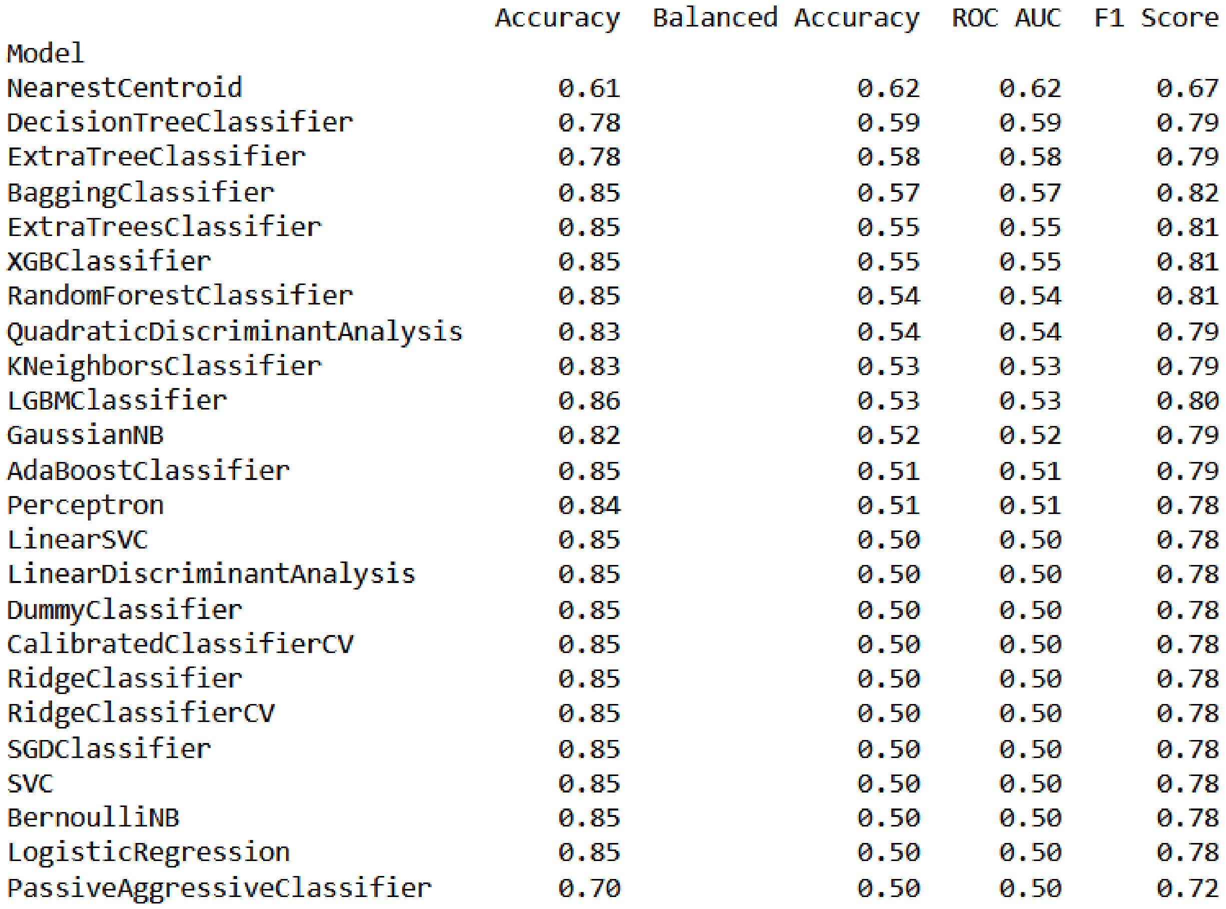
```
(50000, 13) (50000,)
accuracy: 0.855
precision: 0.5513698630136986
recall: 0.10878378378378378
f1_score: 0.18171557562076748
confusion matrix:
 [[8389  131]
 [1319  161]]
```

And here is the Cross Validation Score:



```
[0.53034856 0.46133687 0.42810409 0.40976059 0.33483224 0.3045208
 0.06178658 0.03805835 0.11728253 0.3772205  0.26431299 0.37510419
 0.55633248]
```

And these are the results of predictions with other models:



```
                               Accuracy  Balanced Accuracy  ROC AUC  F1 Score
Model
NearestCentroid                    0.61               0.62     0.62      0.67
DecisionTreeClassifier             0.78               0.59     0.59      0.79
ExtraTreeClassifier                0.78               0.58     0.58      0.79
BaggingClassifier                  0.85               0.57     0.57      0.82
ExtraTreesClassifier               0.85               0.55     0.55      0.81
XGBClassifier                      0.85               0.55     0.55      0.81
RandomForestClassifier             0.85               0.54     0.54      0.81
QuadraticDiscriminantAnalysis      0.83               0.54     0.54      0.79
KNeighborsClassifier               0.83               0.53     0.53      0.79
LGBMClassifier                     0.86               0.53     0.53      0.80
GaussianNB                         0.82               0.52     0.52      0.79
AdaBoostClassifier                 0.85               0.51     0.51      0.79
Perceptron                         0.84               0.51     0.51      0.78
LinearSVC                          0.85               0.50     0.50      0.78
LinearDiscriminantAnalysis         0.85               0.50     0.50      0.78
DummyClassifier                    0.85               0.50     0.50      0.78
CalibratedClassifierCV             0.85               0.50     0.50      0.78
RidgeClassifier                    0.85               0.50     0.50      0.78
RidgeClassifierCV                  0.85               0.50     0.50      0.78
SGDClassifier                      0.85               0.50     0.50      0.78
SVC                                0.85               0.50     0.50      0.78
BernoulliNB                        0.85               0.50     0.50      0.78
LogisticRegression                 0.85               0.50     0.50      0.78
PassiveAggressiveClassifier        0.70               0.50     0.50      0.72
```

References

- https://www.theforage.com/simulations/british-airways/data-science-yqoz
- https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition
- https://github.com/mikhailklassen/Mining-the-Social-Web-3rd-Edition

Thanks for your attention