**Module 5: Data Visualization, and Exploration**

**Syllabus:**

| **Lecture no** | **Content** | **Duration (Hr)** | **Self-Study (Hrs)** |
| --- | --- | --- | --- |
| 1 | Processing and Visualizing | 1 | 1 |
| 2 | Data, Influence Maximization | 1 | 1 |
| 3 | Link Prediction | 1 | 1 |
| 4 | Collective Classification | 1 | 1 |
| 5 | Applications in Advertising and Game Analytics (Use of tools like Unity30 / PyCharm). | 1 | 1 |
| 6 | Applications in Advertising and Game Analytics (Use of tools like Unity30 / PyCharm). | 1 | 1 |
| 7 | Introduction to Python Programming | 1 | 1 |
| 8 | Collecting and analyzing social media data | 1 | 1 |
| 9 | Visualization and exploration | 1 | 1 |

**Theoretical Background:**

* Define Your Objectives:
  + Objective Setting: Clearly define the goals and objectives of your social media analysis. Determine what specific insights you want to gain from the data, such as user sentiment, trends, or user interactions.
* Choose Social Media Platforms:
  + Platform Selection: Choose the appropriate social media platforms based on your objectives. Different platforms cater to different types of content and user demographics.
* Data Collection:
  + API Access: Utilize APIs provided by social media platforms to access data programmatically. APIs offer structured and controlled access to the platform's data.
  + Web Scraping: In cases where APIs are limited, web scraping can be used cautiously to extract data from social media pages. Be mindful of the platform's terms of service and legal constraints.
  + Third-Party Tools: Explore third-party tools that simplify data collection and provide user-friendly interfaces for querying and collecting data.
* Data Storage:
  + Structured Storage: Store collected data in structured formats such as databases (SQL or NoSQL) for efficient retrieval and analysis.
* Data Preprocessing:
  + Cleaning and Normalization: Cleanse the data by removing irrelevant information, duplicates, special characters, and handling missing data. Normalize text data to ensure consistency.
* Sentiment Analysis:
  + Sentiment Classification: Utilize natural language processing techniques to classify social media posts' sentiment into categories like positive, negative, or neutral. Algorithms like TextBlob or machine learning models can be used for sentiment analysis.
* Text Analysis:
  + NLP Techniques: Apply natural language processing techniques for deeper analysis, including topic modeling, keyword extraction, and text summarization. These methods unveil insights within the textual data.
* Visualization:
  + Visual Representation: Create visualizations to represent the processed data effectively. Visualization tools like Matplotlib, Seaborn, or Tableau can help in creating charts, graphs, and other visual representations.
* Trend Analysis:
  + Temporal Analysis: Analyze trends over time by tracking keywords or hashtags' frequency. Time-series analysis helps in identifying patterns and changes in user discussions and interests.
* User Engagement Analysis:
  + Engagement Metrics: Analyze user engagement metrics such as likes, shares, and comments to understand which content resonates most with the audience. Analyzing engagement data provides insights into user behavior.
* Reporting and Insights:
  + Insight Communication: Compile findings into comprehensive reports or presentations. Visualization aids in conveying complex information clearly. Insights should be communicated effectively to stakeholders.
* Continuous Monitoring:
  + Real-time Analysis: Set up processes for continuous data collection and real-time analysis. Social media data is dynamic, and continuous monitoring ensures that insights remain relevant and up-to-date.

**Key Definitions:**

Social Media Analytics: The process of collecting, analyzing, and interpreting social media data to extract meaningful insights and patterns.

Data Mining: The process of discovering patterns, trends, and insights from large datasets.

Sentiment Analysis: Also known as opinion mining, it involves determining and extracting the sentiment (positive, negative, neutral) expressed in social media posts or comments.

API (Application Programming Interface): A set of rules and protocols that allows one software application to interact with another.

Web Scraping: The automated method of extracting large amounts of data from websites quickly.

Machine Learning: A subset of artificial intelligence that enables systems to learn and improve from experience without being explicitly programmed.

Natural Language Processing (NLP): A field of AI that focuses on the interaction between computers and humans through natural language.

Data Visualization: The graphical representation of information and data. Tools like charts, graphs, and dashboards are used in social media analytics to visually present insights, making complex data more understandable.

Big Data: Extremely large and complex datasets that traditional data processing applications can't handle efficiently.

Predictive Analytics: The use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data.

Pattern Recognition: The process of recognizing patterns in data, often used in machine learning.

Dashboard: A visual tool that displays key performance indicators, metrics, and data points in a single interface.

**Lecture 1**

**Learning Objective: Learners shall be able to understand the processing and Visualizing Data**

**Processing and Visualizing Data**

### **Process**

The process of understanding data begins with a set of numbers and a question. The following steps form a path to the answer:

**Acquire**

Obtain the data, whether from a file on a disk or a source over a network.

**Parse**

Provide some structure for the data’s meaning, and order it into categories.

**Filter**

Remove all but the data of interest.

**Mine**

Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.

**Represent**

Choose a basic visual model, such as a bar graph, list, or tree.

**Refine**

Improve the basic representation to make it clearer and more visually engaging.

**Interact**

Add methods for manipulating the data or controlling what features are visible.

Refer link: <https://learning.oreilly.com/library/view/visualizing-data/9780596514556/ch01.html#what_is_the_question_question-id1>

**Visualization:**

Each graph is created to present data in a specific way, so it’s only logical that some of them are better than others. Take a look at the different types of graphs and their common usage:

• Scatter plot – this type of graph shows a correlation.

• Pie chart – this one shows proportions. However, make sure you don’t use it when you have more than 5 slivers. In that case, use another graph.

• Line graph – use this graph to present trends and patterns.

• Table – this type of graph is perfect for presenting specific values.

• Bar chart – this one is excellent for showing comparisons.

For example, a pie chart is great for presenting how the budget for an ad campaign on social media is being used. This type of graph shows how much money is being spent and where.

**Refer Website:** <https://whatagraph.com/>

*Let’s check the take away from this lecture*

Exercise:

1. How does the number of hours studied correlate with exam scores?
2. What proportion of the company's revenue comes from different product categories?
3. How do the average temperatures vary across different months of the year?

**Learning from this lecture:** Learners should be able to understand the processing and Visualizing Data.

**Lecture 2**

**Learning Objective: Learners shall be able to understand the concept of Influence Maximization**

**Influence Maximization**

Maximizing influence in social networks can be explained in a broader sense. To begin, consider a social network as a directed graph with users as nodes and their connections to one another as directed edges. This product is promoted throughout the [social network graph](https://www.sciencedirect.com/topics/engineering/social-network-graph) via the word-of-mouth effect, which is the effect of human-to-human transmission. Promotion can result in either rapid decline or rapid growth. These two outcomes depend on three key questions.

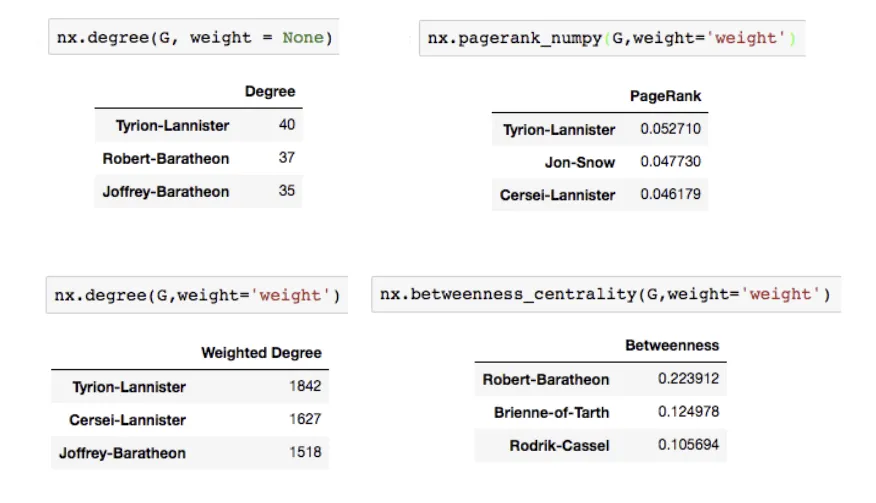
1. The first question is how to choose the people who are initially given the right to try the product. This is the seed node set S mentioned later.
2. The second question is, how does this small group of people get other people to be influenced to buy the product? In fact, this process is the capability of different [diffusion models](https://www.sciencedirect.com/topics/computer-science/diffusion-model).
3. The last question is how to estimate the number of users affected by this small group of people, i.e., the most influential people.

This is actually the problem of estimating the propagation influence of a given seed set. According to the three problems mentioned above, the influence maximization problem in social networks can be defined as finding the seed set composed of the smallest number of nodes in the social network so that they can influence the largest number of users in the social network; that is, influence maximization.

**The influence maximization problem**

Describes a marketing (but not only) setup, where the goal of the marketer is to select a limited set of nodes in the network (seeding set) such that will naturally spread the influence to as many nodes as possible. For example, consider inviting a limited number of influencers to a prestigious product launch event, in order to spread the word to the rest of their network.

Such influencers can be identified with numerous techniques, such as using the centrality measures we’ve mentioned above. Here are the most central nodes in Game of Thrones network, according to different measures:



As we can see, some of the characters re-occur at the top of different measures, and are also well known for their social influence in the show.

By simulating the selection of most central nodes we observe that picking a single node of the network can achieve about 50% of network coverage — That’s how important social influencers might be.

On the other hand, Influence Maximization is Hard. In fact it’s considered an NP-Hard problem. Many heuristics were developed to find the best seeding set in an efficient calculation. Trying a brute-force method to find the best seeding couple in our network resulted in spending 41 minutes and achieving 56% of coverage, a result that would be hard to achieve with centrality heuristics.

**Wrap-up**

Network analysis is a complex and useful tool for various domains, in particular in the rapidly growing social networks. The applications of such analysis include marketing influence maximization, fraud detection or recommender systems. There are multiple tools and techniques that can be applied on network datasets, but they need to be chosen wisely, taking into account the problem’s and the network’s unique properties.

Here's a deeper dive into Influence Maximization:

**Networks:** Influence Maximization is often applied to networks, which could be social networks, communication networks, or any other system where nodes (representing individuals, entities, or points) are connected by edges (representing relationships, interactions, or links).

**Influence Propagation:** The goal of Influence Maximization is to identify a subset of nodes (usually a small number) in the network in such a way that when they are influenced (e.g., by adopting a behavior or idea), the influence spreads most effectively throughout the entire network. This is often modeled as a process where influence "flows" from influenced nodes to their neighbors, who may, in turn, influence others.

**Virality:** Influence Maximization is particularly relevant in contexts where "virality" matters, such as viral marketing or the spread of information on social media. It helps answer questions like, "Who should we target with our marketing campaign to maximize its reach?"

**Complexity:** Identifying the optimal set of nodes for Influence Maximization is computationally challenging, especially in large networks. Various algorithms and heuristics are used to approximate the solution because finding the absolute best solution can be time-consuming.

**Applications:** Influence Maximization has applications beyond marketing and social networks. It's used in epidemiology to identify key individuals for vaccination campaigns, in recommendation systems to suggest products or content to users, and in understanding the dynamics of various networked systems.

**Metrics:** Common metrics for evaluating Influence Maximization strategies include the spread of influence (how many nodes were influenced), the size of the seed set (the number of initial influencers), and various measures of influence, such as the number of new adoptions or the change in network structure.

In summary, Influence Maximization is a complex problem that aims to identify the most influential individuals or nodes in a network to maximize the spread of influence, information, or behaviors. It's a valuable concept in various fields, particularly in marketing, social network analysis, and epidemiology, where understanding and harnessing network effects are critical.

*Let’s check the take away from this lecture*

Exercise:

1. Summarize the methods used for approximating solutions in Influence Maximization.
2. Apply Influence Maximization principles to suggest an approach for maximizing the reach of a social media campaign.?
3. Critically assess the effectiveness of different metrics used to evaluate Influence Maximization strategies.

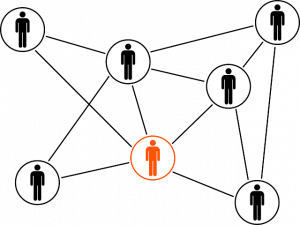
**Learning from this lecture:** Learners should be able to understand the concept of Influence Maximization

**Lecture 3**

**Learning Objective:** **Learners shall be able to understand how link prediction occurs.**

**Link Prediction**

Link prediction is one of the most important research topics in the field of graphs and networks. The objective of link prediction is to identify pairs of nodes that will either form a link or not in the future.

****

Link prediction has a ton of use in real-world applications. Here are some of the important use cases of link prediction:

* Predict which customers are likely to buy what products on online marketplaces like Amazon. It can help in making better product recommendations
* Suggest interactions or collaborations between employees in an organization
* Extract vital insights from terrorist networks

Strategy to Solve a Link Prediction Problem

Overview

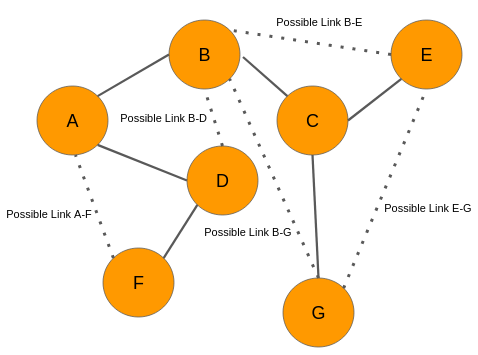
* An introduction to link prediction, how it works, and where you can use it in the real-world
* Learn about the importance of Link Prediction on social media
* Build your first Link Prediction model for a Facebook use case using Python

## 

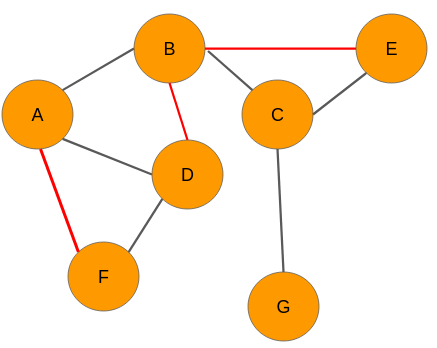
## **Strategy to Solve a Link Prediction Problem**

If we can somehow represent a graph in the form of a structured dataset having a set of features, then maybe we can use [machine learning](https://courses.analyticsvidhya.com/courses/applied-machine-learning-beginner-to-professional?utm_source=blog&utm_medium=link-prediction-how-to-predict-your-future-connections-on-facebook) to predict the formation of links between the unconnected node-pairs of the graph.tcou

Let’s take a dummy graph to understand this idea. Given below is a 7 node graph and the unconnected node-pairs are AF, BD, BE, BG, and EG:



Now, let’s say we analyze the data and come up with the below graph. A few new connections have been formed (links in red):



We need to have a set of predictor variables and a target variable to build any kind of machine learning model, right? So where are these variables? Well, we can get it from the graph itself! Let’s see how it is done.

Our objective is to predict whether there would be a link between any 2 unconnected nodes. From the network at time t, we can extract the following node pairs which have no links between them:

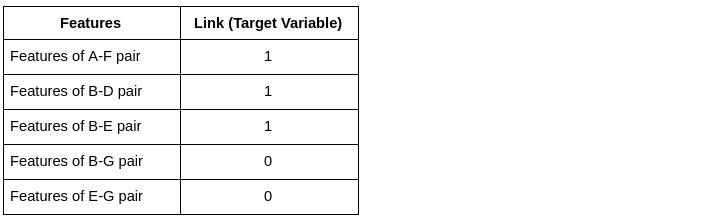
* A-F
* B-D
* B-E
* B-G
* E-G

*Please note that, for convenience, we have considered only those nodes that are a couple of links apart.*

The next step for us is to create features for each and every pair of nodes. The good news is that there are several techniques to extract features from the nodes in a network. Let’s say we use one of those techniques and build features for each of these pairs. However, we still don’t know what the target variable is. Nothing to worry about – we can easily obtain that as well.

Look at the graph at time *t+n*. We can see that there are three new links in the network for the pairs A-F, B-D, and BE respectively. Therefore, we will assign each one of them a value of 1. The node pairs B-G and E-G will be assigned 0 because there are still no links between the nodes.

Hence, the data will look like this:



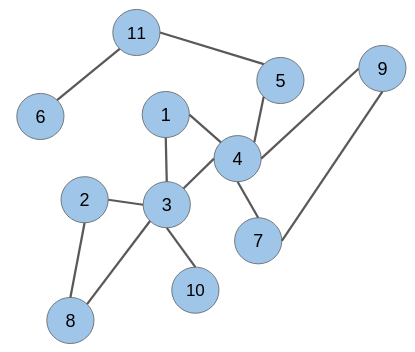
Now that we have the target variable, we can build a machine learning model using this data to perform link prediction.

So, this is how we need to use social graphs at two different instances of time to extract the target variable, i.e., the presence of a link between a node pair. Keep in mind, however, that in real-world scenarios, we will have data of the present time only.

### Extract data from a Graph for Building your Model

In the section above, we were able to get labels for the target variable because we had access to the graph at time t+n. However, in real-world scenarios, we would have just one graph dataset in hand. That’s it!

Let’s say we have the below graph of a social network where the nodes are the users and the edges represent some kind of relationship:



The candidate node pairs, which may form a link at a future time, are (1 & 2), (2 & 4), (5 & 6), (8 & 10), and so on. We have to build a model that will predict if there would be a link between these node pairs or not. This is what link prediction is all about!

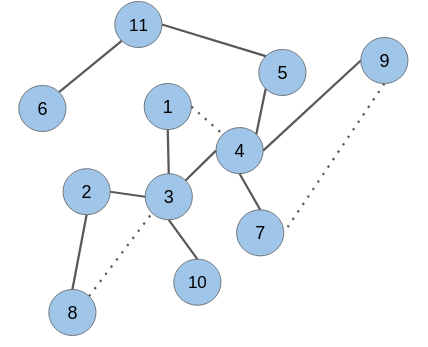
However, to build a link prediction model, we need to prepare a training dataset out of this graph. It can be done using a simple trick.

Picture this – how would this graph have looked at some point in the past? There would be fewer edges between the nodes because connections in a social network are built gradually over time.

Hence, keeping this in mind, we can randomly hide some of the edges from the given graph and then follow the same technique as explained in the previous section to create the training dataset.

Strike Off Links from the Graph

While removing links or edges, we should avoid removing any edge that may produce an isolated node (node without any edge) or an isolated network. Let’s take off some of the edges from our network:



As you can see, the edges in the node pairs (1 & 4), (7 & 9), and (3 & 8) have been removed.

Add labels to extracted data

Next, we would need to create features for all the unconnected node pairs including the ones for which we have omitted the edges. The removed edges will be labeled as ‘1’ and the unconnected node pairs as ‘0’:

| Features | Link (Target Variable) |
| --- | --- |
| Features of node pair 1 – 2 | 0 |
| Features of node pair 1 – 5 | 0 |
| Features of node pair 1 – 7 | 0 |
| Features of node pair 1 – 8 | 0 |
| Features of node pair 1 – 9 | 0 |
| Features of node pair 1 – 10 | 0 |
| Features of node pair 2 – 4 | 0 |
| Features of node pair 2 – 10 | 0 |
| Features of node pair 3 – 5 | 0 |
| Features of node pair 3 – 7 | 0 |
| Features of node pair 3 – 9 | 0 |
| Features of node pair 4 – 8 | 0 |
| Features of node pair 4 – 10 | 0 |
| Features of node pair 4 – 11 | 0 |
| Features of node pair 5 – 6 | 0 |
| Features of node pair 5 – 7 | 0 |
| Features of node pair 5 – 9 | 0 |
| Features of node pair 8 – 10 | 0 |
| Features of node pair 1 – 4 | 1 |
| Features of node pair 3 – 8 | 1 |
| Features of node pair 7 – 9 | 1 |

It turns out that the target variable is highly imbalanced. This is what you will encounter in real-world graphs as well. The number of unconnected node pairs would be huge.

*Let’s check the take away from this lecture*

Exercise:

1. What is the primary objective of link prediction in the context of social networks and graphs?.
2. Describe the importance of creating a training dataset for link prediction. Why is it necessary to remove certain edges during this process?
3. Imagine a social media platform. How could link prediction be utilized to enhance user experience on this platform?

**Learning from this lecture:** Learners should be able to understand link prediction and solve link prediction problems

**Lecture 4**

**Learning Objective: Learners shall be able to understand the concept of collective Classification**

**Collective Classification**

Classification is one of the most studied subjects in machine learning. Most classification methods that were developed this last decade either account for structure (interactions, relationships) or attributes (text, numerical, etc). This leads to ignoring significant patterns in a dataset that could only be captured by analyzing the features of an item and its interactions. Collective classification methods use both structure and attributes, often by aggregating data from neighbors of a node and learning a model on the aggregated data. In social networks, the degree distribution of nodes follows a power law where few nodes have many neighbors. High degree nodes have incoming links from low degree nodes of different classes and many nodes have very few edges. Hence, using only local structure may lead to poor predictions. Also, many social networks allow for different types of interactions (retweet, reply, like, etc.) that affect classification differently.

**Example:**

Collective classification is a way to make predictions about things in a network, like people in a social network or items in an online store, by considering both the connections between them and their individual characteristics.

Here's a simple explanation:

Imagine you have a social network like Facebook. Some people have lots of friends, and some have only a few. Now, if you want to guess something about a person, like whether they like a particular movie, you could just look at their own profile. But that's not always enough information.

Collective classification says, "Let's not just look at one person. Let's also look at their friends." Why? Because sometimes, what your friends do or like can tell us something about you. For example, if many of your friends are big fans of a movie, there's a good chance you might like it too.

But it's not just about counting friends; it's also about considering different kinds of interactions. In a social network, people can do various things like liking posts, sharing pictures, or commenting. Each of these actions might give us clues about what a person is interested in.

So, collective classification combines information from a person's own profile, what their friends are like, and how they interact with others to make better predictions. It's like saying, "Let's look at the big picture and use all the information we have to guess what someone might be into or what they might do next." It's a more powerful way of making predictions, especially in networks where different types of interactions matter.

**Learning from this lecture:** Learners should be able to understand link prediction and solve link prediction problems

*Let’s check the take away from this lecture*

Exercise:

1. Understand how different types of interactions in social networks contribute to collective classification accuracy.
2. Apply collective classification principles to suggest a method for predicting user preferences in a social media platform.
3. Analyze the challenges of incorporating various types of interactions (likes, comments, shares) into a collective classification model.

**Lecture 5**

**Learning Objective: Learners shall be able to apply the knowledge in advertising and Game Analytics with the help of tools like Unity 30**

**Applications in Advertising and Game Analytics (Use of tools like Unity30 / PyCharm)**

Unity is a versatile game development platform, and it can be used for various applications in advertising and game analytics. Here are some ways Unity can be applied in these contexts:

**Applications in Advertising:**

**Interactive Ads**: Unity allows advertisers to create interactive and immersive ad experiences. For example, you can develop mini-games or 3D product showcases within Unity to engage users more effectively than traditional static ads.

**Virtual Product Demonstrations**: Advertisers can use Unity to build virtual simulations or demos of their products or services. This provides potential customers with a realistic experience before making a purchase decision.

**Augmented Reality (AR) Ad**s: Unity's AR capabilities enable advertisers to create AR ad campaigns. These ads can overlay digital content onto the real world through mobile devices, enhancing user engagement.

**Advergames**: Advergames are games created for advertising purposes. Unity is an excellent platform for developing advergames that promote brands or products while providing entertainment.

**User Engagement Analytics**: Unity's analytics tools allow advertisers to track user interactions within their ad experiences. Advertisers can measure user engagement, interaction duration, and other metrics to assess ad effectiveness.

**A/B Testing**: Unity can be used to create variations of ads or ad experiences for A/B testing. Advertisers can compare user engagement and conversion rates between different versions to optimize ad campaigns.

**Applications in Game Analytics:**

**Player Behavior Analysis**: Unity's analytics features provide insights into how players interact with games. You can analyze player behavior, progression, and retention rates to optimize game design.

**Monetization Strategies**: Unity's ad monetization platform can be integrated into games to earn revenue through in-game ads. Analytics can help measure ad performance and its impact on player engagement and revenue.

**Level Design and Balancing**: Game developers can use analytics to fine-tune level design, character balance, and difficulty progression. Data-driven decisions ensure a better gaming experience.

**Player Segmentation**: Unity analytics can segment players based on their behavior. Developers can tailor in-game experiences, offers, or content to different player groups.

**Predictive Analytics**: Advanced analytics can predict player behavior, such as churn prediction (identifying players likely to stop playing) or forecasting in-game item purchases.

**Multiplayer Game Analytics**: Unity supports multiplayer games, and analytics can be used to track player interactions, match outcomes, and overall game performance in a multiplayer environment.

**Performance Optimization**: Unity analytics can monitor game performance metrics like frame rate, memory usage, and crash reports. Developers can identify and resolve performance issues for a smoother gaming experience.

In summary, Unity is a powerful platform that can be applied in advertising to create interactive and immersive ad experiences and in game analytics to understand player behavior, optimize game design, and monetize games effectively. By leveraging Unity's capabilities and analytics tools, developers and advertisers can make data-driven decisions to enhance user engagement and achieve their goals.

*Let’s check the take away from this lecture*

Exercise:

1. Understand the concept of A/B testing in advertising and how Unity is utilized for creating different ad variations.
2. Propose a scenario where Unity's interactive ad features could be effectively used by a specific brand for product promotion.
3. Analyze the impact of augmented reality ads created with Unity on user engagement and brand visibility for a particular product.

**Learning from this lecture:** Learners shall be able to apply the knowledge in advertising and Game Analytics with the help of tools like Unity 30

**Lecture 6**

**Learning Objective:**  Learners shall be able to apply the knowledge in advertising and Game Analytics with the help of tools like Pycharm

PyCharm is an integrated development environment (IDE) for the Python programming language. While it is not a game development platform like Unity, it can still play a role in advertising and game analytics when used in conjunction with other tools and libraries. Here's how PyCharm can be applied in these contexts:

**Applications in Advertising:**

**Data Analysis and Visualization**: PyCharm can be used to write Python scripts for data analysis and visualization. Advertisers can use PyCharm to process advertising data, create visualizations, and gain insights from advertising campaign performance.

**Machine Learning and Predictive Analytics**: PyCharm provides a convenient environment for developing machine learning models. Advertisers can use Python libraries like scikit-learn, TensorFlow, or PyTorch to build predictive models for customer behavior, ad click-through rates, or ad conversion rates.

**A/B Testing**: PyCharm can be used to write code for running A/B tests. Advertisers can design experiments, collect and analyze data, and make data-driven decisions about the effectiveness of different ad strategies.

**Ad Campaign Automation**: Python scripts developed in PyCharm can automate various aspects of ad campaign management, including bid optimization, ad budget allocation, and ad content generation.

**Social Media Data Analysis**: PyCharm can be employed to collect and analyze social media data relevant to advertising campaigns. Python libraries like Tweepy or praw can be used to access social media APIs.

**Applications in Game Analytics:**

**Game Data Analysis:** PyCharm can be used to write Python scripts for analyzing game-related data. Game developers can use it to process player telemetry data, gameplay logs, and other game-related metrics.

**Player Behavior Analysis**: Python scripts can be developed to analyze player behavior within games. PyCharm allows developers to explore player actions, progression, and engagement patterns.

**Content Management**: Game developers can use PyCharm to manage game content, including scripting game events, dialogue, and character behaviors. This is crucial for storytelling and gameplay design.

**Modding and Scripting**: PyCharm can be used for scripting and modding games. Game developers and modders can create custom game features, assets, or modifications using Python scripts.

**Game AI**: Python is often used to develop game AI components. PyCharm provides a coding environment for designing and implementing AI algorithms and behaviors.

**Player Feedback Analysis**: Python scripts developed in PyCharm can process and analyze player feedback from forums, reviews, and surveys, helping game developers make improvements.

While PyCharm is not a game development platform like Unity, it is a versatile tool for data analysis, scripting, and automation. It can complement game development and advertising analytics by providing a robust environment for coding, data processing, and modeling. Developers and analysts can use PyCharm to leverage the power of Python and its libraries for these applications.

*Let’s check the take away from this lecture*

Exercise:

1. Propose a scenario where PyCharm can be used to automate bid optimization in an online advertising campaign. Outline the steps and benefits of automation.
2. Describe the role of PyCharm in automating ad campaign management tasks. Provide examples of tasks that can be automated.
3. Analyze the advantages of using PyCharm for game AI development. How does Python scripting enhance game AI capabilities, and what challenges might developers face?

**Learning from this lecture:** Learners shall be able to apply the knowledge in advertising and Game Analytics with the help of tools like Pycharm

**Lecture 7**

**Learning Objective:** **Learners shall be able to understand basics of Python Programming**

Introduction to Python Programming:

Python is a versatile and widely-used programming language known for its simplicity and readability. It was created by Guido van Rossum and first released in 1991. Python is popular among beginners and experienced developers alike due to its easy-to-learn syntax and powerful capabilities. Here's an introduction to Python programming covering its key features, uses, and basic concepts:

Key Features of Python:

Simple and Readable: Python emphasizes simplicity and readability, which makes it a great language for beginners. Its clean and easy-to-understand syntax allows developers to express their ideas in fewer lines of code compared to other languages.

Interpreted Language: Python is an interpreted language, meaning that code is executed line by line. This makes it easier to debug and test code since errors can be identified as soon as they occur.

High-Level Language: Python is a high-level language, which means it abstracts low-level details like memory management and hardware interactions. This allows developers to focus on solving problems rather than dealing with complex system-specific issues.

Dynamically Typed: Python is dynamically typed, which means you don't need to declare the data type of a variable explicitly. The interpreter automatically determines the type during runtime.

Rich Standard Library: Python comes with a vast collection of libraries and modules that simplify tasks such as file I/O, data manipulation, and web development. These libraries save time and effort, enabling developers to accomplish complex tasks with minimal code.

Object-Oriented: Python supports object-oriented programming (OOP) principles, allowing developers to create classes and objects, encapsulate data, and implement inheritance and polymorphism.

Common Uses of Python:

Web Development: Python frameworks like Django and Flask are widely used for building dynamic and responsive websites and web applications.

Data Science and Analytics: Python is a popular choice for data analysis, data visualization, and machine learning. Libraries like NumPy, Pandas, Matplotlib, and scikit-learn facilitate these tasks.

Artificial Intelligence: Python is extensively used in AI research, natural language processing, and neural networks due to its simplicity and the availability of specialized libraries like TensorFlow and PyTorch.

Automation and Scripting: Python scripts automate repetitive tasks, making it a favorite for system administrators and DevOps engineers.

Game Development: Python is used for game development, both for prototyping and creating full-fledged games. Libraries like Pygame provide a foundation for building interactive games.

Scientific Computing: Python is employed in scientific and mathematical computations due to its numerical and scientific computing libraries.

**Learning from this lecture:** Learners understood the basics of python programming.

**Lecture 8**

**Learning Objective:**  Learners shall be able to understand how to collect and analyze social media data.

**Collecting and analyzing social media data**

Collecting and analyzing social media data is a valuable process for gaining insights into user behavior, sentiment, trends, and more. Here's a step-by-step guide on how to collect and analyze social media data:

**1. Define Your Objectives:**

Determine the goals of your analysis. What questions do you want to answer? Are you looking to understand user sentiment, track trends, or analyze user interactions?

**2. Choose Social Media Platforms:**

Select the social media platforms relevant to your objectives. Common platforms include Twitter, Facebook, Instagram, LinkedIn, Reddit, and YouTube.

**3. Data Collection:**

a. API Access: Most social media platforms provide Application Programming Interfaces (APIs) that allow developers to access data programmatically. You'll need to register for developer accounts on these platforms and follow their API documentation.

b. Web Scraping: If API access is limited or unavailable, web scraping tools and libraries like BeautifulSoup or Scrapy (for websites) can be used to extract data from social media pages. However, be sure to review the platform's terms of service to ensure compliance.

c. Third-Party Tools: There are third-party tools and services that simplify data collection from various social media platforms. These tools often provide user-friendly interfaces for querying and collecting data.

**4. Data Storage:**

Store the collected data in a structured format, such as a database (e.g., MySQL, PostgreSQL) or NoSQL database (e.g., MongoDB). Proper data storage ensures that you can efficiently retrieve and analyze the data.

**5. Data Preprocessing:**

Clean the collected data to remove irrelevant information, duplicates, and noise. This step involves text normalization, removing special characters, and handling missing data.

**6. Sentiment Analysis:**

Perform sentiment analysis to understand the sentiment (positive, negative, neutral) of social media posts or comments. Libraries like NLTK, TextBlob, or spaCy can help with sentiment analysis.

**7. Text Analysis:**

Use natural language processing (NLP) techniques to analyze the content of social media posts. You can perform tasks like topic modeling, keyword extraction, and text summarization to uncover insights.

**8. Visualization:**

Create visualizations (e.g., word clouds, bar charts, heatmaps) to represent the data and insights effectively. Tools like Matplotlib, Seaborn, or data visualization platforms like Tableau can help.

**9. Trend Analysis:**

Analyze trends over time by tracking the frequency of specific keywords or hashtags. This can help you identify popular topics or discussions.

**10. User Engagement Analysis:** - Analyze user engagement metrics, such as likes, shares, and comments, to understand which content is most engaging and what drives user interaction.

**11. Reporting and Insights:** - Compile your findings into a report or presentation. Clearly communicate the insights and trends you've discovered. Visualization is often a crucial part of conveying information effectively.

**12. Continuous Monitoring:** - Social media data is dynamic. Consider setting up processes for continuous data collection and analysis to keep your insights up-to-date.

*Let’s check the take away from this lecture*

Exercise:

1. Describe the process of web scraping for social media data collection. What are the limitations and ethical considerations associated with web scraping?
2. Understand the significance of trend analysis in social media data. How can tracking specific keywords help identify popular topics?
3. Design a hypothetical social media data collection strategy for analyzing user interactions on Twitter. Include the choice of platform, data collection methods, and storage techniques.

**Learning from this lecture:** Learners understood how to collect and analyze social media data.

**Lecture 9**

**Learning Objective:**  Learners shall be able to understand exploration and visualization of data.

As we all know that social media generates huge amounts of data; the explosive growth of social media is one of the reasons that 90% of data in the world has been generated in the last 3 years alone. When we use Social Media services like facebook and twitter; we let the companies store our data like photos, public observation, comments and/or communications. In exchange, they structure that information for our comfort and make it easy for us to access data of today as well as yesterday comfortably. Visualization is the technique that makes all this possible. Visualizing social networks in an interactive format offers faster and more accurate access to the network analysis. It helps us to group our relationships and friend groups by visualizing the concepts and links that were used.

**Visualization approaches:**

1. Slopegraphs  —  Slopegraphs are a special type of a line chart where two or more sets of values are compared by connecting each group’s values on one scale to their values on the second scale. The two scales have identical maximum and minimum values to make it very easy to see whether each group increases, decreases, or remains similar between the two categories: highlighted to show interest and graying out for disinterest.

2. Parallel Coordinates  —  A parallel coordinates graph shows multiple variables alongside one another with each scaled from highest to lowest value :highest at the top, lowest at the bottom and with lines connecting each entity’s position for each variable, horizontally across the graph.

3. Alluvial Diagrams  —  Alluvial diagrams show how various nodes flow together or apart across stages representing multiple time periods. In these diagrams, the width of the streams shows size within each category.

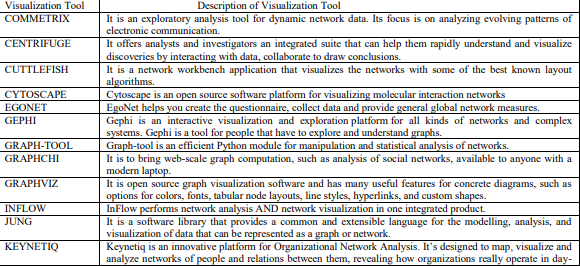
4. Sunbursts  —  Sunbursts show a hierarchical structure in a circular layout, with each ring outward representing a deeper level of the hierarchy. Ring segments are usually sized by the number of members within that segment.

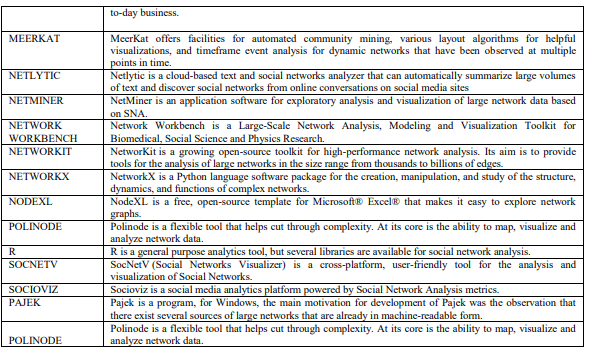
5. Circle Packing  —  Circle packing diagrams show groups as tightly-organized circles, and are often used to show hierarchies where smaller groups are either colored similarly to others in the same category, or nested within larger groups.

6. Horizon Charts  —  Horizon charts show time-series data with both negative and positive values on the vertical scale, using coloring or shading to show negative values.

7. Streamgraph  —  Streamgraphs show how the size or proportions of groups vary over time, with vertical width of the “stream” representing the size of that entity. Streamgraphs can use either a fixed scale, where change in the overall size of all groups can be seen, or a relative scale, where all groups consistently add to 100% (similar to an area chart).

**TOOLS OF VISUALIZATION**





*Let’s check the take away from this lecture*

Exercise:

1. What is the primary reason behind the explosive growth of data in the last few years, and how is social media contributing to this phenomenon?
2. Compare the characteristics of slopegraphs and parallel coordinates as visualization approaches.
3. Develop a circle packing diagram to represent a hierarchical organization within a company, with different departments and teams nested within larger groups.

**Learning from this lecture:** Learners understood visualization of social network data.

**Objective Questions:**

What are the initial steps in processing data according to the lecture?

A) Mining

B) Acquiring

C) Filtering

D) Parsing

Visualization Techniques: Which visualization method is suitable for showing trends and patterns?

A) Scatter plot

B) Pie chart

C) Line graph

D) Table

Link Prediction: What is the primary objective of link prediction?

A) Identify nodes in a network

B) Predict pairs of nodes that will form a link or not in the future

C) Count the number of nodes in a graph

D) Analyze the influence of nodes in a network

Collective Classification 4. Collective Classification Definition: What does collective classification involve?

A) Considering only individual characteristics of nodes

B) Ignoring both structure and attributes

C) Considering both connections between nodes and their individual characteristics

D) Focusing only on the most influential nodes in a network

Application of Collective Classification: In what context can collective classification be used?

A) Analyzing chemical reactions

B) Making predictions about items in a network considering connections and individual characteristics

C) Calculating prime numbers

D) Sorting elements in a database

Applications in Advertising and Game Analytics 6. Unity in Advertising: What can Unity be used for in advertising?

A) Only creating static ads

B) Creating interactive ads, virtual product demonstrations, and advergames

C) Analyzing text-based ads

D) Running email marketing campaigns

PyCharm in Game Analytics: How can PyCharm be used in game analytics?

A) Game content management and player behavior analysis

B) Creating 3D game models

C) Testing network connections in games

D) Monitoring hardware performance during gameplay

**Subjective Questions:**

1. Explain the steps involved in processing data, starting from acquisition to visualization.
2. Discuss scenarios where a bar graph would be a better choice than a pie chart for visualizing data.
3. Provide an example of a real-world scenario where collective classification could be applied effectively. Explain the nodes, connections, and individual characteristics involved.
4. Elaborate on the advantages of using collective classification over traditional classification methods based solely on individual attributes.
5. Interactive Ads: Describe an interactive ad concept that utilizes Unity. How does interactivity enhance user engagement in advertising?
6. Explain the steps involved in using PyCharm to analyze player behavior in a gaming environment. How can this analysis be utilized for game improvement?