**1. News Article:**

a. **Data to Extract:**

* Headlines: Extract the main headlines of news articles.
* Article Content: Retrieve the content or summary of each news article.
* Publication Date: Collect the date when each article was published.
* Source: Capture the name or URL of the news source.

b. **Websites to Crawl:**

* CNN (<https://www.cnn.com/>)
* BBC News (https://www.bbc.com/news)
* The New York Times (<https://www.nytimes.com/>)

c. **Challenges:**

* Different HTML structures: Each website may have a different structure for presenting news articles, requiring tailored parsing strategies.
* Pagination: Handle multiple pages of articles if the chosen websites paginate their news content.

d. **Media Types:**

* Include images associated with news articles.
* Optionally, consider extracting embedded videos or multimedia elements.

**Queries:**

1. **Trending Topics:**
   * How many articles are currently trending based on the number of views in the last 24 hours?
   * What are the top 3 trending topics with the highest engagement?
   * Which authors have contributed the most to the trending topics?
2. **Author Contributions:**
   * Who are the top 5 authors with the highest number of published articles?
   * What is the average word count for articles written by each author?
   * Which authors have the highest engagement based on reader comments?
3. **Multimedia Coverage:**
   * What percentage of articles include images or multimedia content?
   * Can we identify a correlation between multimedia content and higher user engagement?
   * How has the inclusion of multimedia evolved over the past month?
4. **Geographical Coverage:**
   * What regions or countries are most frequently covered in the news articles?
   * Are there specific topics that are more popular in certain geographic locations?
   * How has the geographical coverage changed over the past year?
5. **Temporal Analysis:**
   * What is the distribution of articles over different days of the week?
   * Are there specific months that see a surge in news coverage?
   * How has the publishing frequency changed during major events or holidays?

**2. Movie Database:**

a. **Data to Extract:**

* Movie Details: Retrieve information such as title, release date, and duration.
* Cast and Crew: Collect details about actors, directors, and other crew members.
* Genre: Extract the genre(s) of each movie.
* Ratings: Include ratings from sources like IMDb or Rotten Tomatoes.

b. **Websites to Crawl:**

* IMDb (<https://www.imdb.com/>)
* Rotten Tomatoes (<https://www.rottentomatoes.com/>)
* The Movie Database (TMDb) (<https://www.themoviedb.org/>)

c. **Challenges:**

* Dynamic Content: Some movie databases use dynamic loading of content, requiring you to handle AJAX requests or other dynamic loading mechanisms.
* User Reviews: If extracting user reviews, consider the structure of the review section on each website.

d. **Media Types:**

* Movie Posters: Download and store movie posters associated with each film.
* Trailers: Optionally, extract and store links to movie trailers.

**Queries:**

1. **Top-Rated Movies:**
   * What are the top 10 movies with the highest IMDb ratings?
   * Which genres are most commonly associated with high-rated movies?
   * Are there any notable differences in ratings across different decades?
2. **Director's Impact:**
   * How many movies has Christopher Nolan directed, and what is the average rating of his films?
   * Who are the top 3 directors with the highest average ratings for their movies?
   * Is there a correlation between a director's filmography size and average ratings?
3. **Genre Distribution:**
   * Which genres have the highest number of movies in the database?
   * What is the average rating for movies in each genre?
   * Are there trends in genre popularity over the past decade?
4. **Actor's Influence:**
   * Which actors have appeared in the most movies in the database?
   * What is the average rating of movies featuring Tom Hanks, Leonardo DiCaprio, and Scarlett Johansson?
   * Are there any patterns in the genres of movies a specific actor tends to star in?
5. **Release Analysis:**
   * How many movies were released each year over the past decade?
   * Can we identify any seasonal trends in movie releases?
   * What are the most recent movies released in the last month?

**Topic: Cricket News and Statistics**

Data to Extract:

1. **Headlines:** Extract the main headlines of cricket-related news articles.
2. **Article Content:** Retrieve detailed content or summaries of cricket news.
3. **Publication Date:** Collect the date when each article was published.
4. **Author Information:** Extract details about the authors of the articles.
5. **Team and Player Stats:** Gather statistics related to cricket teams and players.
6. **Match Results:** Capture information about match results and scores.
7. **Upcoming Matches:** Extract details about upcoming cricket matches.
8. **Media Content:** Include images and multimedia content related to cricket events.
9. **Social Media Mentions:** If available, gather data on social media mentions and reactions to cricket news.
10. **Player Transfers and Trades:** Collect information on player transfers and trades between teams.

Websites to Crawl:

1. **ESPN Cricinfo (**<https://www.espncricinfo.com/>**):**
   * A comprehensive source for cricket news, scores, and statistics.
2. **Cricbuzz (**<https://www.cricbuzz.com/>**):**
   * Provides live scores, news, and commentary on cricket matches.
3. **BBC Sport - Cricket (https://www.icc-cricket.com/):**
   * Offers cricket news, analysis, and coverage from BBC Sport.

Media Types:

* **Images:** Extract images related to cricket events, players, and match highlights.
* **Videos:** If available, include videos such as match highlights, player interviews, and analysis.
* **Infographics:** Capture infographics that depict statistics and trends in cricket.

Challenges:

1. **Dynamic Content:** Cricket websites often use dynamic content loading. Ensure your crawler handles dynamic content using tools like Selenium.
2. **Variability in Data Structure:** Different websites may have distinct HTML structures. Adjust your scraping logic accordingly.
3. **Real-time Updates:** Cricket news and statistics are constantly updated. Plan for regular updates in your crawling and scraping processes.

Queries:

1. **Top Run Scorers:**
   * Retrieve the top 5 run-scorers in international cricket for the current year.
2. **Best Bowling Performances:**
   * Find the top 3 bowling performances in T20 matches based on wickets taken.
3. **Team Rankings:**
   * Obtain the current ICC rankings for One Day International (ODI) cricket teams.
4. **Player of the Match History:**
   * Identify the players with the most "Player of the Match" awards in Test cricket.
5. **Upcoming Matches Schedule:**
   * Provide a list of upcoming international cricket matches, including teams and venues.
6. **Recent Match Results:**
   * Retrieve the results of the most recent five international cricket matches.
7. **Player Transfers Analysis:**
   * Analyze the trends in player transfers between cricket teams over the past year.
8. **Social Media Buzz:**
   * Explore the sentiment analysis of social media mentions for a specific cricket event.
9. **Player Stats Comparison:**
   * Compare the batting and bowling statistics of two specific cricket players.
10. **Most Popular Cricket News:**
    * Identify the most popular or shared cricket news articles based on social media interactions.

These queries cover a range of aspects related to cricket, including player performances, team rankings, upcoming matches, and social media engagement. Adjust these queries based on the specific needs and structure of your cricket database.