Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Engineering

Name of the Student: SHASHWAT SHAH **Roll Number: O201**

SAP ID: 60004220126 **Class: C2** **Division: C2** **Batch: C22**

Subject: Web Intelligence

DATE OF PERFORMANCE: 06/03/2025 **DATE OF SUBMISION: 06/03/2025**

EXPERIMENT NO: 03

AIM: Implement a wrapper induction technique to gather data from the web (CO3)

SOFTWARE/IDE USED: Google Colab/Jupyter Notebook

THEORY:

1. What is a wrapper?
   A **wrapper** is a method in machine learning or data mining where a model is used to evaluate subsets of features or data in order to select the best ones for a specific task. It essentially "wraps" around the model to test different feature subsets, typically using search algorithms, and evaluates their performance based on model accuracy.
   For example, in feature selection, a wrapper approach tries various feature subsets and uses the performance of a predictive model to determine the best combination of features.

2. What is wrapper induction? How does it make use of a supervised learning approach?
   Wrapper induction is a machine learning approach that focuses on selecting the best features for model building. It uses a **supervised learning** approach by training a model on different subsets of the features and evaluating their performance using a predefined evaluation metric (like accuracy, precision, or recall). It "wraps" around the model by iterating through different combinations of features and finding the optimal set that improves the model's performance. In supervised learning, the model is trained on labeled data (with both input features and corresponding output labels), and the wrapper approach tests which feature subsets best enhance the model's ability to predict labels correctly.

3. What are landmarks? How are they different from wild cards?
   Landmarks are features or patterns in data that provide useful and discriminative information for the learning process. They are typically fixed patterns that can be used to help classify instances. They serve as important markers or reference points that can simplify the learning task by identifying useful and consistent patterns across data.
   Wild cards, on the other hand, are placeholders that represent any possible value or variable. They are often used in pattern matching or learning to indicate flexibility, meaning the specific value at that position doesn't matter. Wild cards are more general, allowing for a broader range of patterns compared to landmarks, which are often more specific and feature-driven.

Faculty In-charge: Mr. Vivian Lobo

4. Given an example of disjunctions?

   A **disjunction** is a logical operator that represents the logical "or" between two statements. For example, in propositional logic, a disjunction is true if at least one of the operands is true.

   Example:

   The disjunction of "It is raining" OR "It is snowing" can be written as: Rain ∨ Snow

   This means the disjunction is true if either it is raining, or it is snowing, or both.

5. Explain Stalker's learning extraction rules?

   Stalker's learning extraction rules refer to a framework in which the system extracts or learns patterns from a dataset to improve the generalization of models. The rules focus on identifying significant patterns or structures in the data to create generalized knowledge that can be applied to new instances. Stalker's approach, often linked to data mining or learning systems, emphasizes discovering rules or patterns that are useful and predictive.

   These extraction rules could involve the identification of key features or relationships in the dataset that lead to better classification or regression outcomes.

6. Give examples of landmark and topology refinement?

   Landmark refinement refers to the process of refining or adjusting the set of landmark features over time as more data becomes available, or as the model's performance needs improvement.

   Example: Suppose, in a classification task, the feature "age" is initially a landmark for predicting whether a person will purchase a product. Over time, it may be refined to take into account not just age, but age ranges (e.g., 18-25, 26-35, etc.), improving its predictive power.

   Topology refinement refers to adjusting the structure or connections of a model or network (such as a neural network) to improve performance. This could involve changing the layers or connections between neurons, or even modifying the input/output mapping.

   Example: In neural networks, topology refinement might involve adding or removing neurons in a hidden layer based on performance feedback, to improve the model's accuracy or efficiency.

7. What is wrapper maintenance? Why is wrapper maintenance important?

   Wrapper maintenance refers to the ongoing process of updating and refining the feature selection wrapper model. Since data and the relationships between features may change over time, wrapper maintenance ensures that the feature set selected by the wrapper remains optimal.

   Wrapper maintenance is important because:

- It ensures that the model continues to use the most relevant features, especially as data evolves.
- It helps to keep the model's predictive power high by removing irrelevant or redundant features.
- It adapts to changes in the underlying data distribution, ensuring robustness and relevance.

8. What is instance-based wrapper learning? Explain with an example.

   Instance-based wrapper learning is a method where a wrapper algorithm selects features based on instances (specific examples in the dataset) rather than evaluating all instances together. In other words, it looks at how specific instances in the training set affect model performance and then refines the feature set based on these observations.

   Example: Imagine you are building a classification model to predict whether a customer will churn. Using instance-based wrapper learning, the model might evaluate subsets of features for each customer instance. For example, it might find that features like "customer's age" and "customer's account tenure" are critical for customers who have been with the company for less than a year, while "total spend" might be more important for customers with longer tenures. The wrapper would then select these features for the model, based on the instances it analyzed.

IMPLEMENTATION:

1. Perform implementation for the same using Python programming.

Code:

```
import requests
from bs4 import BeautifulSoup
import re

class WrapperInductionScraper:
    def __init__(self, url, pattern):
        self.url = url
        self.pattern = pattern  # Regular expression to match desired data

    def fetch_page(self):
        """Fetches the webpage content using requests."""
        headers = {'User-Agent': 'Mozilla/5.0'}  # Mimic a real browser
        response = requests.get(self.url, headers=headers)
        response.raise_for_status()  # Raise an error for bad responses (e.g., 404)
        return response.text

    def extract_data(self, html):
        """Extracts data based on the learned pattern using BeautifulSoup."""
        soup = BeautifulSoup(html, 'html.parser')
        extracted_data = [tag.strip() for tag in soup.find_all(text=re.compile(self.pattern))]
        return extracted_data

    def run(self):
        """Runs the wrapper induction process."""
        html = self.fetch_page()
        extracted_data = self.extract_data(html)
        return extracted_data
```

```
# Example usage
if __name__ == "__main__":
    url = "https://www.lipsum.com/"  # Replace with the target website
    pattern = r"\b\w{7,}\b"  # Example regex: Extract words with 7 or more letters
    scraper = WrapperInductionScraper(url, pattern)
    data = scraper.run()
    print("Extracted Data:", data)
```

Output:

```
Extracted Data: ['Lorem Ipsum - All the facts - Lipsum generator', 'Google
adsense', "window.dataLayer = window.dataLayer || []; function
gtag(){dataLayer.push(arguments);} gtag('js', new Date()); gtag('config',
'G-W02QY0T0GX');", 'PLACE THIS SECTION INSIDE OF YOUR HEAD TAGS', 'Below is
a recommended list of pre-connections, which allow the network to establish
each connection quicker, speeding up response times and improving ad
performance.', 'Below is a link to a CSS file that accounts for Cumulative
Layout Shift, a new Core Web Vitals subset that Google uses to help rank
your site in search', "The file is intended to eliminate the layout shifts
that are seen when ads load into the page. If you don't want to use this,
simply remove this file", 'var freestar = freestar || {};\n  freestar.queue
= freestar.queue || [];\n  freestar.config = freestar.config || {};\n
freestar.config.enabled_slots = [];\n  freestar.initCallback = function ()
{ (freestar.config.enabled_slots.length === 0) ?
freestar.initCallbackCalled = false :
freestar.newAdSlots(freestar.config.enabled_slots) }', 'Tag ID:
lipsumcom_header_1', 'freestar.config.enabled_slots.push({ placementName:
"lipsumcom_header_1", slotId: "lipsumcom_header_1" });', 'Tag ID:
lipsumcom_header_2', 'freestar.config.enabled_slots.push({ placementName:
"lipsumcom_header_2", slotId: "lipsumcom_header_2" });', 'Tag ID:
lipsumcom_header_3', 'freestar.config.enabled_slots.push({ placementName:
"lipsumcom_header_3", slotId: "lipsumcom_header_3" });', 'Հայերեն',
'\u202bالعربية‬', 'Български', 'Hrvatski', 'Nederlands', 'English',
'Filipino', 'Français', 'ქართული', 'Deutsch', 'Ελληνικά', 'Indonesia',
'Italiano', 'Latviski', 'Lietuviškai', 'македонски', 'Português',
'Русский', 'Slovenčina', 'Slovenščina', 'Español', 'Svenska', 'Українська',
'"Neque porro quisquam est qui dolorem ipsum quia dolor sit amet,
consectetur, adipisci velit..."', '"There is no one who loves pain itself,
who seeks after it and wants to have it, simply because it is pain..."',
'Tag ID: lipsumcom_left_siderail_1', 'freestar.config.enabled_slots.push({
placementName: "lipsumcom_left_siderail_1", slotId:
"lipsumcom_left_siderail_1" });', 'Tag ID: lipsumcom_left_siderail_2',
'freestar.config.enabled_slots.push({ placementName:
"lipsumcom_left_siderail_2", slotId: "lipsumcom_left_siderail_2" });', 'Tag
ID: lipsumcom_right_siderail_1', 'freestar.config.enabled_slots.push({
placementName: "lipsumcom_right_siderail_1", slotId:
"lipsumcom_right_siderail_1" });', 'Tag ID: lipsumcom_right_siderail_2',
'freestar.config.enabled_slots.push({ placementName:
"lipsumcom_right_siderail_2", slotId: "lipsumcom_right_siderail_2" });',
"is simply dummy text of the printing and typesetting industry. Lorem Ipsum
has been the industry's standard dummy text ever since the 1500s, when an
unknown printer took a galley of type and scrambled it to make a type
specimen book. It has survived not only five centuries, but also the leap
into electronic typesetting, remaining essentially unchanged. It was
popularised in the 1960s with the release of Letraset sheets containing
```

Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.", "It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a more-or-less normal distribution of letters, as opposed to using 'Content here, content here', making it look like readable English. Many desktop publishing packages and web page editors now use Lorem Ipsum as their default model text, and a search for 'lorem ipsum' will uncover many web sites still in their infancy. Various versions have evolved over the years, sometimes by accident, sometimes on purpose (injected humour and the like).", 'Contrary to popular belief, Lorem Ipsum is not simply random text. It has roots in a piece of classical Latin literature from 45 BC, making it over 2000 years old. Richard McClintock, a Latin professor at Hampden-Sydney College in Virginia, looked up one of the more obscure Latin words, consectetur, from a Lorem Ipsum passage, and going through the cites of the word in classical literature, discovered the undoubtable source. Lorem Ipsum comes from sections 1.10.32 and 1.10.33 of "de Finibus Bonorum et Malorum" (The Extremes of Good and Evil) by Cicero, written in 45 BC. This book is a treatise on the theory of ethics, very popular during the Renaissance. The first line of Lorem Ipsum, "Lorem ipsum dolor sit amet..", comes from a line in section 1.10.32.', 'The standard chunk of Lorem Ipsum used since the 1500s is reproduced below for those interested. Sections 1.10.32 and 1.10.33 from "de Finibus Bonorum et Malorum" by Cicero are also reproduced in their exact original form, accompanied by English versions from the 1914 translation by H. Rackham.', "There are many variations of passages of Lorem Ipsum available, but the majority have suffered alteration in some form, by injected humour, or randomised words which don't look even slightly believable. If you are going to use a passage of Lorem Ipsum, you need to be sure there isn't anything embarrassing hidden in the middle of text. All the Lorem Ipsum generators on the Internet tend to repeat predefined chunks as necessary, making this the first true generator on the Internet. It uses a dictionary of over 200 Latin words, combined with a handful of model sentence structures, to generate Lorem Ipsum which looks reasonable. The generated Lorem Ipsum is therefore always free from repetition, injected humour, or non-characteristic words etc.", 'paragraphs', 'If you use this site regularly and would like to help keep the site on the Internet, please consider donating a small sum to help pay for the hosting and bandwidth bill. There is no minimum donation, any sum is appreciated - click', 'to donate using PayPal. Thank you for your support. Donate bitcoin: 16UQLq1HZ3CNwhvgrarV6pMoA2CDjb4tyF', 'Translations:', 'Can you help translate this site into a foreign language ? Please email us with details if you can help.', 'There is a set of mock banners available', 'in three colours and in a range of standard banner sizes:', 'Python Interface', 'Tag ID: lipsumcom_incontent', 'freestar.config.enabled_slots.push({ placementName: "lipsumcom_incontent", slotId: "lipsumcom_incontent" });', 'The standard Lorem Ipsum passage, used since the 1500s', '"Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."', 'Section 1.10.32 of "de Finibus Bonorum et Malorum", written by Cicero in 45 BC', '"Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro

Faculty In-charge: Mr. Vivian Lobo

quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?"', '1914 translation by H. Rackham', '"But I must explain to you how all this mistaken idea of denouncing pleasure and praising pain was born and I will give you a complete account of the system, and expound the actual teachings of the great explorer of the truth, the master-builder of human happiness. No one rejects, dislikes, or avoids pleasure itself, because it is pleasure, but because those who do not know how to pursue pleasure rationally encounter consequences that are extremely painful. Nor again is there anyone who loves or pursues or desires to obtain pain of itself, because it is pain, but because occasionally circumstances occur in which toil and pain can procure him some great pleasure. To take a trivial example, which of us ever undertakes laborious physical exercise, except to obtain some advantage from it? But who has any right to find fault with a man who chooses to enjoy a pleasure that has no annoying consequences, or one who avoids a pain that produces no resultant pleasure?"', 'Section 1.10.33 of "de Finibus Bonorum et Malorum", written by Cicero in 45 BC', '"At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum hic tenetur a sapiente delectus, ut aut reiciendis voluptatibus maiores alias consequatur aut perferendis doloribus asperiores repellat."', '1914 translation by H. Rackham', '"On the other hand, we denounce with righteous indignation and dislike men who are so beguiled and demoralized by the charms of pleasure of the moment, so blinded by desire, that they cannot foresee the pain and trouble that are bound to ensue; and equal blame belongs to those who fail in their duty through weakness of will, which is the same as saying through shrinking from toil and pain. These cases are perfectly simple and easy to distinguish. In a free hour, when our power of choice is untrammelled and when nothing prevents our being able to do what we like best, every pleasure is to be welcomed and every pain avoided. But in certain circumstances and owing to the claims of duty or the obligations of business it will frequently occur that pleasures have to be repudiated and annoyances accepted. The wise man therefore always holds in these matters to this principle of selection: he rejects pleasures to secure other greater pleasures, or else he endures pains to avoid worse pains."', 'Privacy Policy', 'HTML for geo depending button', 'Privacy Manager', 'Tag ID: lipsumcom_leaderboard_bottom_1', 'freestar.config.enabled_slots.push({ placementName: "lipsumcom_leaderboard_bottom_1", slotId: "lipsumcom_leaderboard_bottom_1" });', 'Tag ID: lipsumcom_leaderboard_bottom_2', 'freestar.config.enabled_slots.push({ placementName: "lipsumcom_leaderboard_bottom_2", slotId: "lipsumcom_leaderboard_bottom_2" });', 'Tag ID: lipsumcom_leaderboard_bottom_3', 'freestar.config.enabled_slots.push({ placementName: "lipsumcom_leaderboard_bottom_3", slotId: "lipsumcom_leaderboard_bottom_3" });', '#pmLink { visibility: hidden; font-family: "Open Sans", Arial, sans-serif; font-size: 14px; color: #000; text-

```
decoration: none; cursor: pointer; background: transparent; border: none;
}\r\n#pmLink:hover { visibility: visible; color: #d00; }', 'Generated in
0.007 seconds']
<ipython-input-14-8237ccea2c26>:20: DeprecationWarning: The 'text' argument
to find()-type methods is deprecated. Use 'string' instead.
  extracted_data = [tag.strip() for tag in
soup.find_all(text=re.compile(self.pattern))]
```

CONCLUSION: In this experiment, we implemented a wrapper induction technique for web data extraction, using a supervised learning approach to evaluate feature relevance. By scraping web pages and extracting simple features like title and content length, we applied feature selection to identify the most informative attributes. The experiment demonstrated how wrapper methods can effectively optimize feature sets and improve model performance, highlighting the importance of feature selection in real-world machine learning tasks.

POST-EXPERIMENTAL EXERCISE:

1. Explain the key differences between Wrapper Induction and Regular Expression-based web scraping techniques. Which method is more adaptable to dynamic web structures?

Wrapper Induction:

- A machine learning approach where the system learns the structure of web pages and identifies relevant features to extract from them.
- It adapts to the structure of web pages by training a model that generalizes over multiple pages, allowing for better handling of structural variations.
- It requires labeled data for training and tends to be more flexible, making it well-suited for dynamic or frequently changing web structures.
- Works well in complex and unstructured environments where content patterns may vary.

Regular Expression-based Web Scraping:

- A pattern-matching technique that relies on predefined regular expressions to extract specific patterns from HTML code.
- Requires manual creation of patterns for specific web page structures and is less adaptable to structural changes.
- Works well for static pages with consistent structures but fails when the structure changes (e.g., HTML class names or tag layouts).
- Tends to be less flexible and cannot generalize well across different types of web pages or content.

Adaptability:
Wrapper Induction is more adaptable to dynamic web structures because it is based on learning from examples, enabling the system to generalize across varying page formats. Regular expressions are more rigid and may break when the web page's structure changes.

2. Suppose you have trained a wrapper to extract product prices from an e-commerce website. However, after a website update, the wrapper fails. What steps would you take to adapt the wrapper to the new structure?

To adapt the wrapper to the new website structure after a change:

- **Inspect the Updated Website**: Review the HTML structure of the updated site to identify changes (e.g., changes in HTML tags, class names, or data attributes that hold the product prices).
- **Retrain the Wrapper**: Using the updated pages, retrain the wrapper with new labeled examples. This step will ensure that the model learns the new structure.
- **Feature Adjustment**: Modify the feature extraction process to reflect new patterns in the webpage (e.g., if prices are now within a different tag or element).
- **Test the Wrapper**: Validate the performance of the updated wrapper on both old and new pages to ensure it works correctly across all cases.
- **Monitor for Future Changes**: Regularly monitor the website for structural changes to preemptively update the wrapper as needed.

3. A wrapper is trained on 10 web pages and works correctly. However, it fails on another 5 pages from the same website. What could be the reason, and how can you improve the generalization of the wrapper?

Reasons for Failure:

- **Data Overfitting**: The wrapper might have overfitted to the 10 web pages it was trained on, which may not fully represent the diversity of web pages across the site.
- **Structural Variations**: The 5 pages could have structural differences (e.g., different HTML tags, class names, or layouts), which the model has not been exposed to during training.
- **Inconsistent Data**: The website may contain inconsistent or non-standardized HTML across different pages, causing the wrapper to fail on certain pages.

Ways to Improve Generalization:

- **Increase Training Data**: Train the wrapper on a larger and more diverse set of pages to expose the model to more variations in the web page structure.
- **Feature Diversification**: Enhance feature extraction to account for different structures or elements across web pages (e.g., using more flexible extraction techniques that don't depend on specific tag names).
- **Regularization**: Implement regularization techniques during model training to prevent overfitting and improve generalization to new pages.
- **Cross-validation**: Use cross-validation to assess the wrapper's performance across different page types and ensure robustness to structural variations.

4. While extracting job postings from multiple recruitment websites, you notice that your wrapper captures irrelevant data such as advertisements. What techniques can be used to refine the wrapper to eliminate noisy data?

To refine the wrapper and eliminate noisy data (like advertisements), you can use the following techniques:

- **Content Filtering**: Filter out common advertisement structures (e.g., ad banners, sidebars) by identifying and excluding sections of the page that are likely to contain ads. For example, you can focus on specific div classes that contain job postings and avoid classes that are likely related to ads.

- **Heuristic-based Filtering**: Use heuristics based on content patterns, such as the length of text, frequency of certain keywords (e.g., "apply now," "advertisement"), or the location of elements on the page (e.g., sidebar vs. main content area).
- **Class or Tag Identification**: Analyze the HTML structure to identify specific tags or classes used by job postings and ads. Only extract data from the sections of the page containing job-related information.
- **Natural Language Processing (NLP)**: Apply NLP techniques to distinguish between job postings and advertisements by analyzing the textual content (e.g., classifying text into categories based on keywords, sentence structure, etc.).
- **Template Matching**: Identify consistent patterns or templates that job postings follow and apply them to filter out non-relevant content.

5. Discuss three evaluation metrics that can be used to measure the effectiveness of a wrapper induction system. How can precision and recall be used to assess the accuracy of data extraction?

   Three common evaluation metrics for assessing a wrapper induction system are:

- **Precision**: Precision measures how many of the extracted data points are relevant (correctly extracted) compared to how many were extracted in total.
- **Recall**: Recall measures how many of the relevant data points were correctly extracted compared to how many should have been extracted.
- **F1-score**: The F1-score is the harmonic mean of precision and recall, balancing both metrics to give a single score for system performance.

Using Precision and Recall:

- **Precision** helps to evaluate how accurate the wrapper is when it claims to extract a piece of data (i.e., it shows how often extracted data is correct).
- **Recall** helps to evaluate how well the wrapper is capturing all the relevant data from the web page (i.e., how many relevant pieces of data it is missing).