Part I - Websites for training and testing

- We collected 20 publicly accessible websites for crawling and content extraction.
- Five websites will be reserved as our testing set, while the remaining fifteen will be used for training purposes.
- Website links are shown as below:
    - 1.https://www.cnn.com/travel/hagia-sophia-istanbul-hidden-history/index.html
    - 2.https://www.cnn.com/2025/04/25/style/labubu-plush-toy-buying-craze-hnk-intl/index.html
    - 3.https://www.cnn.com/2025/04/28/style/chrysler-building-art-deco-centennial/index.html
    - 4.https://www.nad.org/resources/how-to-file-a-complaint/
    - 5.https://www.nad.org/resources/international-advocacy/
    - 6. https://www.nad.org/about-us/priorities/2022-2024/
    - 7.https://towardsdatascience.com/when-openai-isnt-always-the-answer-enterprise-risks-behind-wrapper-based-ai-agents/
    - 8.https://towardsdatascience.com/retrieval-augmented-generation-rag-an-introduction/
    - 9.https://towardsdatascience.com/googles-new-ai-system-outperforms-physicians-in-complex-diagnoses/
    - 10.https://www.nasa.gov/news-release/nasa-astronaut-don-pettit-to-discuss-seven-month-space-mission/
    - 11. https://www.nasa.gov/newsletters/
    - 12. https://www.tripsavvy.com/the-worlds-coldest-cities-4067210
    - 13. https://www.tripsavvy.com/the-best-ski-towns-in-the-us-6834692
    - 14.https://www.tripsavvy.com/inside-annual-competition-that-determines-best-baguette-in-paris-7569537
    - 15. https://www.gutenberg.org/cache/epub/75978/pg75978-images.html
    - 16. https://www.allrecipes.com/recipe/268091/easy-korean-ground-beef-bowl/
    - 17. https://www.allrecipes.com/recipe/277945/spicy-baked-shrimp/
    - 18. https://www.allrecipes.com/gallery/easy-chinese-recipes/
    - 19.https://computer.howstuffworks.com/internet/tips/how-to-see-and-delete-google-history.htm
    - 20.https://history.howstuffworks.com/historical-events/10-things-missing-without-trace.htm

Part II - Grading rubrics

-

Priority 0-0	Main Title	The title of the main content, the core theme of the page
Priority 0-1	Main Content	The main information block of the page
Priority 0-2	Primary Action Button	Directly related action buttons
Priority 1-0	Subtitles	Section headings in the body of the text

Priority 1-1	Main Navigation	Top navigation of the web page
Priority 2-0	Search Box	Site search input box
Priority 2-1	Secondary Buttons	Buttons not directly related to the task
Priority 3-0	Related Content	Links to pages with the same or similar topic
Priority 3-1	Related Info	Publication date, author signature. etc
Priority 4	Unrelated	Content that should be ignored

- Level 0 would be marked as 5 out of 5
- Level 1 would be marked as 4 out of 5
- Level 2 would be marked as 3 out of 5
- Level 3 would be marked as 2 out of 5
- Level 4 would be marked as -5 out of 5

- Priority score formula:
    - $S = \frac{(5*C0 + 4*C1 + 3*C2 + 2*C3) - (2.5\times E0+2\times E1+1.5\times E2+1\times E3)-(5\times W4)}{5*T0 + 4*T1 + 3*T2 + 2*T3} * 100\ \%$
        - Ci -> Corrected_priority_i
        - E -> Error_priority_i (missing or mis-selected)
        - W -> Wrong_priority_i
    - S score will be considered as 70 percent of the evaluation

Part III - Human Evaluation

- Our team member will conduct a human evaluation to see if our project and grading criteria meet the real human interaction process.
    - 1. Read through the website
    - 2. Mark down the content by its priority to the audience
    - 3. Compare the results to our project outcome
    - 4. Compute the S score manually to see if there are any huge differences
- Human_S score will be considered as 30 percent of the evaluation

Part IV - Evaluation

- We will evaluate our project with the baseline, which we choose to be the traditional information extraction method.
- We will compute the S score for each method, and compare the results. The S score obtained there will be considered as 70% of the final score.
- We will also do a human evaluation, and the results from human evaluation will be used as 30% for the final score.
- We will compare each final score to see how our project is better than the traditional ways of information extraction.