# Project Proposal: HyperLink Classification

***Group Name: TTIS 5***
Captain: Zhengkai Zhang (zz68)
Yuan Chung Ho (ych11)
Wan Feng Cai (wfcai2)
Xinqian Xiang (xinqian6)
Zheng Ma (zhengma3)

## Abstract

We receive a mass amount of information everyday, for example advertisements from email and push notifications from search engines, and it comes with a link to another page. We scan through titles and quickly make decisions whether we are interested or not. It is time consuming in searching through the interesting information. Therefore, we are encouraged to seek a way to pick out useful information to make our life more efficient.

The HyperLink Classification is a project that helps users to identify whether a link they receive from email or search engine (eg: google) meets their needs. The user profile will be built based on their selected interested categories and some other helping data like relevance feedback. It will help the system identify if the link is useful or interesting to them.

## Scope of Work

The task consists of two subtasks. The first part is to build a web based/terminal based tooling that allows an user selects their interesting categories and enters a link. The second part is to judge whether the link meets user's interest and to update the system based on users feedback or click through.

We will need to build client side and server side for this tool and based on the entered hyperlink we will parsing the web page in our server side. With this parsed data from the web, we will compare and run the algorithm we learned to decide if the link fits to the current user's needs and return the result to the user.

In addition, we plan to add some user feedback systems based on the user feedback or click through if the user wants to do it. We could potentially use these data to improve our algorithm for later judgements.

# Introduction

## *Why is it important or interesting?*

There are many systems that provide recommendations or classification for different purposes, but we haven't seen any tools that help to link classification.

This project is a fundamental one with potential to extend to a start-up project. It could potentially be used for helping any pages/email to check their sublink or content link, marking the qualified link with some highlights to the user and even give tooltips. It could be a very useful plugin and generate important user data for different web-pages or companies.

## *Our planned approach*

1. Build the client(web/terminal) and server with APIs and database.
2. Obtain the initial data for each category and write the script to parsing the web page.
3. Clean the data and develop algorithms to evaluate the parsed data to compare with the user database for outputting user results.
4. Make use of user's relevance feedback to update the evaluation algorithms and user database for future judgement.

# Methods

## *Tools, systems and dataset*

We will need javascript for some front end and python for server side. We will also use scripts like beautiful soup and databases like mysql. For natural language processing, we plan to make use of some functions from Google Nature Language AI/Microsoft Cognitive Services and NLP Algorithms.

We will create a user dataset based on ourselves and our friends to examine the tools and systems created.

# Programming languages

Javascript, html, python, flask, beautiful soup, mysql/sqlite.

# Outcome and Evaluation

Outcome of the system will be a binary judgement (yes or no) for the link for the particular user and help them decide whether the link meets their interest. Then they can make a decision whether they want to click/read the content to save their time.

Users will use this tool to see if it helps to filter out the links they are interested in. It's not easy to get many users but all our team members could use the tool and give feedback. We could also make it a real web tool to get feedback from others.

# Process

In this project, we will work as a group to build the tool. It will take over 100 hours for our team with 5 members. We will follow our processes shown below and approximate our time spending for each section. Note that the list only includes the implementation time, we will need to spend time learning the language.

1. Build the front end interface (allow user to select category, input link): ~10 hours
2. Build the server side api and database (provide api for inserting data/do classification, build the database for the user): ~20 hours
3. Scripting all potential web link and be able to evaluate the content text inside and integrate with server: ~15 hours
4. Prepare each category's data with different source and integrate with backend/front end: ~15 hours
5. Write the algorithm and integrate with server side: ~ 25 hours
6. Test and evaluate the tooling: 15 hours
7. Other optional features may need: x hours

# Timeline

| Execution Schedule | Event |
|---|---|
| October 25, 2021 | Proposal submission |
| November 1, 2021 | Prepare data, scrape links, build api/interface |
| November 8, 2021 | Prepare data category and Implement algorithm |
| November 16, 2021 | Project Progress Report Submission |
| November 22, 2021 | Test and evaluate tool, add other features if needed |

| | |
|---|---|
| November 29, 2021 | Finish up project code and prepare the presentation material |
| December 10, 2021 | Project Code, Documentation and Presentation submission |