FOCUS TRACKING

PROBLEM DESCRIPTION:

In these modern times, working remotely, we all have blurred our professional and personal lives. Most of us do not have an idea/track of how much time is being allotted to work, making time management even more challenging. This project utilizes modern natural language processing strategies to track down a user's web usage and help the user manage time better.

Computational Perspective:

One of the ways to approach such a problem is by understanding the user's web usage. Using Natural Language Processing, we can classify and cluster all the web pages the user has visited. By scraping all the web pages, we will have text data, which can be fed to models that classify or cluster sequential text inputs.

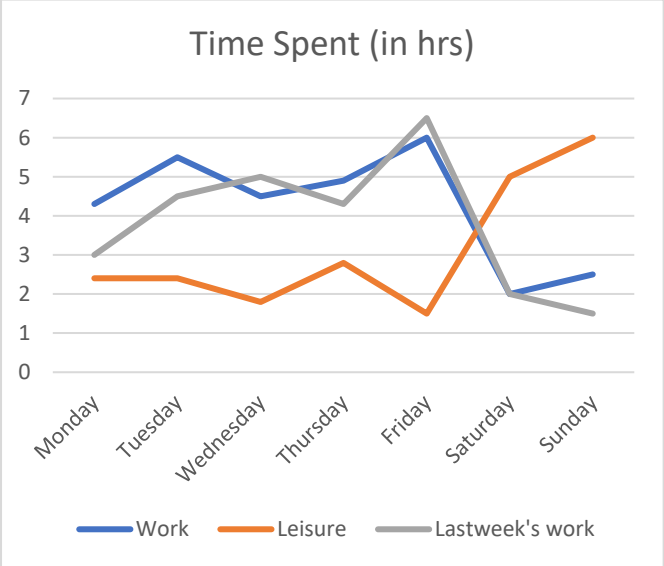
Inputs:

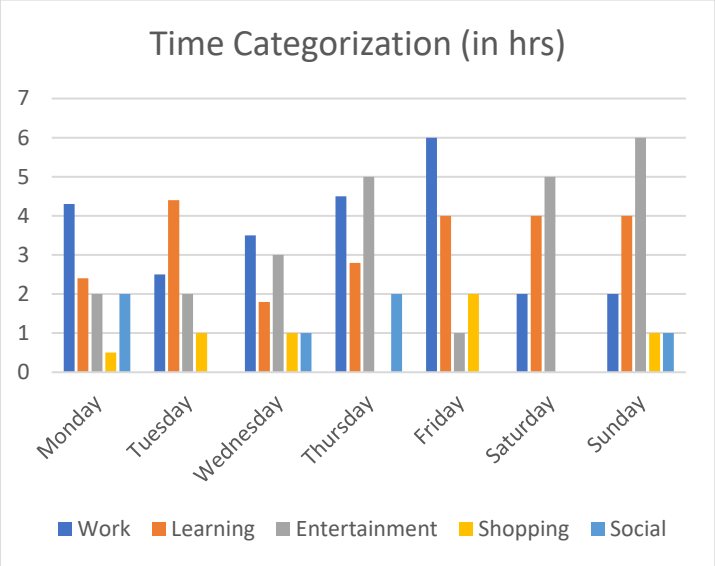
To analyze and track down a user's web usage, we need the following inputs:

- User's web history – Web pages that the user has visited.
- User's active time – Time spent on each web page.
- Information about the user's professional life in natural language.

Outputs:

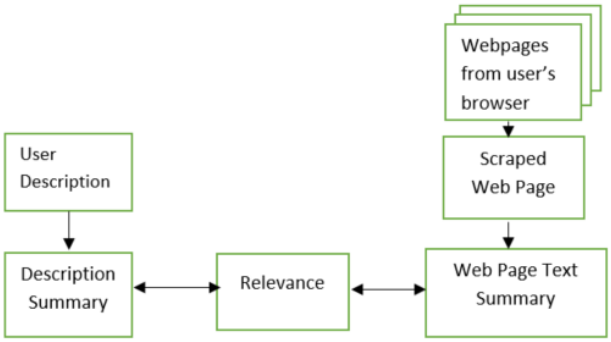
Graphical representation of user's time usage. The graphs contain information on how much time is spent on professional work, time taken by each topic of work and the amount of leisure time. The following are example outputs we expect as results:





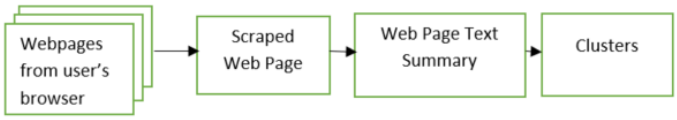
Classification Model:

The classification model will be designed to take the information of the user's professional life (such as a LinkedIn profile page) and all the web pages the user has visited through the browser. It performs text summarization of each web page and user description and calculates relevance (similarity) between the two summaries. Based on the relevance score, the web page is classified into work or non-work. The total time spent on work and non-work can then be represented on a graph. The following diagram shows the workflow:



Clustering Model:

In the above step, to calculate relevance, text summaries are extracted from each web page. These text summaries are then encoded by an encoder network and then clustered into categories. Categories such as Work, Entertainment, Social, Shopping and more. The time spent on each cluster can then be represented on a graph.



Algorithms:

We will search for different algorithms in the following areas. We also mentioned a few potential algorithms we might end up using. Based on the results, we will increase or decrease the complexity of these models.

1. Name Entity Recognition: Bert, Hidden Markov Models, Support Vector Machines.
2. Text Summarization: Bert, Seq2Seq(encoder-decoder).
3. Unsupervised Clustering: DB-SCAN, Hierarchical Clustering.
4. Similarity Metrics: Cosine Similarity, Euclidean Distance, Manhattan Distance.
5. Embedding Models: Word2Vec, GloVe.

Why it is interesting:

We all keep complaining about how we couldn't achieve our goals due to lack of time.

- An Ideal focus tracking program can help a user balance his work life and personal life, prioritize time and achieve more.
- Impact: an Ideal program can improve the digital life of every computer user.

As students working on this project:

- Helps us explore modern Strategies used in NLP, Such as Text Summarization, Named Entity Recognition and more.
- Exposes us to Deep-Learning-based NLP such as Transformer Networks and utilize them for different purposes.
- Helps us understand the cleaning and preparation process of text data.
- Makes us work with large corpora of text data and retrain the pre-trained networks (using Transfer Learning).