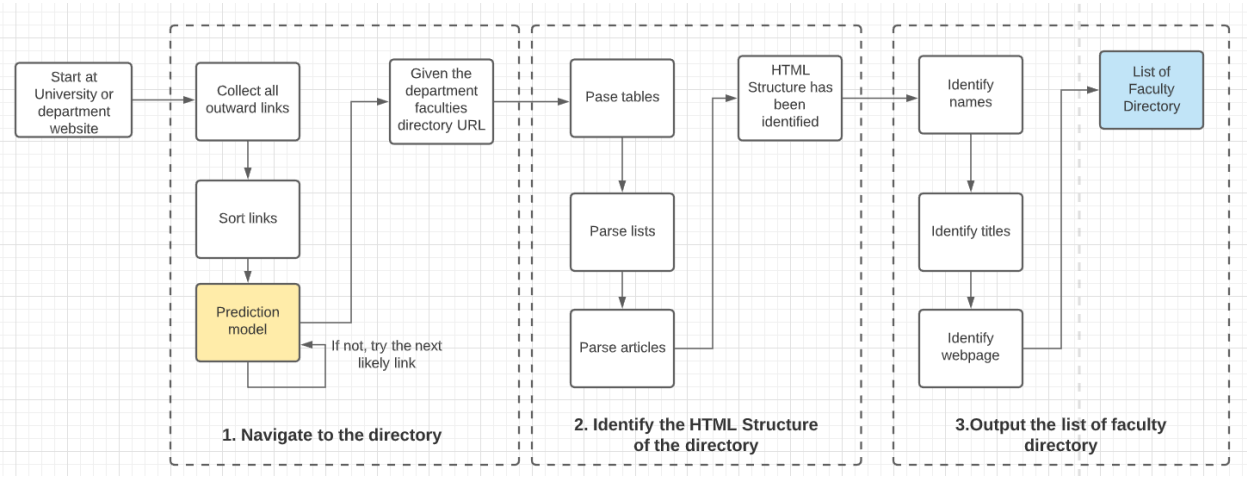
CS410 Project Proposal - Automatic Crawler of Faculty Pages

Net IDs: ap41, gangyao2, sg54
Group Name: Team Texas

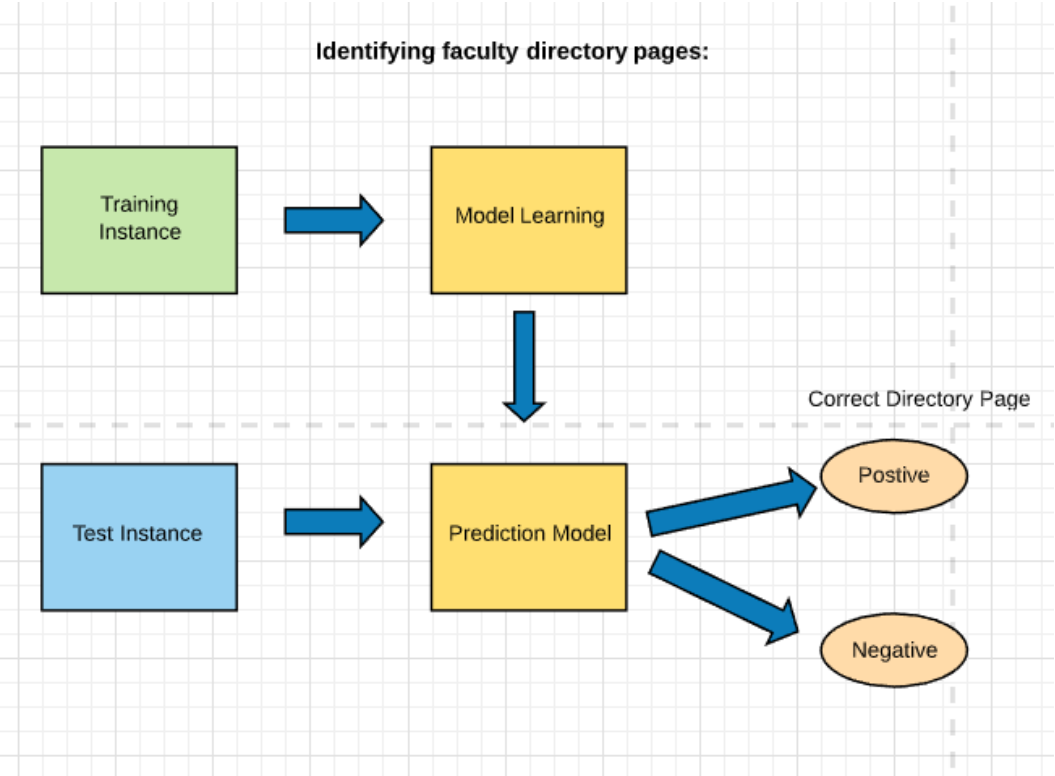
Project Scope:

1. Develop an automatic crawling tool that grabs URLs from a given university web page.
2. Train a classification model with a training instance for determining if a URL corresponds to a faculty directory page.
3. Identify the URLs corresponding to faculty member biographies.
4. Validate the classification models.

One possible solution is depicted in the following figure, showing how we can combine a Classification Model with page structure analysis.



The following figure describes how the Classification Model will use a Training Set to build the model and a Test Set to measure the accuracy.

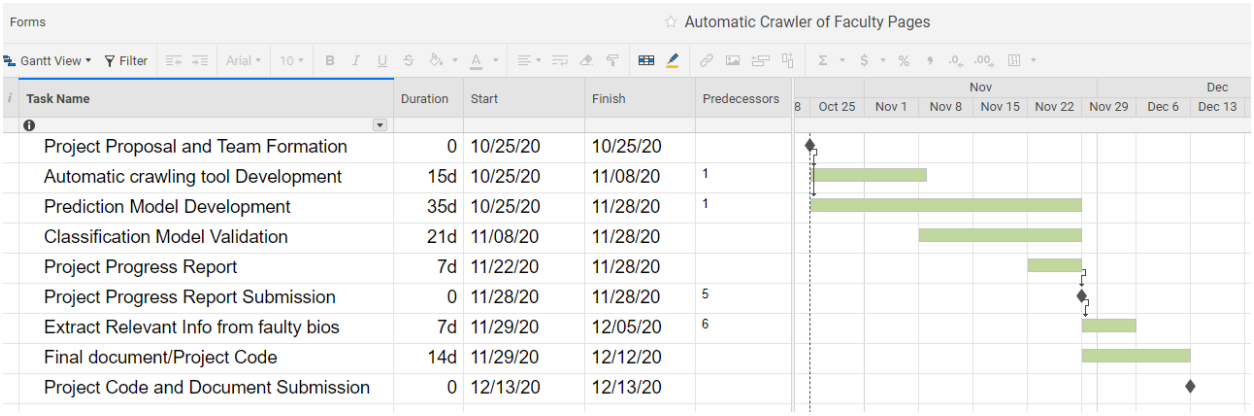


Project Overview:

- What is the function of the tool?
    - Crawling a university home page to find a URLs of faculty directory pages. An example of this would be: https://anthropology.stanford.edu/people/faculty
    - Given the faculty directory page, the automatic crawler should again identify each faculty URL. Example: https://aa.stanford.edu/people/faculty/grid
- Who will benefit from such a tool?
    - We believe our tool can be expanded upon for a variety of use cases involving a full repository of faculty across universities.
    - Some use case examples include:
        - Users trying to establish a faculty census at a university.

        - Students looking to contact people working on common areas of research. Once the automatic crawler is integrated with a system like the *ExpertSearch*, such use case will be possible.
        - PhD applicants looking for specific instructors/PIs to work with based on the field they are focused on.
        - Organizations looking for people with the expertise to solve problems faced in the industry.
- Does this kind of tool already exist? If similar tools exist, how is your tool different from them? Would people care about the difference?
    - In terms of finding research papers and experts at universities around the world, it seems like *GoogleScholar* has been of great use in recent years.
    - There are instances of very robust commercial *topic crawlers* assisted by AI: https://www.scrapestorm.com
    - Even in the specific topic of interest (faculty pages), there have been efforts: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6114776/
    - Our tool would be different from these in that it would serve to build a framework with which to access the information for any faculty member belonging to any university. It will be general and easily extended to fit any use case requiring this information.
- What existing resources can you use?
    - A tool like scrapy (https://scapy.net/) would allow us to make the process of finding URLs and anchor texts easier, which will be the main input for our *Prediction Models (Classifiers)*.
    - The existing faculty pages in the *ExpertSearch* system can be used as positive examples in training our Prediction Model. We still need to acquire negative example by crawling/scraping different kinds of websites.
    - A library such as scikit-learn (https://scikit-learn.org/stable/) is a good option to train the Classification Model.
        - Scikit-learn SVM library to train the SVM classification model or
        - Scikit-learn Naïve Bayes library to train the Naïve Bayes classification model.
- What techniques/algorithms will you use to develop the tool? (It's fine if you just mention some vague idea.)
    - We will use a Machine Learning Model such as Logistic Regression or SVM, Naïve Bayes or Pattern-Based Classification like DPClass to develop the tool.
    - To evaluate the classifier, we will use F-measure and a test set which will be independent from the training set.
    - The URLs should be tokenized and passed to our Classifier. We could also use the anchor text as an input to avoid problems such as an *"empty URL"*.
    - The crawler would have courtesy considerations to avoid robot exclusion.

    - In the implementation we might need a mechanism where the crawler should abandon fruitless explorations.
    - We are considering that the crawler will follow a Breadth-First exploration logic.
- How will you demonstrate the usefulness of your tool?
    - With a set of University URLs (preferably not in the training set) the tool will output the faculty URLs.
- A very rough timeline to show when you expect to finish what. (The timeline doesn't have to be accurate.)



References:

1. Allison C. Morgan, Samuel F. Way, Aaron Clauset, Automatically assembling a full census of an academic field. PLoS ONE, April 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6114776/

2. A Comprehensive Study of Features and Algorithms for URL-Based Topic Classification, https://ingmarweber.de/wp-content/uploads/2013/07/A-Comprehensive-Study-of-Features-and-Algorithms-for-URL-Based-Topic-Classification.pdf