1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

This is a solo project. My name is Kiranmayee Nimashakavi. My NetId is kn13@illinois.edu.

1. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

WebCrawler and classifier for MenuPages.com or some site that offers restaurant menus. The objective is to crawl the entire site and collect all pages and then classify the pages that has menu information. This information can be used to create a dish specific search on the restaurant menu.

My approach is to build a non-recursive crawler that extracts all pages from a given list of sites and use some classifier as (naïve bayes) to classify the pages. I’ve not decided on the tools to use on this project yet. The specific set of systems that are used are the list of websites that will be given as an input to the program.

Expected outcome is to store the Menu pages to a specific directory. I will manually test the output data by supplying some random collection of URLs.

If time permits, I will implement the dish search on the dataset that was classified as a menu page.

1. Which programming language do you plan to use?

Python

1. Please justify that the workload of your topic is at least 20\*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

The following activities has to be performed to accomplish this project.

* Design and implementation of the non-recursive crawler. – 5 hours
* Design and implementation of the classifier – 3 hours
* Unit and manual testing of the crawler and classifier - 5 hours
* Performance tuning of crawler and classifier – 3 hours
* Documentation of the project – 4 hours
* Presentation of the project -2 hours