Project To Do List:

Focussing web crawler on recently popular topics:

* Add function to website summariser that uses keywords found from current website to determine a genre for the webpage
* This can then be stored by the crawler into the database of websites
* Then, after a webpage is viewed by the user and hence stored in their search history, the genre will aso be stored in this table
* Upon new crawl cycles, the crawler will look at all of the search history entries from within the last week, determining the most popular genres ( can also use keywords as well), and using these to focus its search to make the crawl list more accurate and relevant
* This process can be used to get individual suggestions for all users – ie by looking at the most popular genres for each person and suggesting similar things – or by identifying groups for which the suggestions can be made for
* Neural network:
* Would be run independently of the crawler and web browser
* Will use search history and personal data like age, address, gender, etc to create the recommendations
* Preprocessing – take user data and ensure all of it can be used in the model
* Design – define architecture for the NN
* Training – Use collected data to train model, define a loss function that captures the accuracy of recommendation compared to user preferences – need to define user preferences
* Test on actual system

Server Side Processing:

* Use genres/keywords/urls from websites in a user’s search history in order to generate recommendations for them – these can be shown in the main page and used to ask the user if they want to further explore a topic that they are interested in
* Can also use a similar technique to the web crawler’s group trends functionality in order to suggest stuff that other people with similarities are searching/show popular searches – maybe by area as well??
* Currently have a simple implementation that takes a url and finds similar websites to it in the database made by the crawler
* url passed to this function can be chosen out of the most popular searches from the user recently

Testing:

* Testing the web crawler:
* Test for how good diverse the crawl space is after a crawl cycle has ended – this can be fairly simple because the genre of each website is saved so I can write a program to determine the number of each genre – I would expect there to be more websites from the genre of the starting point but it is good for it to explore the web more and get some variation. Will have different starting points in each cycle so need to test at the very end to make sure lots of genres are chosen and the crawler isn’t biased
* Test how it uses gernes to focus its search – again fairly simple because we can use the function to count genres to see if the genre of the crawled websites tend towards the more popular genres in recent searches
* Speed – crawler works in the background so not number one priority to have a fast crawler but still good to test – measure number of sites crawled per second, how adjusting depth of the search effects speed
* Accuracy of summaries and keywords generated from each website. Take a sample a crawled sites and ensure that the summaries match the content of the page
* Testing search engine:
* Speed is important because it will affect the user experience
* How does the size of the crawl list impact performance
* How does the introduction of data processing affect performance – ie test the speed when no user is signed in – hence no extra data is saved – compared to when a user is signed in and different algorithms are being run by the browser
* Test actual functionality of the browser – does it return a relevant sample of webpages based on the user’s query? Is it easy to use? Are there any bugs? Does it feel like the search experience is personalised? If so, in what ways?
* For testing functionality, can get people to test the system and report back on how their experience was and if they felt that any areas could be improved

Additions to Front End:

* Change password option
* Privacy settings – enable/disable certain cookies – this is possible when the user first interacts with the page but after that there is no option to change these preferences
* Add section in the main page that gives the user different suggestions – currently the interface just ahs the search box but I can add in sections underneath that offer different suggestions based on the user’s recent searches and group trends