CS410 Project Proposal: Food Recipe Search Engine

Team name: F20_TIS_DEV_TEAM

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 member*s**

Team Members:

Jonathan LaFlamme (Captain)	jml11@illinois.edu
Pradeep Sakhamoori	ps44@illinois.edu
Rahul Sharma	rahul9@illinois.edu
Rohan Khatu	khatu2@illinois.edu

Project Overview (*What is the function of the tool?)***:**

2. ***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?***

What is your free topic? Please give a detailed description. What is the task?

Our project proposal is a vertical search engine that specializes in retrieving Indian Cuisine food recipes in a collection aggregated from multiple sources. Our goal is to support both push and pull retrieval models. We think that the specialization of this search tool is what will allow us to return more relevant documents to our users compared with existing food recipe search platforms that are broader in scope. For instance, we can build a user profile (essentially a customized background language model) with a brief initialization phase (ask the user a series of questions when they set up their profile) that should improve accuracy with each subsequent query (since we will gain a better understanding of our user's palette in the form of language modeling with user feedback/relevance judgments). We may also include sentiment analysis of recipe reviews as part of our document ranking algorithm at a later stage, but we are not yet committed to that. It will depend on available time. This project proposal is being submitted under the "Free Topics" category.

Why is it important or interesting?

This project holds tremendous academic value for our group in the context of this course as it requires both a high level understanding information retrieval as well as many opportunities to wrestle with challenging design questions specific to our context and application, such as language modeling, feedback mechanisms, ranking functions, classification, recommender systems and evaluation techniques.

In a broader sense, we think that our application should outperform more generic food recipe search engines because we will be able to tune our search results based on a narrower context. But more specifically, we think that in the context of food recipes this will be especially advantageous due to the lower amount of noise present in our user's background language model. For instance, a user who routinely searches for American, Indian and Italian cuisine would add a lot of misdirection to their background language model compared with a user who only ever searched for Indian cuisine. Our system would begin to develop a more sophisticated picture of a user's Indian cuisine preferences and consequently retrieve more relevant recipes.

Target User (*Who will benefit from such a tool?)***:**

We think this application would appeal to at-home cooks and chefs who are interested in Indian Cuisine.

Originality of Tool (*Does this kind of tool already exist? If similar tools exist, how is your tool different from them?* Would people care about the difference?*)***:**

We could not find a major food recipe search engine that specifically focuses only on Indian Cuisine. However, many of the most popular food recipe websites do provide search filters for querying or browsing Indian Cuisine. Yet, as mentioned in an earlier section, we think that our limited scope of topic coverage will allow us to tune our retrieval models to a greater degree of accuracy than would be possible on a platform that supports all types of cuisine. Another advantage of our tool is that it aggregates recipes in this one category across multiple sources. This adds a level of convenience and efficiency for our users who can query or browse a much larger collection of recipes from a single source that would otherwise be a tedious process working across multiple websites with different types of interfaces and retrieval models.

Resource Utilization (*What existing resources can you use? What techniques/algorithms will you use to develop the tool?)***:**

We think that we can use the META toolkit for many of the preprocessing tasks involved, including web scraping, tokenizing, POS tagging, indexing and ranking. The BM25F ranking algorithm is an obvious candidate for our ranking algorithm because recipes naturally translate well into structured documents (eg. list of ingredients; title; description; cook time; etc.). As alluded to earlier, our document collection will be aggregated from multiple sources. One obvious early source will be a Kaggle dataset of 250,000 food recipes, but we hope to find additional recipes through other sources or from scraping food recipe websites.

We hope to dynamically update our background language model in some fashion based on user feedback, but the exact formula/algorithm we will employ is still a matter under consideration. For our push recommendation system, we hope to test and compare both content and collaborative models of filtering. We will implement the better performing model for our final submission or consider implementing a combination of the two models.

Validation (*How will you demonstrate the usefulness of your tool?)*

For the validation phase, we will solicit 10 volunteers to generate and run their own set of 10 queries and make explicit relevance judgments about the top 10 ranked results for each query. Five of the queries will be fed to competing search engines with top ten results returned and judged by our users. Ideally, this will be done at random so as to obfuscate to the user which search engine they are using at any given time. We will be using statistical methods to evaluate our search engine and compare with competing engines (using precision and recall as the primary metrics).

Once the queries and judgments are complete, we will feed our participant's user data into our recommendation system and send an email to each of the participants with five recipe recommendations. This email will ask for relevance judgments from each of the five recipes. We will also send another set of emails based on a collaborative model once all participants have completed their queries and associated relevance judgments.

3. ***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.***

Project Timeline (A very rough timeline to show when you expect to finish what.):

Complete Proposal (By Oct 25)
<u>Architecture document with API level information (Python) and Test Scenarios</u> (By Nov 1, 12 hours expected)
  Architecture block diagram showing end2end pipeline (Interface to search results representation)
  High level Input and Output data for each function block
  Sample TestCase ("Queries") and TestResults ("Expected")
Document collection formation: tokenizing, POS tagging, indexing, etc. (By Nov 1)
  (*20 total hours expected to aggregate our document collection for Indian Cuisine recipes and associated review/comments*)
  <u>Associated preprocessing tasks</u> (*2 hours expected to tokenize, POS tag, build out the document collection, and build an inverted index*)

Recipe Classification (By Nov 20)

Classify recipes by cuisine style, meal type, dish type, etc. (20 hours)
This process will help provide filters for advanced search functionality, and could also be integrated into our background language models.

Develop test queries and test/select optimal scoring function (By Nov 8)

Develop test queries (*4 hours expected*)
Test/select/optimize scoring function (*3 hours expected*)

Develop user profile and feedback model (By Nov 15)

Research, build and test/select optimal user feedback model for recommendation system (*15 hours expected*)

Working user interface (By Nov 22)

Locally hosted Web Interface/Python GUI with search button and results window (*20 hours expected*)

Trial Phase (Nov 22-Nov 29)

Build automated system to pull query results from competing search engines *(5 hours expected)*
Solicit volunteers for our application *(4 hours expected to solicit volunteers, explain the tool and the participation requirements*)
Evaluate and document/model query relevance judgments (1 hour expected)
Evaluate push recommendation relevance judgments and adjust recommendation model (2 hours expected)

Progress report completed (By Nov 29)

Final software code with documentation and software presentation (By Dec 9)

Debug/test source code (2 hours expected)
Construct Readme (2 hours expected)
Create presentation (4 hours expected)

Total Hours expected: 116 hours
Total Hours required: 80 hours

4. ***Which programming language do you plan to use***?
Primarily Python

Additional Details and Expected Outcome:

We expect that our tool will outperform competing search engines within our specified topic/domain of Indian food recipes.