CSE 435/535: INFORMATION RETRIEVAL

PROJECT 4: Multi-topic Information Retrieval Chatbot

Final Deadline: 7th Dec, 11:59 PM ET

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Overview of previous projects

- The first 3 projects dealt with:
  - *Project 1*: Indexing and Crawling
    - How do you gather data from a particular reddit thread? How do you retrieve the comments?
    - How do you effectively index this data using Solr?
  - *Project 2*: Scoring - How does query scoring work?
  - *Project 3*: Relevance - How do you tune relevance for specific information needs?
  - *Project 4*: Seeks to unify these subtasks into a single end-to-end IR chatbot.

Datasets

- The data that you have collected in Project 1, please index the (<submission>, <comment>) or (<comment>, <comment>) pairs.
- Along with that you can use: https://github.com/BYU-PCCL/chitchat-dataset , this is a chit-chat corpus which will enable your chatbot to reply to general utterances like: “How are you?”, “What is your name?”, etc.
- You are free to use/collect more data.

Project Goal

Basic Requirements:

1. Functional Chatbot
   - Based on the described datasets build a fully functional IR chatbot which is capable of carrying out at least 6 turns of coherent utterances.
   - Develop a full functional chat UI and host a Web App.
2. Topic Mode
   - Your chatbot should be able to able to converse in the defined topics(***Politics, Environment, Technology, Healthcare, and Education***) as introduced in Project 1.
   - Use faceted search feature to restrict your chatbot to converse in a particular topic.

Advanced Requirements:

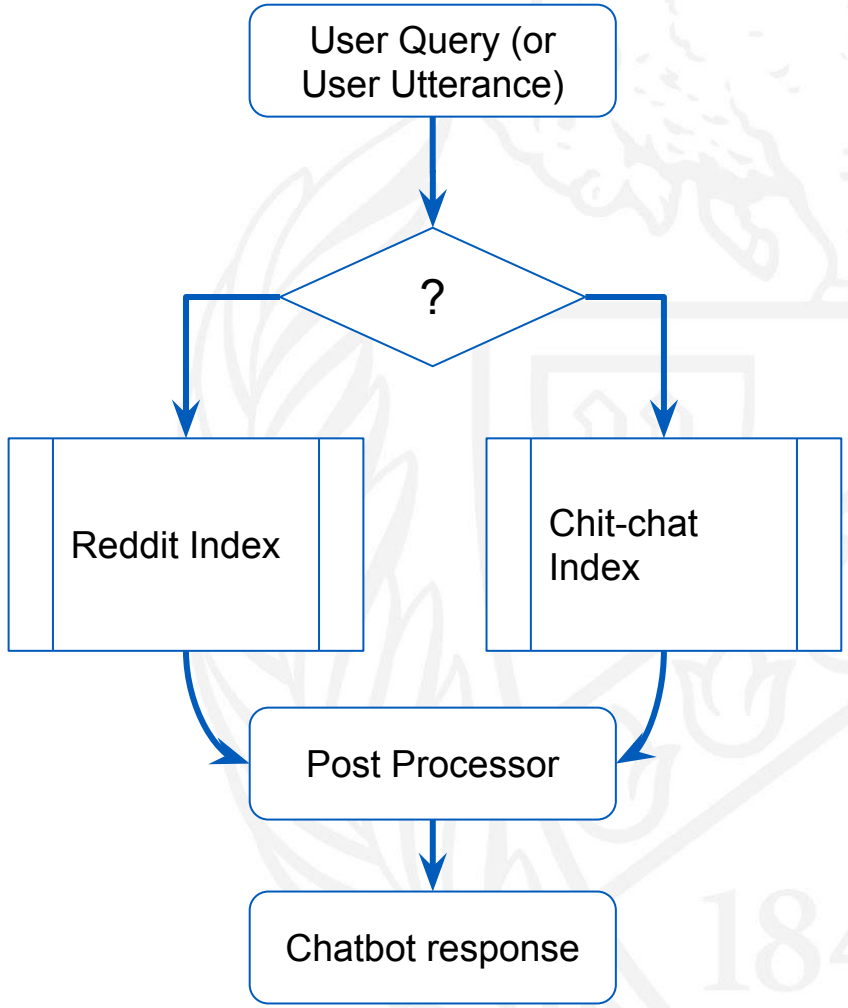
- Your chatbot is able to coherently converse for >= 15 turns.
- Detailed analysis/visualization of chats, across topics, entities etc.

Groups and Dataset Sharing

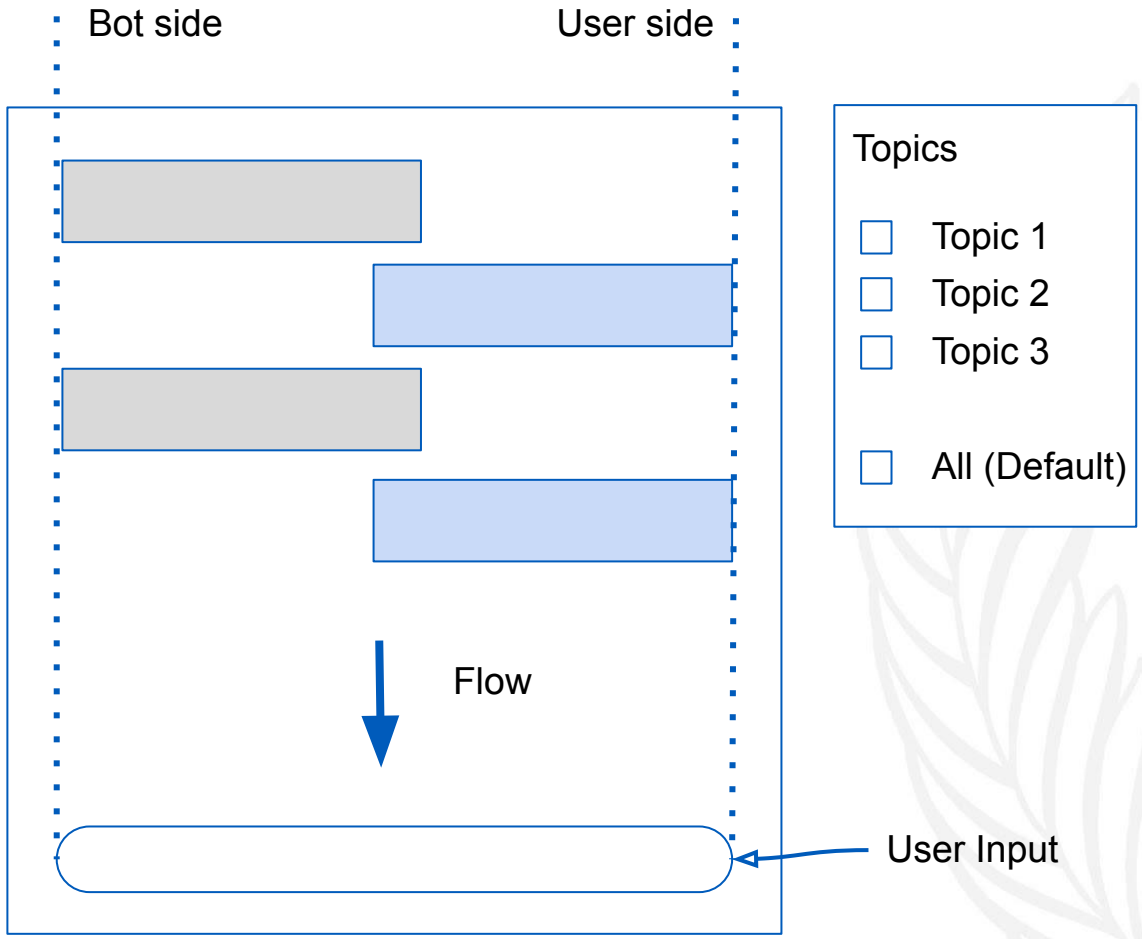
- You need to form your own groups of 3-4 members.
- Sign-up your team using the Google Form(https://forms.gle/4myB416vE4XLLUWJ6) posted on Piazza before 10th Nov, 8 PM EST.
- You are allowed to share your data within the group.
- You are free to collect more data.

End to End IR Chatbot

- Build End-to-End IR chatbot pipeline as shown beside.
- Decide which index to query based on the user utterance, some ideas are :
  - Write explicit rules.
  - Train a classifier.
- Option to write a post-processing module, after you have obtained the results from the indexes, some ideas are:
  - Rerank the results.
  - Combine results.



Chatbot UI and Faceted Chat



Visualization and Analytics Ideas

- Save N number of conversations for analytics and visualization.
- Main purpose is to understand how well your chatbot is able to converse in the defined topics and how diverse are the responses.
- Understand which topics the chatbot is able talk more about and where most of the errors are occurring.
- Also analyse the user queries and what type of queries are fetching more relevant utterances.
- Based the previous point can we group certain queries? Do query reformulation increases the chat coherency?
- Be creative and come up with your own ideas.

Final Deliverables

- A short demo video (at most 3 minutes)
- A working web application URL hosted on GCP
- A short report detailing all work done and member contributions.
  - You can use the double column ACL 2022 or single column ICLR 2022 Latex template.
  - You can also use word, if you are not comfortable with Latex.
  - The report should contain the following broad sections: *(i) Introduction, (ii) Methodology (iii) Sample screenshots (iv) Work breakdown by teammates (v) Conclusion*
  - More details on how to submit will be shared closer to the deadline.

Grading

- Grading is based on relevancy of chat, duration, ranking techniques and topic coverage.
- Points distribution (total **30 points**):
  - Meet basic requirements – **22 points**
    - UI and basic chatbot functionality – **12(6+6) points**
    - Topic mode/ faceted chat – **5 points**
    - Chat coherence and duration – **5 points**
  - Meet advanced requirements – **5 points**
  - Report – **3 points**
- We will select best performing groups to present their work in the class
- 7 groups will be selected to present their work in 8 minutes with additional 2 minutes for Q&A
- Each team member of the selected groups will receive **2 bonus points**.
- More details to be shared later.

Timeline

1. 7th November: Project released
2. 10th November: Deadline for team formation.
3. 4th December, *before 5 PM*: Interested groups submit videos for class presentations. Sign-up sheet will be released 3 days before.
4. 5th December: In-class presentation for selected groups (2 bonus points).
5. 7th December: Final submissions due.

Resources

- Machine learning / clustering / topic modelling:
  - Python : Scikit-learn, nltk (NLP specific)
  - Java : Spark/Mahout, Weka, Mallet
  - C++ : Shogun, mlpack
- Word embeddings (pre-trained)
  - http://nlp.stanford.edu/projects/glove/
  - Pointers to download links: https://www.quora.com/Where-can-I-find-some-pretrained-word-vectors-for-natural-language-processing-understanding
- Translation : Google and Bing APIs, several free to download dictionaries

Resources

- Visualization / analytics examples and ideas
  - http://www.tableau.com/stories/gallery
  - https://www.census.gov/dataviz/
  - https://app.powerbi.com/visuals/
  - https://github.com/d3/d3/wiki/Gallery
  - https://developers.google.com/chart/interactive/docs/gallery
  - https://developers.google.com/chart/interactive/docs/more_charts