Information Retrieval - 2021-22 Autumn Semester
Term projects

7 term projects are given below. The last one is a “default” project that specifically involves the concepts discussed in the class. The other projects are more research-oriented. Every project has a project code and an associated mentor (one of the TAs of the course).

- Students need to form their own groups. Each group should have 3 or 4 students and work on one project chosen by the group.

- One member from each group should mail TA Abhisek Dash (assignmentad@gmail.com) -- (i) the names and roll numbers of all the group members, and (ii) the code of the chosen project -- by Saturday (Sep 25) end of day. All group members must be cc-ed in the mail. The subject line of the email should be **“IR Term project choices".**

- If students have questions about any project, they can mail the corresponding mentor.

- **Every project will have two evaluations.**
  (1) An intermediate evaluation during October 20--22, weightage will be 60% for the default project and 50% for the other projects.
  (2) The final evaluation during November 14--16, weightage will be 40% for the default project and 50% for the other projects.

==========

Project Code: WebQA Mentor: Paramita Koley

WebQA - a NEURIPS 2021 Competition

WebQA is a new benchmark for multimodal multihop reasoning in which systems are presented with the same style of data as humans when searching the web: snippets and images. Upon seeing a question, the system must identify which candidates potentially inform the answer from a candidate pool. Then the system is expected to aggregate information from selected candidates with reasoning to generate an answer in natural language form. Each datum is a question paired with a series of potentially long snippets or images that serve as "knowledge carriers" over which to reason. Systems will be evaluated on both supporting fact retrieval and answer generation to measure correctness and interpretability. To demonstrate multihop multimodal reasoning ability, models should be able to 1) understand and represent knowledge from different modalities, 2) identify and aggregate relevant knowledge fragments scattered across multiple sources, 3) make inference and do natural language generation.

Link Data

==========

Project Code: CODWOE Mentor: Paramita Koley

CODWOE: COmparing Dictionaries and WOrd Embeddings

The CODWOE shared task invites you to compare two types of semantic descriptions: dictionary glosses and word embedding representations. Are these two types of representation equivalent? Can we generate one from the other? To study this question, we propose two subtracks: a **definition modeling** track (Noraset et al., 2017), where participants have to generate glosses from vectors, and a **reverse dictionary** track (Hill et al., 2016), where participants have to generate vectors from glosses.

Dictionaries contain definitions, such as Merriam Webster's:

cod: *any of various bottom-dwelling fishes (family Gadidae, the cod family) that usually occur in cold marine waters and often have barbels and three dorsal fins*

The task of definition modeling consists in using the vector representation of $\vec{\text{cod}}$ to produce the associated gloss, "*any of various bottom-dwelling fishes (family Gadidae, the cod family) that usually occur in cold marine waters and often have barbels and three dorsal fins*". The reverse dictionary task is the mathematical inverse: reconstruct an embedding $\vec{\text{cod}}$ from the corresponding gloss.

These two tracks display a number of interesting characteristics. These tasks are obviously **useful for explainable AI**, since they involve converting human-readable data into machine-readable data and back. They also have a **theoretical significance**: both glosses and word embeddings are also representations of meaning, and therefore involve the conversion of distinct non-formal semantic representations. From a practical point of view, the ability to infer word-embeddings from dictionary resources, or dictionaries from large unannotated corpora, would prove **a boon for many under-resourced languages**.

https://competitions.codalab.org/competitions/34022

Data

==========

Project Code: LEGSUMM Mentor: Paheli Bhattacharya

Legal document summarization

Description : Summarizing lengthy legal documents is an important problem to solve. In this project, the task is to summarize Indian Supreme Court case documents and UK Supreme Court case documents, using the following methods:

a. Discourse-Aware Unsupervised Summarization of Long Scientific Documents [Paper] [code]

b. Unsupervised Extractive Text Summarization with Distance-Augmented Sentence Graphs [Paper] -- only single-document summarization part; implementation to be done by yourself

==========

Project Code: LEGSTAT Mentor: Paheli Bhattacharya

Legal Statute Retrieval

In countries following the Common Law system (e.g., India, UK, Canada, Australia, and many others), there are two primary sources of law -- (a) Statutes which are the written laws (e.g. IPC Section 302, Constitution Article 19) and (b) Precedents or judgements of prior cases delivered by a court, which involve similar legal facts and issues are the current case, but are not directly indicated in the written law.
While working on a new case a legal practitioner often relies on these statutes and precedents to understand how the Court has discussed, argued and behaved in similar scenarios. This task is aimed at creating retrieval systems capable of addressing this problem

Task Description : Given a query (short description of a legal situation), identify relevant statutes (laws e.g. IPC Section 302, Constitution Article 19), from a pool of 197 statutes. For training, you will be provided 50 queries and their relevant statutes. For test, you will be provided 10 queries, on which your method will be evaluated.

This was a shared task at Artificial Intelligence for Legal Assistance (AILA). Several approaches have been tried out. You can refer to the leaderboard here (Table 2).

==========

Project Code: LEGPREC Mentor: Paheli Bhattacharya

Legal Precedent Retrieval

In countries following the Common Law system (e.g., India, UK, Canada, Australia, and many others), there are two primary sources of law -- (a) Statutes which are the written laws (e.g. IPC Section 302, Constitution Article 19) and (b) Precedents or judgements of prior cases delivered by a court, which involve similar legal facts and issues are the current case, but are not directly indicated in the written law.
While working on a new case a legal practitioner often relies on these statutes and precedents to understand how the Court has discussed, argued and behaved in similar scenarios. This task is aimed at creating retrieval systems capable of addressing this problem

Task Description : Given a query (short description of a legal situation), identify relevant prior case documents from a pool of 3257 documents. For training, you will be provided 50 queries and their relevant precedents. For test, you will be provided 10 queries, on which

your method will be evaluated.
This was a shared task at Artificial Intelligence for Legal Assistance (AILA). Several approaches have been tried out. You can refer to the leaderboard here (Table 1).

==========

Project Code: RUMOR Mentor: Rajdeep Mukherjee

Rumour Detection from Social Media Conversations

Detecting rumours or unverified claims from social media conversations is a crucial task for protecting netizens from the detrimental consequences of misinformation. In this project, each team will be randomly assigned 1 paper out of the following 3 papers:

Paper 1: Rumor Detection on Twitter with Tree-structured Recursive Neural Networks [Link] [Github]
Paper 2: Tree LSTMs with Convolution Units to Predict Stance and Rumor Veracity in Social Media Conversations [Link][Github]
Paper 3: Cascade-LSTM: A Tree-Structured Neural Classifier for Detecting Misinformation Cascades [Link][Github]

Common task for all teams: Incorporate *stance classification* from "All-in-one: Multi-task Learning for Rumour Verification" [Link][Github] into Tree LSTM-based *rumour detection* from "Going Beyond Content Richness: Verified Information Aware Summarization of Crisis-Related Microblogs" [Link][Github]

You are required to obtain the results for the assigned paper and the common task on the PHEME-RNR dataset.

==========

Project Code: DEFAULT Mentor: Rajdeep Mukherjee

The default project consists of two parts - Part-A and Part-B.

- Part-A:
  Task 1: Building an Inverted Index and Boolean Retrieval - Problem Statement
  Task 2: TF-IDF vectorization and Evaluation - Problem Statement
  Weightage: 60%
  Deadline: 20th October 2021, 11:59 pm

- Part-B:
  Relevance Feedback - Problem Statement
  Weightage: 40%
  Deadline: 14th November 2021, 11:59 pm

The same dataset is to be used for both the parts.