Project 2

Background

Thanks to technological advances, animal geneticists have an ever-expanding tool chest with which to study the inheritance of traits in livestock in order to improve production. With the rise of high-throughput technology, vast amounts of genotype/phenotype data are being rapidly generated. Animal QTLdb and other genotype/phenotype databases would greatly benefit from automated and expedited curation tools.

In this project, we want to identify relevant papers retrieved through keyword searching. Here, the relevant papers mean that these papers contains important geno-trait results that can be curated in the QTLdb.

Data

You are given two datasets: QTL_text.json and QTL_test_unlabeled.tsv
both files are available in /work/classtmp/NLP/project_data

QTL_text.json is in the following format (all fields are strings)

{"PMID":
"Journal":
"Title":
"Abstract":
"Category":}

Each entry in this json file corresponds to a paper

1. "Title": this is the title of the paper
2. "Abstract": this is the abstract of the paper. Usually it contains many sentences.
3. "Category": this is either '0' or '1'. '0' means the paper is not related to animal QTLdb curation, and '1' means the paper is related to animal QTLdb curation. In this project, this is the label.

QTL_test_unlabeled.tsv is a tab-separated values (TSV) file where each row represents a research paper. It consists of four columns with the following data types:

1. PMID (int): A unique numerical identifier assigned to each publication in the PubMed database.
2. Title (string): The title of the research paper
3. Abstract (string): this is the abstract of the paper. Usually it contains many sentences.
4. Label (int): An integer value, which is currently unlabeled (e.g., a placeholder 0 for future classification). 0 means the paper is not related to animal QTLdb curation, and 1 means the paper is related to animal QTLdb curation.

Each field is separated by a tab (\t).

Pre-processing and other requirements

You will need to pre-process the QTL_text.json to the proper format and split it into a training and development set if needed. You can use any existing libraries/implementations for this project. Please add a readme file in your submission so that we can run your code.

Task 1

Train a document classifier to predict the "Category" of each paper in the QTL_test_unlabeled.tsv file.

Submission

You will need to submit the prediction result in Kaggle.com.

Kaggle Competition Invitation Link: https://www.kaggle.com/t/bff80d3fde1a499189c10d9670e61578

Your submission should look like “sample_submission_random.csv”, which contains a header and has the following format:

PMID,Label

34902587,0

35268189,1

35451025,0

etc.

Note that your QTL_test_unlabeled.tsv is a .tsv file (‘\t’ separated), but your Kaggle submission needs to be .csv file (comma separated). For more details, please check the Kaggle Competition page.

Your grade is partially based on the score of private leaderboards. See the following link for more information. https://www.kaggle.com/docs/competitions#leaderboard

You will also need to submit the following documents on Canvas

1. your code (optional: readme): if your code is available on Nova, please enter the path in comment and no need for submission
2. a report (max 2 pages) in pdf using this template: https://www.overleaf.com/latex/templates/association-for-computational-linguistics-acl-conference/jvxskxpnznfj

Your report needs to include:

Title: project 2

Author: your name, email, and Kaggle username (as it appears on the Kaggle Competition leaderboard)

Method: describe the method you used.

Discussion (optional): If you trained several models, you can discuss your findings.

Appendix (optional, does not count into the 2-page limit): Include links to all tutorials/websites/github that you referred to for this project. If you used AI assistant for coding, please include screenshots.