Project 1

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. While it is well recognized that the adoption of common controlled vocabularies and ontologies facilitates data integration and reuse, it is nontrivial to extract such terms from scientific texts automatically.

In this project, we want to analyze the terminologies used in the collected corpus and understand the trend of studies.

Data

You are given two datasets: QTL_text.json and Trait dictionary.txt

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

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

Each entry in this json file is an abstract of a paper. In this project, you will need to use "Abstract" and "Category", and you can ignore the other fields.

1. "Abstract": this is the abstract of the paper. Usually it contains many sentences.
2. "Category": this is either '0' or '1'. '0' means the paper is not related to animal QTL, and '1' means the paper is related to animal QTL. In this project, you need to ignore papers in Category '0'.

Trait dictionary.txt:

Each line in this file is a trait term we collected from domain dictionaries. Note that not all terms appear in the QTL_text.

Pre-processing and other requirements

You will need to preprocess the text with 0) collect all abstracts with Category:1, 1) split sentences and tokenize, 2) convert to lower case, and 3) remove stop words.

You can use any existing libraries/implementations for this project. If you use a github repo, please add a readme file in your submission so that we can run your code.

Task 1

Use wordcloud to visualize words in this corpus. The figure should be 800*800, with white background color. You will need to generate two word cloud images: 1) use word frequency, and 2) use tf-idf

You may find the following libraries useful for this task: WordCloud

Task 2

Train a Word2Vec model on this corpus, with the following parameters

```
vector_size=100, window=5, min_count=10
```

For each of the top 10 tf-idf words, print the 20 most similar words.

Task 3

Extract phrases and repeat task 1 and 2. You can be creative in phrase extraction.

Submission

You will need to submit on Canvas

1. your code (optional: readme)
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 1

Author: your name and email

Method: describe the phrase mining method you used.

Main result:

1. Comparing the word cloud images, what observation do you have? What insight can you provide?
2. Comparing the word2vec results (you do not need to include the full lists in the report), do you think the results are correct? List some positive examples and some negative examples. What interesting observation do you have? What insight can you provide?
3. How many phrases you extracted can be found in the trait dictionary (by exact string matching)?

Discussion (optional): If you tried some other experiments in additional to the required tasks, you can discuss your findings. For example, if you tried more than one phrase mining method, you can discuss their performance here. Or if you found some methods are surprisingly slow/fast, you can also discuss it here.

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.