Project 1

Moulica Vani Goli
moulica9@example.com

Abstract

We preprocess abstracts from QTL_text.json, generate word clouds based on word frequency and TF-IDF, train a Word2Vec model to capture semantic associations, and extract key phrases using a bigram collocation approach. Extracted phrases are evaluated by matching against a domain-specific Trait dictionary. Our analysis reveals insights into domain-specific vocabulary and the semantic structure of the corpus.

1 Method

1.1 Data Preprocessing

We begin by filtering the abstracts from **QTL_text.json** to retain only those with Category "1". Each abstract is segmented into sentences using NLTK's sentence tokenizer. The sentences are tokenized into words, converted to lowercase, and cleaned by removing stopwords. These processed sentences serve as documents for further analysis.

1.2 Word Cloud Generation

Two word clouds are generated:

- **Word Frequency Cloud:** Displays overall term frequencies.
- **TF-IDF Cloud:** Emphasizes terms that are statistically significant in the corpus.

1.3 Word2Vec Training

A Word2Vec model is trained on the tokenized corpus with parameters: **vector_size = 100, window = 5, and min_count = 10.** For the top 10 TF-IDF words, the model returns the 20 most similar words, allowing us to assess the semantic associations captured by the model.

1.4 Phrase Mining

For phrase extraction, we employ a bigram collocation approach using NLTK's **BigramCollocationFinder**. We tokenize the entire corpus and compute the likelihood ratio for each bigram. The top 50 bigrams are selected as candidate phrases by joining word pairs with a space. These candidate phrases are evaluated through exact string matching against the provided Trait dictionary **Trait dictionary.txt**.

2 Main Results

2.1 Word Cloud Comparison

The word frequency word cloud provides a broad view of common terms across the corpus. In contrast, the TF-IDF word cloud highlights more domain-specific vocabulary, such as "qtl", "snps", and "traits". This indicates that terms with high TF-IDF scores capture important, less frequent domain-specific concepts.

2.2 Word2Vec Evaluation

The Word2Vec model generally returns semantically relevant similar words.

- **Positive Examples:** For "snps", similar words such as "snp", "polymorphism", and "intron" were returned.
- **Negative Examples:** Occasionally, generic words like "found" or "one" appear, which are less informative.

These results suggest that while the model effectively captures domain-specific semantic relationships, further tuning or an expanded training corpus might improve precision.

2.3 Phrase Extraction Evaluation

Out of the 50 candidate phrases extracted using our bigram method, 7 phrases match exactly with entries in the Trait dictionary (e.g., "fatty acid", "body weight", "litter size", "backfat thickness", "intramuscular fat", "abdominal fat", "milk yield").

3 Discussion

Additional experiments were conducted using alternative phrase mining methods, such as n-gram frequency filtering and part-of-speech (POS) tagging. Although

these methods generated slightly different candidate phrases, the bigram collocation approach provided the most robust balance between recall and precision for extracting domain-specific phrases. Future work will explore more sophisticated techniques (e.g., dependency parsing) to further enhance phrase extraction quality.

4 Conclusion

This project presents a comprehensive pipeline for the automated curation of animal QTL abstracts. Through preprocessing, visualization, semantic modeling via Word2Vec, and phrase extraction evaluated against a Trait dictionary, our approach reveals key domain-specific insights. These methods can be further refined to improve the accuracy of automated literature curation in genomic research.