Bee Species Prediction Using Feature Specifications in Text Format

Tejas Dubhir (2018110) Shivangi Dhiman (2018265)
Indraprastha Institute of Information Technology, Delhi

Motivation

Bees play a significant role in the pollination of food crops and the production of honey and wax; hence are a crucial ecological and economic resource. There are over 20,000 known species of bees around the globe and each species has its own nique characteristic traits. However, many species possess similar characteristics which can make the task of bee identification or bee taxonomy difficult.

Objective

Due to such a large number of known species, it is infeasible to manually identify and classify all the bees, based on their physical traits. Thus, our goal is to create an automated bee identification system, using Information Retrieval, Machine Learning and Natural Language Processing techniques.
Using a list of features as queries/inputs, we aim to predict the species of the bee.

Data Collection & Data Description

Since there was no availability of a database of bee species along with a description of their physical traits, we collected and compiled our own dataset.

- We used two books - 'The Bees of the World' [3] and 'Apis & The Fauna of British India, Including Ceylon and Burma. Hymenoptera (Vol. 1) Wasps and Bees' [2]. We implemented **Image to Text conversion** methods using pytesseract and ocrmypdf.
- We also scraped data from several websites and online resources [5,7,8,9,10,11]
- Finally, we compiled a list of 102 species of bees along with their physical description.
- The data was entirely in text format and consisted of two fields - 'Title' and 'Description'

There were 1842 distinct words in our dataset. Most common words- genus, species, bee, female.

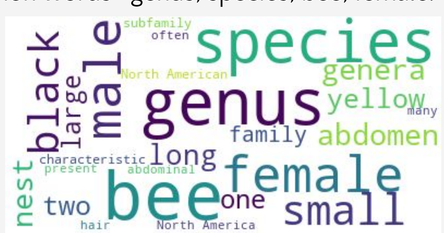

Fig 1: 25 most frequently used words

Methodology

1) Data Preprocessing:

a) Tokenization
b) Lower Casing
c) Removing extra spaces
d) Removing stopwords
e) Lemmatization

2. Rank Prediction [1,4]

- Term Frequency - Inverse Document Frequency (TF-IDF) is numerical statistic that is used to represent the importance of a word in a document in a collection or corpus.
- It is an important tool for scoring and ranking a document's relevance given a user query (in our case, features of the bee).

We used the following 3 TF-IDF methods:

a. **Raw TF-IDF:** TF × IDF
b. **Binary TF-IDF:** Binary TF × IDF
c. **Log-Normalised TF-IDF:** Log-norm TF × IDF

The final rank list of 15 of the most probable species was prepared by combining the outcomes of these 3 algorithms.

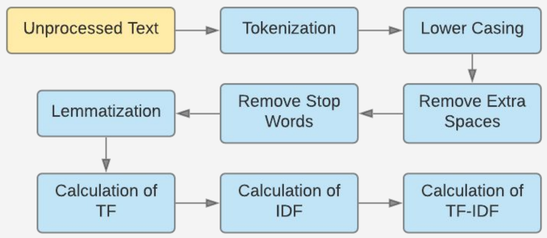

Fig 2: Data Preprocessing & TF-IDF Calculation

3. Machine Learning Models [6]

The following models were run to classify the vectorized text (feature description) into categories (species) and their performance was noted:

a. Multinomial Naive Bayes
b. SVM Classifier
c. Random Forest Classifier
d. K-Neighbour Classifier
e. Logistic Regressor
f. Multilayer Preceptron Classifier

Results

- The accuracy obtained with each model has been reported below:

Model	Accuracy (in %)
Multinomial Naive Bayes	77.78
SVM Classifier	77.78
Random Forest Classifier	55.54
K-Neighbour Classifier	11.12
Logistic Regression	88.89
MLP Classifier	77.78

- Logistic Regression gave us the best accuracy of **88.89%** and K-Nearest Neighbour Classifier gave us the lowest accuracy of 11.12%. (Fig 3)

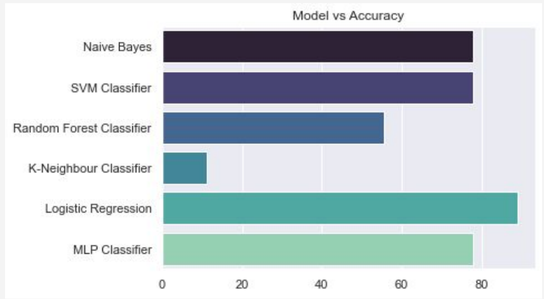

Fig 3: Models and their Accuracies

- We also created a webapp using Flask (bees-identifier.herokuapp.com) that shows the working of our model.
- The user will enter a list of features of a bee and the output will be a list of 15 species that best fit the description. The list is ranked in the descending order of the likelihood.

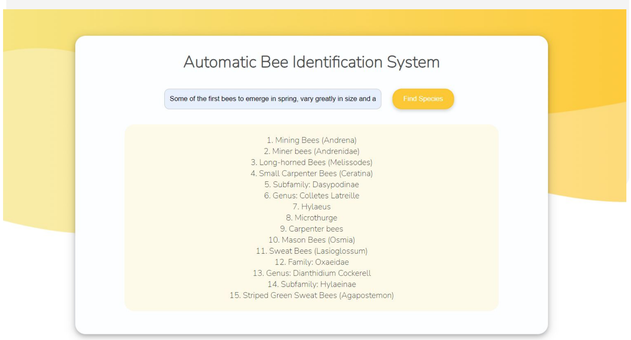

Fig 4: Bee Species Identification - WebApp

Future Work

- Compilation of a larger database for better results
- Bi-grams or Multi-grams indexing for ranked retrieval.
- Computer Vision techniques to read and identify the species of bees from an image.
- Use of deep learning models like LSTM and RNNs for better predictions.

Source Code

- The source code is available in this GitHub repository.
- Instructions to set up the environment and perform predictions are present in the README file.

References

[1] B. Armouty and S. Tedmori. 'Automated keyword extraction using support vector machine from Arabic news documents'. In: 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). IEEE. 2019, pp. 342–346.
[2] C. Bingham. 'The Fauna of British India, Including Ceylon and Burma. Hymenoptera-Vol.1. Wasps and Bees.' In: (1897).
[3] C. D. Michener. The bees of the world. Vol. 1. JHU press, 2000.
[4] J. Ramos et al. 'Using tf-idf to determine word relevance in document queries'. In: Proceedings of the first instructional conference on machine learning. Vol. 242. 1. Citeseer. 2003, pp. 29–48.
[5] About Bees, https://idtools.org/id/bees/exotic/bees classification.php.
[6] Alper. NLP: Classification and Recommendation Project. July 2020, towardsdatascience.com/nlp-classification-recommendation-project-cae5623ccaae.
[7] S. Bob. HYMENOPTERA Bees, Wasps, Sawflies &; Ants. May 2014, bobs-bugs.info/2014/01/02/hymenoptera-bees-wasps-ants-etc.
[8] S. Kincaid. Common Bee Pollinators of Oregon Crops. 2017, oregon.gov
[9] Palaearctic Osmiine Bees, blogs.ethz.ch/osmiini/phylogeny-and-classification.
[10] The Most Beneficial Types of Bees (With Identification Guide and Pictures). Apr. 2021, leafyplace.com/types-of-bees.
[11] Wood Wasps and Sawflies, amentsoc.org/insects/fact-files/orders/hymenoptera-symphyta.html.

Acknowledgements

1. Dr Swapna Purandare for guiding us throughout the project.
2. Gursimran Kaur for helping us with TF-IDF implementation.