An Ensemble Recognition Algorithm for Duplicated Product Post on Shopee

## Group21-Members

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | Qinren Zhou | Lu Zhang | Yongrui Chen | Zhaoyuan Qiu | Zhibao Li |
| **NetID** | Qz142 | Lz468 | Yc910 | Zq37 | Zl424 |
| **Project Github** | <https://github.com/ZhaoyuanQiu/ANLY590.git> | | | | |

## Project Goal & Objective Summary-精简

Our goal is to achieve an algorithm which can automatically find out the original post for a given product post, if any previous post pointing to the same item is found, and returns FALSE if the product is posted for the first time.

This model can be used to help prevent duplicate post on e-shopping websites and thus improve product posting efficiency. This algorithm can also be used to improve recommending system and reduce duplicated recommendations for users.

## Proposed Data Source

Our data sources from [Kaggle](https://www.kaggle.com/competitions/shopee-product-matching/overview)

两行简单介绍一下数据来源，数据类型，数据量，数据用途

## Methods

To accurately predict the label for a given , we will define a similarity score to measure the similarity between and any other in the dataset. Then we return the most similar post. Since the original data consists of image data and text data, the task can be divided into 2 parts. Calculating image similarity score and calculating text similarity score . And then we combine the 2 metrics via ensemble method and get the final similarity score .

To calculate score , we can apply pre-trained model, which maps images to dense vectors. The similarity between two images can be obtained by calculating the cosine similarity between their embedded vectors. The calculate score , first we can use Bag-of-Words or TF-IDF value to digitize the text data. Then we apply Latent Semantic Index (LSI) to compute the similarities between text data. We will compare different approaches including SVD, NMF and LDA.

**Modeling Workflow**

The final stop is how to determine the weights and . We can apply general linear regression with softmax to get predictions, then do gradient descent with as loss function. The work scheme is as below,

|  |  |  |  |
| --- | --- | --- | --- |
| **Phase 1**  Preparing Data | **Phase 2**  Similarity Score Modeling | **Phase 3**  Combining Metrics | **Phase 4**  Test & Evaluation |
| Image Data Standardization | Transfer Pre-trained Model | GLM and Gradient Descent | Tuning Hyperparams  Writing Reports |
| Text Data Cleaning | Topic Models & LDA |

## Expected Results-精简

With the input of product post with text descriptions, our model is expected to identify if there is any repeated or similar posts. If there is a duplicate, the model will return the IDs of the first n duplicate posts with the highest probability, otherwise return False.

## Reference

搞一下Refer的格式，用超链接插入文字节省空间，随意添加觉得切题的资源

<https://www.tensorflow.org/hub/common_signatures/images#feature-vector>

<https://www.kaggle.com/code/hamditarek/similar-image-cnn-cosine-similarity#4.-Finding-the-similar-image-through-LSH-and-cosine-similarity>

<https://peltarion.com/blog/data-science/image-similarity-explained>