**Group Name:** 400 Bad Request

**Group Members:** Zhengqian Xu, Hongyang Zheng, Heng Zhou

**NLP Project2 – Product Search Relevance**

**Motivation:**

When a user wants to buy something online, they might not be able to input the exact name for that product, and the company needs to figure out which products are most relevant to the words the user inputs and return a page of products with high relevance. For example, if a user wants to buy some “AA battery”, then the website should return “pack of size AA batteries” or at least “cordless drill battery” if they don’t have AA battery, rather than “snow shovel” which is totally irrelevant. Based on this scenario, our group wants to use some NLP techniques to calculate the relevant score between what a user enters (key words) and what the company has (products), with returning the top 10 most relevant products to the users.

The reason why this question is worthy to investigate is that we can help the online website match customers’ needs by analyzing the key words. For example, if the user searches something that we don’t have, we need to introduce new products. The more accurately a company can satisfy the customers’ demand; the more business profits/customers a company can gain.

**Relevant Scientific Literature**

The original idea for this project comes from a paper [1], in which the author describes the question about matching similar products using their title, description and image. Due to the complexity and lack of data, we choose a similar but easier topic. However, this paper provides us with some useful hints about how to process the data. For example, in that paper, the author uses dictionary-based method, conditional random field (CRF) model, CRF with text embeddings to extract features; applies cosine similarity on TF-IDF, paragraph2vec and Silk as products matching baseline; build random forest, SVM, Naïve Bayes and logistic regression to solve the classification problem. Other papers [2-4] also provide helpful guidance for this project implement.

**Data**

The data is from Kaggle (<https://www.kaggle.com/c/home-depot-product-search-relevance/data>), which contains a number of products and real customer search terms from Home Depot's website. The training data has more than 70,000 rows and each row has a ‘product unique id’, ‘product title’, ‘search terms’ as well as ‘relevance score’. There are also other supplementary datasets containing detailed information about each product such as product description, product value, product image and product URL. We will mainly focus on product description and product title and use product value if we need more features.

**Plan**

The overall plan for this project is to predict relevance score for the provided combinations of search terms and products, validate our results using the given true labels, and finally apply the model to the test data. Since the score is in the range [1, 3], we need to build a regression model rather than a classification model. Besides that, we may also try to calculate the cosine similarity score for each combination of keyword and product as a baseline.

Therefore, our project may follow the next steps:

1) Extract useful information from product title and product description, such as brand, material, functionality, color, size and other features which may be included in the searching keywords.

2) Preprocess to normalize our data such as lowercase and stemming/lemmatization.

3) Create feature vectors using proper NLP methods.

4) Calculate cosine similarity for each pair.

5) Build regression model to predict relevance score.

6) Evaluate the model and improve the performance.

7) Apply the final model to test data.

Roughly, each team member will be responsible for 2 steps. The tentative plan is Zhengqian for 1) and 4), Heng for 2) and 5), Hongyang for 3) and 6). However, we will keep in touch and talk with each other throughout the process.

**Expected Results**

By the end of the project, we will successfully train a model to predict a relevance score for each pair of search keywords and the product. We expect that we can use the results to develop a mini application when a user inputs some keywords, we can return the top 10 (if applicable) most relevant products to him/her based on the test dataset.

**References**

[1] Petar R. Petar P., Peter M., Heiko P.: A Machine Learning Approach for Product Matching and Categorization, <http://www.heikopaulheim.com/docs/swj2018_product_data.pdf>

[2] Petrovski, P., Bryl, V., Bizer, C.: Learning regular expressions for the extraction of product attributes from e-commerce microdata (2014)

[3] Ghani, R., Probst, K., Liu, Y., Krema, M., Fano, A.: Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter 8(1), 41–48 (2006)

[4] Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. arXiv preprint arXiv:1405.4053 (2014)