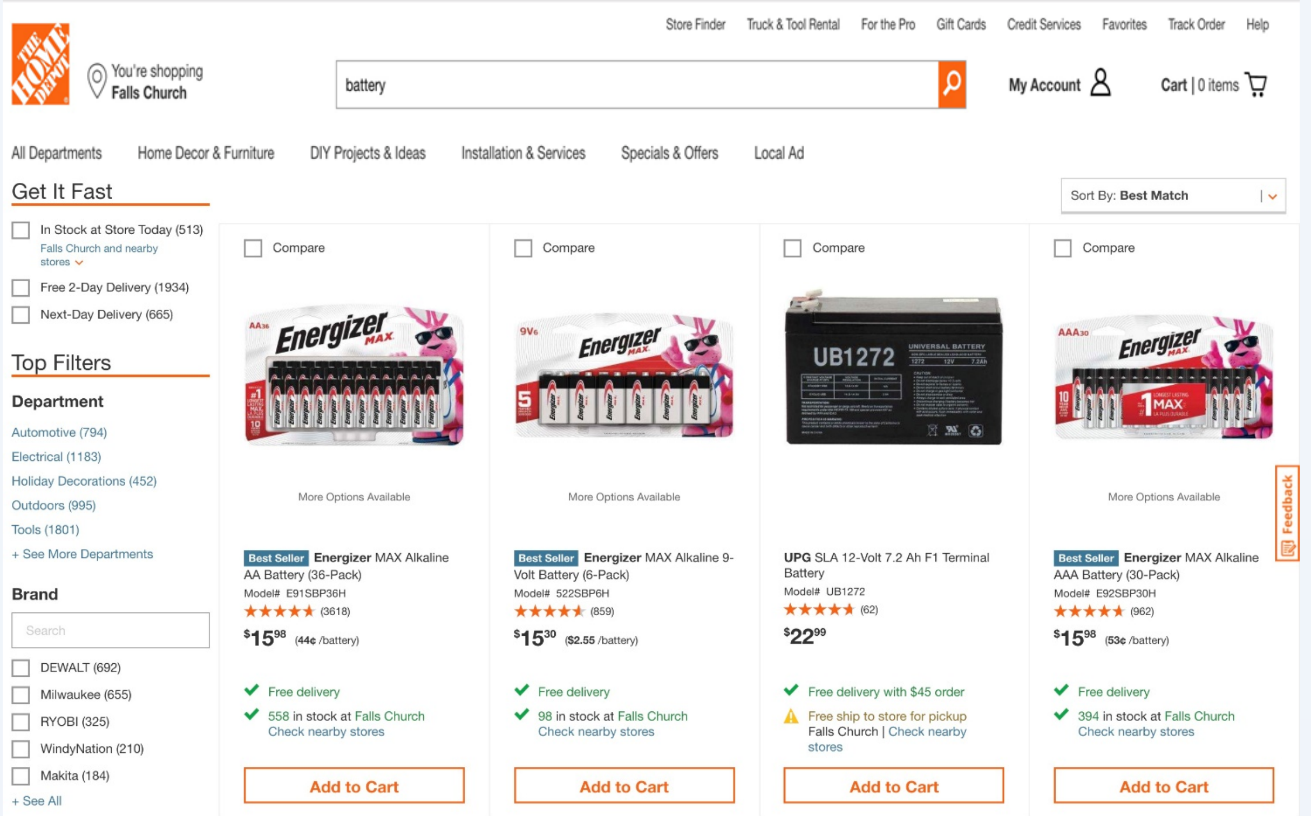
Home Depot Product Search Relevance

400 Bad Request - Hongyang Zheng, Heng Zhou, Zhengqian Xu

Georgetown University

INTRODUCTION



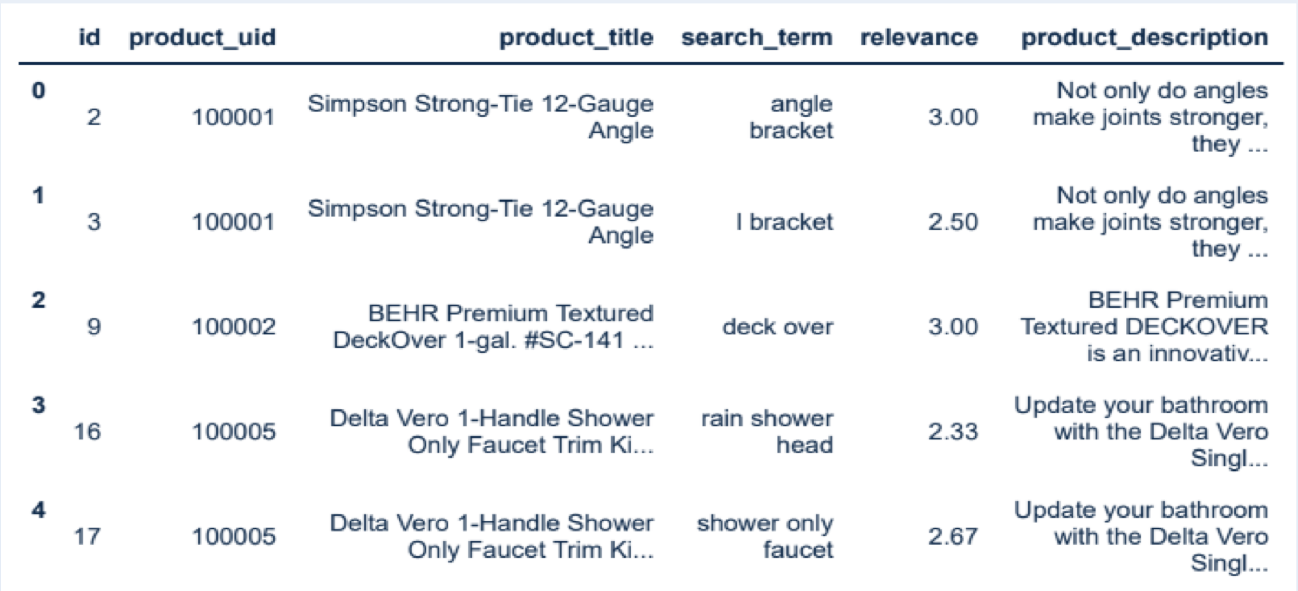
When a customer wants to buy something online, he/she might not be able to input the exact name for that product, and the company needs to figure out which products are most relevant to the search words in order to return the high relevant products to that customer.

Based on that scenario, our group first predicted the relevant score between what a user enters (keywords) and what the company has (products) by building supervised machine learning models with NLP techniques. And then on the top of our best model, we developed a mini search recommendation application to return the top 10 high relevant products in the database given a keyword.

DATA

Data and Features

- Train data: id, product_id, product_title, search_term, relevance
- Descriptions: product_description



	id	product_uid	product_title	search_term	relevance	product_description
0	2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	3.00	Not only do angles make joints stronger, they ...
1	3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	2.50	Not only do angles make joints stronger, they ...
2	9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 ...	deck over	3.00	BEHR Premium Textured DECKOVER is an innovativ...
3	16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	rain shower head	2.33	Update your bathroom with the Delta Vero Singl...
4	17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Ki...	shower only faucet	2.67	Update your bathroom with the Delta Vero Singl...

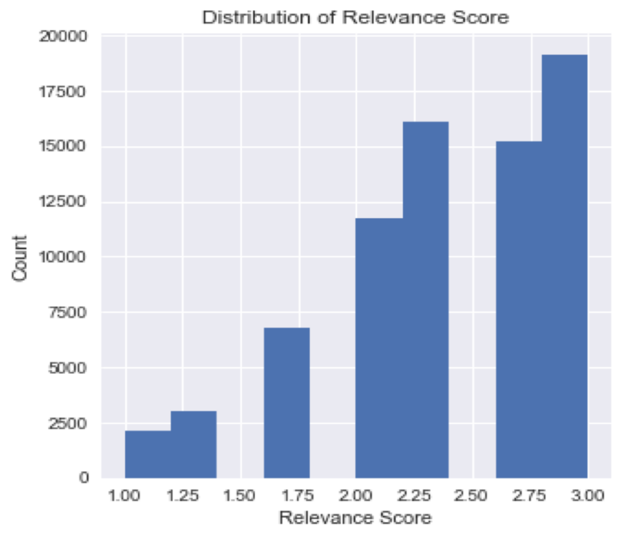
ANALYSIS

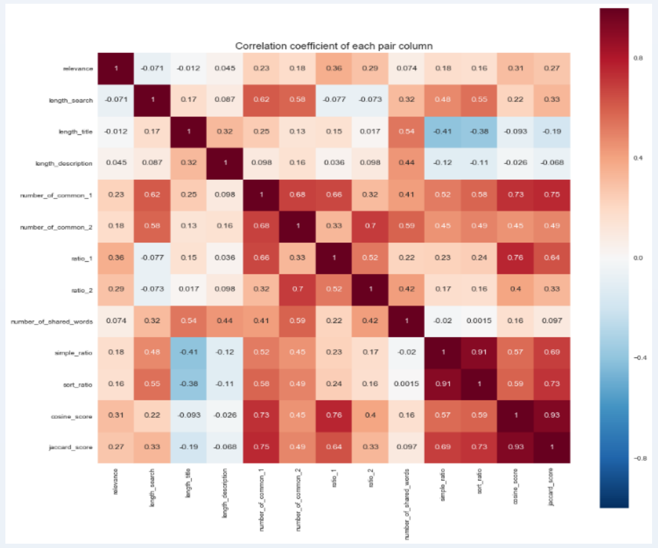
Feature Engineering

We create 12 numerical variables as following:

- **length_search**: length for `search term`
- **length_title**: length for `product title`
- **length_description**: length for `product description`
- **number_of_common_1**: the number of common words between token of `search term` and that of `product title`
- **number_of_common_2**: the number of common words between token of `search term` and that of `product description`
- **number_of_shared_words**: the number of shared words between token of `search term`, `product title` and `product description`
- **ratio_1**: number_of_common_1 / length_search
- **ratio_2**: number_of_common_2 / length_search
- **simple_ratio**: simple fuzzywuzzy similarity between `search term` and `product title`
- **sort_ratio**: sorted fuzzywuzzy similarity between `search term` and `product title` without considering the order
- **cosine_score**: cosine similarity between `search term` and `product title`
- **jaccard_score**: jaccard similarity between `search term` and `product title`

Exploratory Data Analysis



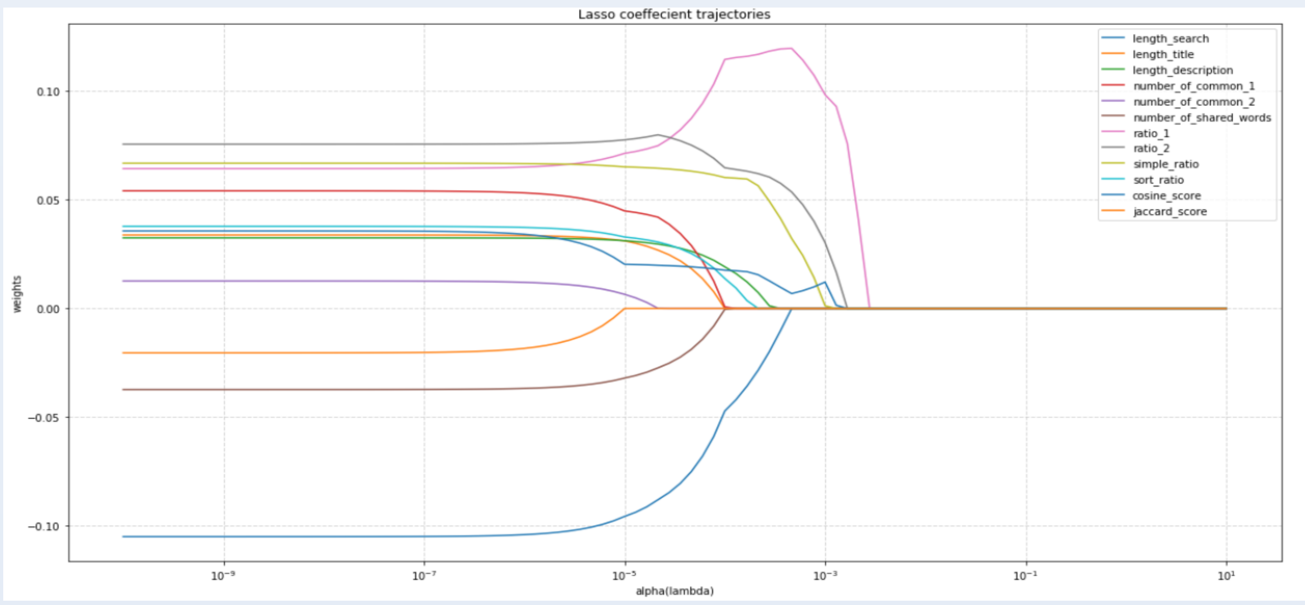


Model Building

We built four models by using 12 new created numerical variables to predict relevance score.

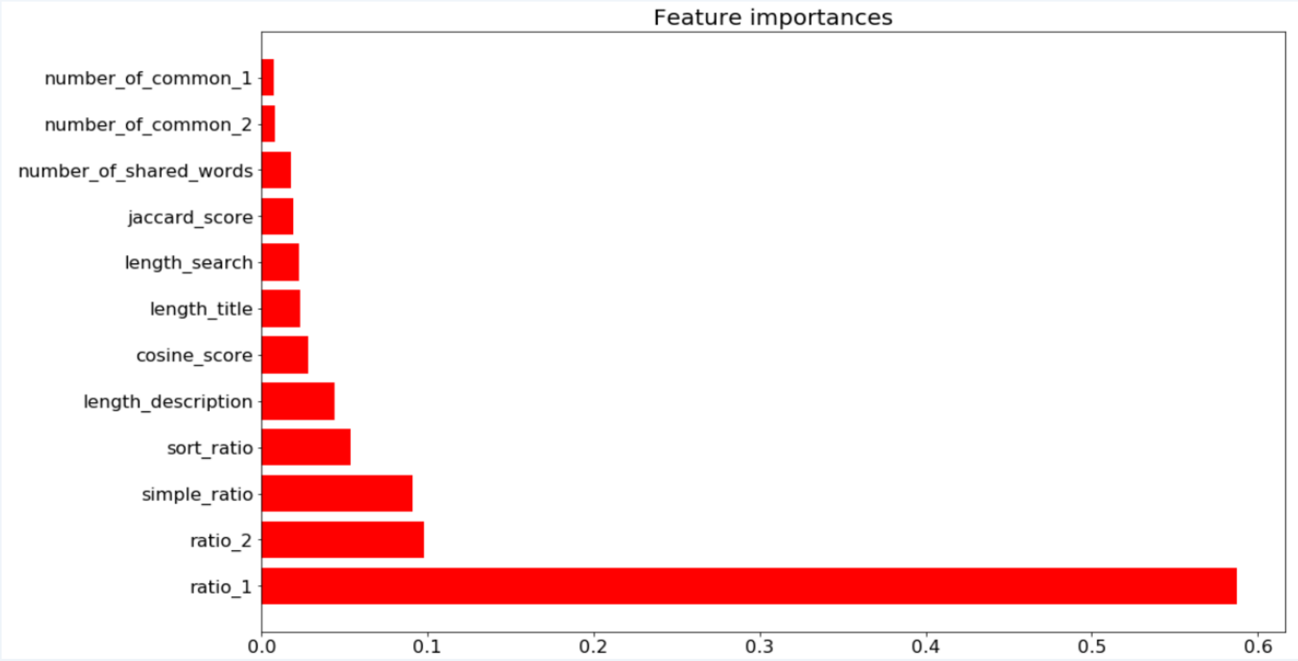
- Lasso:

It helps to make an automatic feature selection. We first made a trajectory plot of lasso coefficients and then used LassoCV to choose best parameter to fit the training data.



- Random Forest:

For this model, we used GridSearchCV to find the best parameters. We set max_depth as [5,6,7] and n_estimators as [200,400,600], then got the optimal parameters with max_depth at 7 and n_estimators at 400. Here is the feature selection visualization.



- Xgboost:

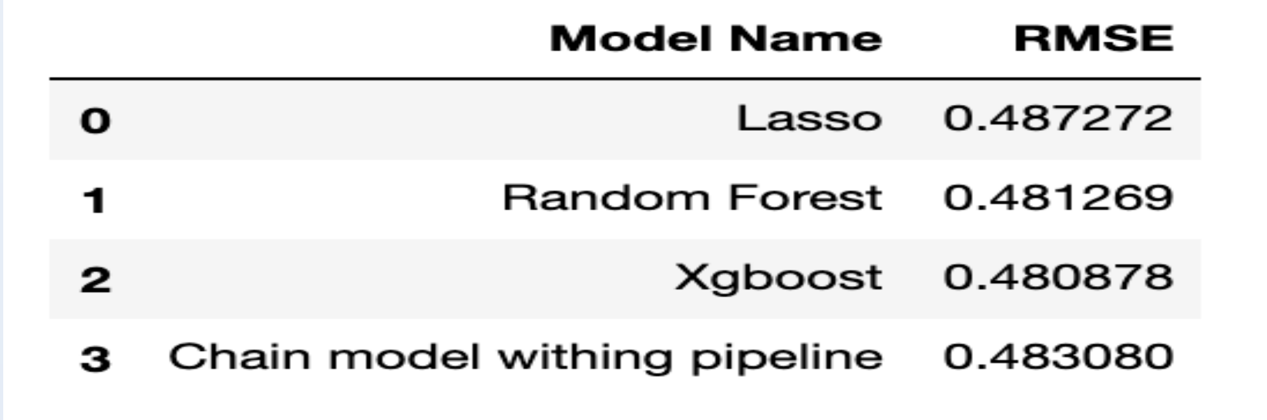
We also used GridSearchCV to find the best parameters. We set max_depth as [2,4,6,8] and n_estimators as [20,50,100,200]. Then got the optimal parameters with max_depth at 4 and n_estimators at 100.

- Chain model with pipeline:

We picked two best models - Random Forest and Xgboost, from the previous three models and built a pipeline to fit the training data.

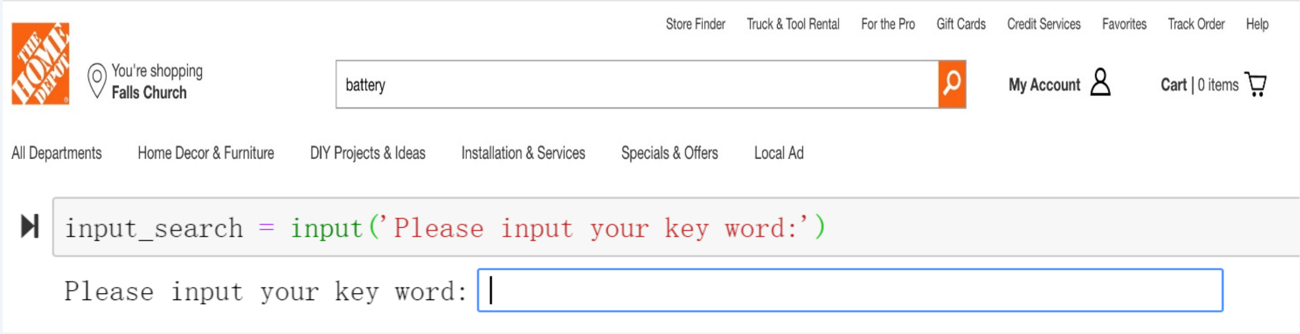
RESULTS

From the previous results, we can see that ratio_1 and ratio_2 are the top two important variables. Also, we use RMSE as our evaluation metrics and find our best model is Xgboost with test RMSE = 0.48. Compared with the best RMSE score 0.43 in Kaggle competition leaderboard, our result is pretty good since we trained our model based on a smaller dataset.



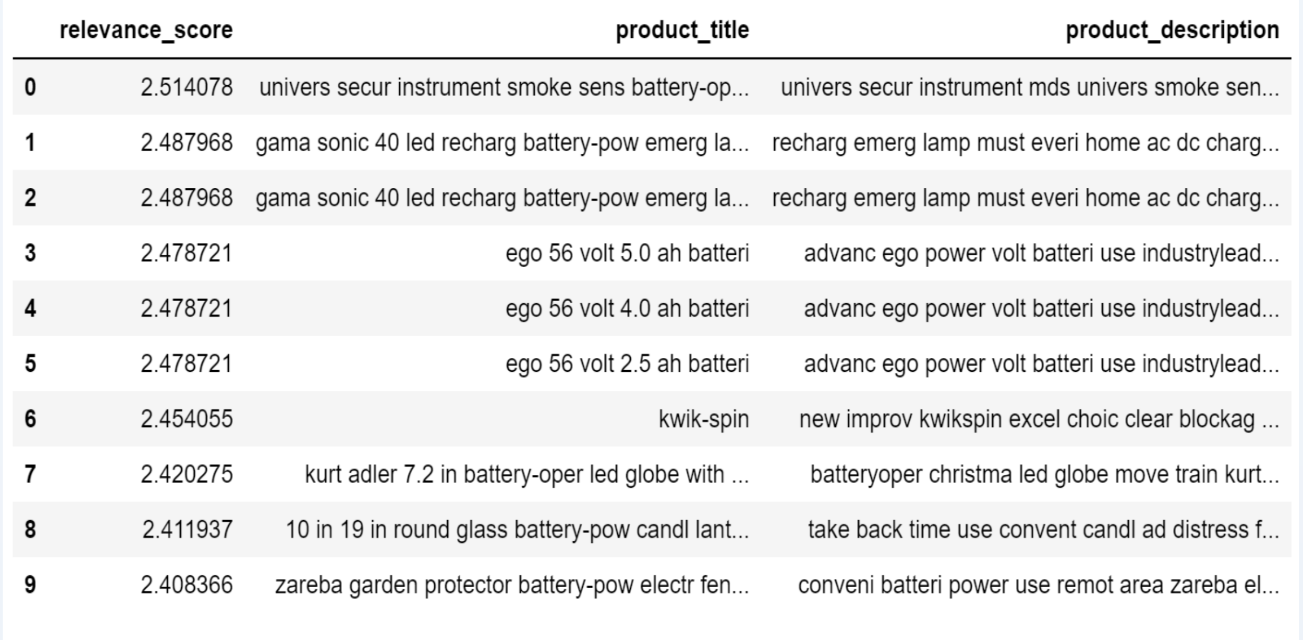
	Model Name	RMSE
0	Lasso	0.487272
1	Random Forest	0.481269
2	Xgboost	0.480878
3	Chain model withing pipeline	0.483080

APPLICATION



```
input_search = input('Please input your key word:')
Please input your key word:
```

- Ask the user to input a search term: battery
- Recalculate the new features
- Apply the best model Xgboost to the new data
- Get the products with the top 10 relevance score



	relevance_score	product_title	product_description
0	2.514078	univers secur instrument smoke sens battery-op...	univers secur instrument mds univers smoke sen...
1	2.487968	gama sonic 40 led recharg battery-pow emerg la...	recharg emerg lamp must everi home ac dc charg...
2	2.487968	gama sonic 40 led recharg battery-pow emerg la...	recharg emerg lamp must everi home ac dc charg...
3	2.478721	ego 56 volt 5.0 ah batteri	advanc ego power volt batteri use industrylead...
4	2.478721	ego 56 volt 4.0 ah batteri	advanc ego power volt batteri use industrylead...
5	2.478721	ego 56 volt 2.5 ah batteri	advanc ego power volt batteri use industrylead...
6	2.454055	kwik-spin	new improv kwikspin excel choic clear blockag ...
7	2.420275	kurt adler 7.2 in battery-oper led globe with ...	batteryoper christma led globe move train kurt...
8	2.411937	10 in 19 in round glass battery-pow candl lant...	take back time use convent candl ad distress f...
9	2.408366	zareba garden protector battery-pow electr fen...	conveni batteri power use remot area zareba el...

CONCLUSION

Our project starts from feature engineering, model building, and finally developed a mini search recommendation application. Most of the products are related to the keyword, but there are still some unrelated products. However, since we only used a subset from the original data, this result is better than what we expected at the beginning.

REFERENCES

[1] Petar R. Petar P., Peter M., Heiko P.: A Machine Learning Approach for Product Matching and Categorization(2016)

[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)