# Kaggle - Home Depot Product Search Relevance

1. Описание задачи
2. Очистка исходных данных
3. Генерация признаков
4. Отбор моделей
5. Поиск оптимальных параметров
6. Построение ансамбля

# Описание задачи

train.csv

| uid | product title | search query | relevance |
|---|---|---|---|
| 100010 | Valley View Industries Metal Stakes (4-Pack) | steele stake | 2.67 |
| 100119 | Purdy 2 in. A. P. All Paints Brush Set (3-Pack) | paint roller inserts | 1 |

description.csv

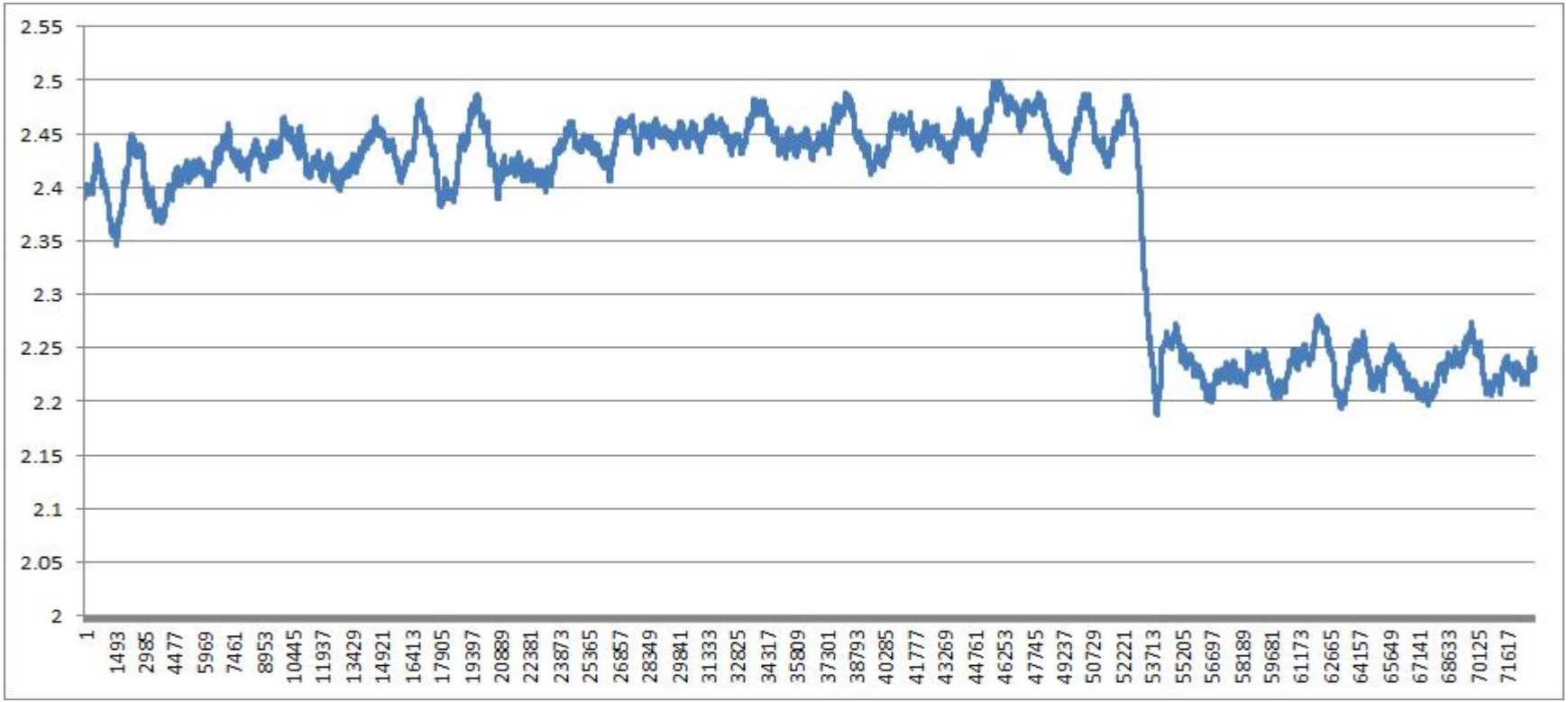
| |
|---|
| Valley View Industries Metal Stakes (4-Pack) are 9 in. galvanized steel stakes for use with all Valley View lawn edgings and brick and paver edgings. These utility stakes can also be used for many other purposes. It is recommended that anchor stakes are used every five feet on designs that have the edging in straight lengths. Where there are curved designs for edgings, additional anchor stakes are recommended at the curve points. |

attributes.csv

| | |
|---|---|
| Color Family | Metallic |
| Color/Finish | Silver/Gray |
| Material | Steel |
| MFG Brand Name | Valley View Industries |
| Package Quantity | 4 |
| Product Depth (in.) | 1 |
| Product Height (in.) | 9 |
| Product Width (in.) | 1 |
| Width | 1 in |

# Описание задачи

MA(relevance, 1000)



## Очистка исходных данных

- Удаление не значащих символов
- Удаление стоп-слов
- Коррекция ошибок
- Поиск синонимов
- Процесс нахождения основы слова
  - stemming
  - lemmatization

## Генерация признаков

1. Кол-во символов и кол-во слов в объектах
2. Кол-во прямых совпадений слов
3. Мера схожести слов (Levenshtein Distance)
4. Матрица схожести
5. Анализ структуры предложения
6. TF-IDF
7. Word2Vec

# Генерация признаков - Матрица схожести

Valley View Industries Metal Stakes (4-Pack) steele stake

| | Industries | Metal | Stakes | (4-Pack) |
|---|---|---|---|---|
| steele | 0.1 | 0.7 | 0.1 | 0.1 |
| stake | 0.1 | 0.1 | 1 | 0.1 |

# Генерация признаков - Анализ структуры предложения

“I shot an elephant in my sleep”

| | | | |
|---|---|---|---|
| I | PRP | 2 | nsubj |
| shot | VBD | 0 | root |
| an | DT | 4 | det |
| elephant | NN | 2 | dobj |
| in | IN | 7 | case |
| my | PRPs | 7 | nmod |
| sleep | NN | 2 | nmod |