Home Depot Product Search Relevance

•••

JaJan Hsu
Leilei Liu
Zhaoyan Ai

Content

Data Analysis Process

- Clean-up data
- Stem data
- Feature selection

Result

- What method worked and what didn’t work
- Scores

Future Improvement

- Future Work

Clean-up and Stem data

- Converted all the characters to lowercase

  Python built-in method

- Remove or replace special characters

  ", “, "$", "Ã¥Â¡", "+",";" "?", "-", "#" ,"(" ,")"

  " sq." to " 1sq ", " sq " to " 1sq ", " v. " to " 1volt "., " cm " to " 1cm "

  convert edall number words to integers - one to 1

- Spelling check

  JAVA jSpellCorrect library

  Scripts thread “Fixing Typos” - spelling check dictionary

- Stem data

  Nltk SnowballStemmer

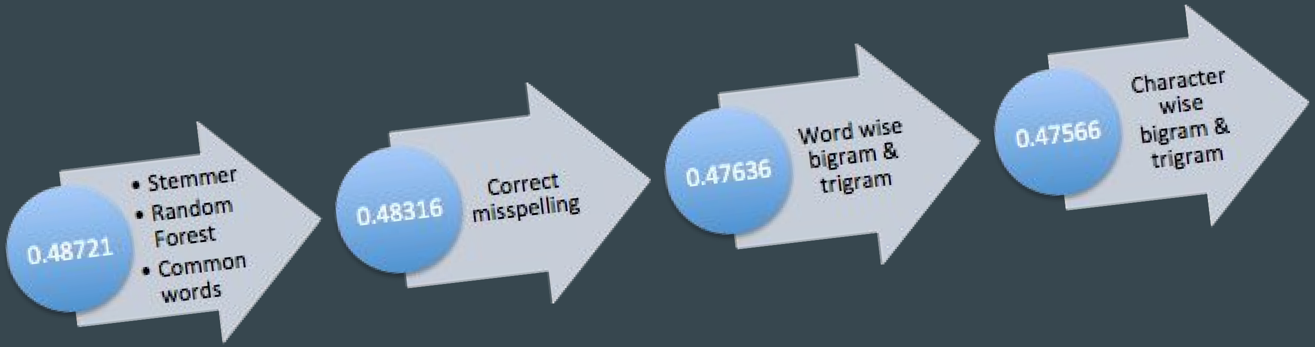
Feature Engineering

- Count how many words in search item also appear in product title and how many words in search item also appear in product description
    - Word in title & word in description
- Ngrams - consecutive words may bear semantic meanings therefore may contribute to the model
    - Word-wise bigrams & trigrams
        - “Hello world this is Python” => ('Hello', 'World'), ('World', 'this'), ('this', 'is'), ('is', 'Python')
        - “Hello world this is Python” => ('Hello', 'World', 'this'), ('World', 'this', 'is'), ('this', 'is', 'Python')
    - Character-wise bigrams & trigrams
        - “Hello world this is Python” => ('H', 'e'), ('e', 'l'), ('l', 'l'), ('l', 'o'), ('o', ' '), (' ', 'W'), ('W', 'o') ...
        - “Hello world this is Python” => ('H', 'e', 'l'), ('e', 'l', 'l'), ('l', 'l', 'o'), ('l', 'o', ' '), ('o', ' ', 'W'), (' ', 'W', 'o'), ('W', 'o', 'r'), ('o', 'r', 'l'), ('r', 'l', 'd'), ('l', 'd', ' '), ('d', ' ', 't') ...

Progress Roadmap

- Worked
  - Data cleaning
    - Stemming
    - Typo correction using JAVA jSpellingCheck
    - Typo correction using Google
  - Ngram
    - Bigram and trigram
    - Word - wise and character - wise
- More or less worked
  - Tweaking the parameter of random forest and bagging regressor
    - Number of estimator
    - Max depth
- Not worked
  - Extracting brand information from attributes raw data and investigate relationship between search item and brand information
  - Stop word

Scores



0.48721

- Stemmer
- Random Forest
- Common words

0.48316

Correct misspelling

0.47636

Word wise bigram & trigram

0.47566

Character wise bigram & trigram

Future Work

- Normalizing data
- Feature transformation
- Using attributes

Thank You!