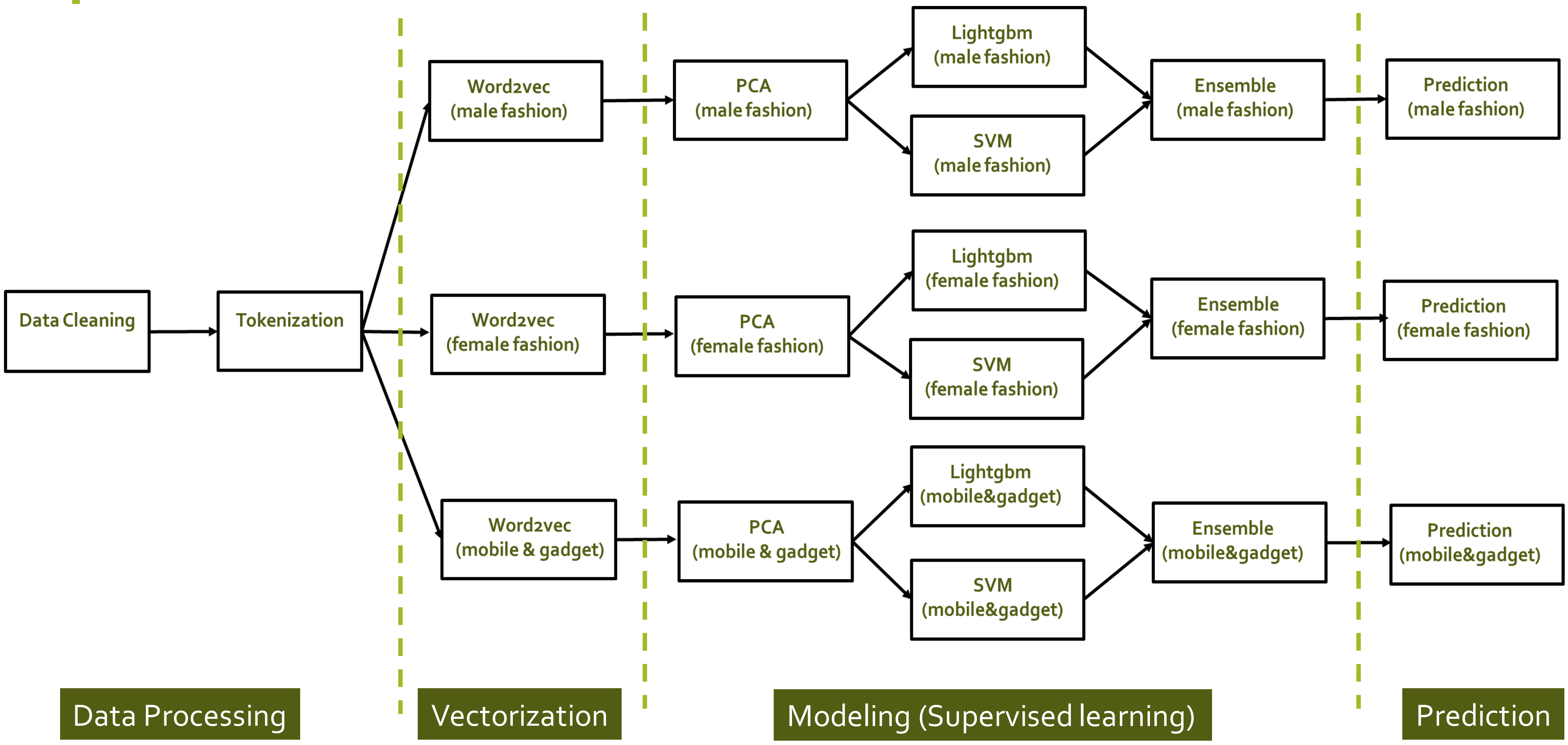
PRODUCT KEYWORDS EXTRACTION

LI ZHIJUN

Pipeline


Data Cleaning
Tokenization
Word2vec (male fashion)
Word2vec (female fashion)
Word2vec (mobile & gadget)
PCA (male fashion)
PCA (female fashion)
PCA (mobile & gadget)
Lightgbm (male fashion)
SVM (male fashion)
Lightgbm (female fashion)
SVM (female fashion)
Lightgbm (mobile&gadget)
SVM (mobile&gadget)
Ensemble (male fashion)
Ensemble (female fashion)
Ensemble (mobile&gadget)
Prediction (male fashion)
Prediction (female fashion)
Prediction (mobile&gadget)
Data Processing
Vectorization
Modeling (Supervised learning)
Prediction

Data Processing

1. Convert English characters to lower case, and Chinese characters to simplified version
2. Remove emoji, punctuations, stopwords
3. Tokenization

   - Tool used: jieba (python)

   - new word identification: HMM model

   - Performance improvement: customized dictionary

Example

Input	🙈現貨👩♡👩創意新款 可愛文藝 大理石充電寶10000毫安 花紋 行動電源 超薄簡約卡通移動電源
Output	现货 创意 新款 可爱 文艺 大理石 充电宝 10000毫安 花纹 行动电源 超薄 简约 卡通 移动电源
Input	OPPO R9/R9S/R9plus/R9S-plus/R11日韓潮牌鉚釘卡通兔子手機殼/手繩全包軟殼防摔（2色）預購！
Output	oppo r9s r9plus r9splus r11 日韩 潮牌 铆钉 卡通 兔子 手机壳 手绳 全包 软壳 防摔 2色 预购

Vectorization

- **Word embedding**

  1) Algorithm: Word2vec (skip-gram)

  2)Train three word2vec model by three categories

  (reason: product titles from different categories construct different corpus)

  3) Output: Each word would be represent as an vector with size of (100,1)

- **String vectorization**

  1) Statistical summarization of each dimension in word vector: min, max, mean, skewness, kurkosis

  2) Each string (either product title or query) would be translated into a vector with size (500, 1)

Example

Input

```
>> print(model_mf['沙滩裤]
```

Output

```
array([-0.19924697,  0.6276916 ,...
     ..., 0.16667709,  0.01390105],
   dtype=float32)
```

Input

```
>> print(model_ff.similarity('短裤','长裤'))
>> print(model_ff.similarity('短裤','t恤'))
```

Output

```
0.860566388725l421
0.7665001355901444
```

Input

```
>> model_ff.most_similar(['t恤'])
```

Output

```
[('素t', 0.9832006692886353),
 ('短袖上衣', 0.9797725677490234),
 ('圆领', 0.97728562355O415),
 ('短袖t恤', 0.9683908224105835),...]
```

...

Modeling

- **Objective**: Build a predictive model to better capture effective (product title, keyword) pair.
- **Method used**: Binary Classification by using supervised learning. Users' log data could be a kind of natural label indicating whether a keyword is effective or not. Therefore, supervised learning, which always outperforms unsupervised learning, is chosen.
- **Binary Target Variable Construction**: The same (product title, query) pair may have both 'click' or 'impression' records. Here, target variable 'is_effective' is defined as a binary variable, which has value of 1 if the number of click is greater than 0, otherwise 0.

Product Name	Query	Event	Date
Name1	keyword1	Click	30/7/17
Name1	keyword1	Impression	31/7/17
Name1	keyword2	Impression	31/7/17

Product Name	Query	Is_effective
Name1	keyword1	1
Name1	keyword2	0

- **Dimension reduction**: PCA (from 1014 features → 214 features), reducing the complexity of the model
- **Algorithms**: lightgbm + svm (ensemble), ensure the robustness of the model

Feature engineering

14 new features are created based on an assumption that position could indicate the name entity of a word to some extent (for example, seller's names always occur at the beginning of the title, followed by adjective and product category)

Store name	Adjective	Product name
(A05)棉花糖女孩 ❤	M-3XL大碼大尺碼M-3XL大尺碼夏春韓國歐美大碼	洋裝連身裙長版T短袖T恤顯瘦中長款短袖

Feature Name	Description
Product_length	Length of product title
Query_length	Length of query
Min_pos	Position of the first occurrence of the query in the title.
Min_pos_por	Proportion of the first occurrence of the query in the title. (min_Pos/ length(title))
Max_pos	Position of the last occurrence of the query in the title
Max_pos_por	Proportion of first occurrence of the query in the title. (max_Pos/ length(title))
Mean_pos	Average position of the occurrence of the query in the title
Mean_pos_por	Proportion of average position of the query in the title. (mean_Pos/ length(title))
T_min_pos_por	Among all occurrence of a query in all product titles, what is the minimum proportion of position
Tmax_pos_por	Among all occurrence of a query in all product titles, what is the maximum proportion of position
Tmean_pos_por	Among all occurrence of a query in all product titles, what is the average proportion of position
tf	Term frequency (number of occurrence of the query in the title)
idf	Inverse document frequency
tfidf	Tf * idf

Model measurement

- To maximize the ROI of sellers, **precision** is important
- To maximize the revenue of ecommerce platform, **recall rate** is important
- F1, recall rate, and precision are both chosen as a measurement of model

	Male Fashion	Female Fashion	Mobile & gadget
F1	0.409	0.366	0.379
Recall	**0.915**	**0.908**	**0.929**
Precision	0.263	0.229	0.238

Prediction

Step 1: Tokenization

夏季情侶款三色拼接OVERSIZE寬鬆落肩短袖T恤

夏季, 情侶款, 三色, 拼接, oversize, 寬鬆, 落肩, 短袖t恤

Step 2: Construct (product tile, query) pair

Step 3: Predict the probability of click for each (title, query) pair

Product Name	Query	Is_effective
夏季情侶款三色拼接OVERSIZE寬鬆落肩短袖T恤	短袖t恤	0.8
夏季情侶款三色拼接OVERSIZE寬鬆落肩短袖T恤	寬鬆	0.7
夏季情侶款三色拼接OVERSIZE寬鬆落肩短袖T恤	落肩	0.6

Step4: Give a list of recommended words ordered by predicted probability

Product Name	Recommend words
夏季情侶款三色拼接OVERSIZE寬鬆落肩短袖T恤	[短袖t恤,寬鬆,落肩, …]

Output: A list of recommended words ordered by predicted probability, for sellers' reference, to allow them pick either single keyword or combination of keywords they want to buy.