FDA BI Research Weekly Report

Yage Wang, Zheng Gong 20170830

1. **Integrated BI text mining method report**

**Step 1: Extract data from data source**

The current dataset comes from MAUDE database and all records whose product code equals to “FWM” or “FTR” are selected out from table “Foidev” as the initial dataset.

Then we go through table “Foitext” and “Mdrfoi” to select all records who matches the initial dataset on [MDR\_REPORT\_KEY] column, and add these records to our dataset.

We now have 3 tables in our dataset, each having around 27,000 records using [MDR\_REPORT\_KEY] column as the primary identifier.

Currently, we focus on following columns in each table and we will use Table\_Name[Column\_Name] to stand for the column in following content. E.g Foidev[BRAND\_NAME] stands for BRAND\_NAME column in Foidev table.

|  |  |
| --- | --- |
| Table | Column |
| Mdrfoi | - |
| Foidev | MANUFACTURER\_D\_NAME, BRAND\_NAME, GENERIC\_NAME |
| Foitext | FOI\_TEXT |

Note: all columns above is text type. Our method will mainly focus on natural language processing techniques.

**Step 2: Data preprocessing and cleaning**

1. Convert all text to lower/upper case

All contents in MAUDE database are already upper case, so we skip this sub-step.

1. Remove stop words

Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words.

The default stop words list of English language has 153 words.

1. Remove punctuations

Punctuation list: '''!()[]{};:'"\,<>./?@#$%^&\*\_~'''

Note: considering that we will search biomarkers such as “CD30+”, so we don’t remove the punctuation “+” and “-“

**Step 3: Keywords search**

1. Keywords preprocessing - stemming

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form.

Considering we only need to select the keyword out from plain text, it’s efficient to stem only the keywords rather than the whole text.

* 1. For single word, we stem the word to it root, e.g.

"stemmer", "stemming", "stemmed" -> "stem"

"fishing", "fished", "fisher" -> "fish"

"argue", "argued", "argues", "arguing", "argus" ->"argu"

* 1. For phrase, we stem the last word of the phrase, e.g.

“New Yorkers” -> “New York”

1. Standardize manufacturer name

We use a manufacturer name dictionary offered by FDA to standardize Foidev[MANUFACTURER\_D\_NAME], e.g.

“PIP INC”, “PIP USA”, “PIP/USA INC” -> “Poly Implant Prothese”

1. String matching method
   1. For single word

We use space as a delimiter to split the preprocessed text data into a word list, then match stemmed keyword to each word in the word list.

Example:

1. Target: to find if the sentence contains the keyword “apple”.
2. Sentence: I eat three sweet apples today.
3. Stemmed keyword: “appl”
4. Split sentence: [“I”, “eat”, “three”, “sweet”, “apples”, “today”]
5. Match “appl” to each word in split sentence.
6. String “appl” is in string “apples”, consider the sentence contains keyword “apple”
   1. For phrase

We use the whole text from a record as string, then match stemmed phrase to the whole text.

Example:

1. Target: to find if the sentence contains phrase “breast pain”.
2. Sentence: I got serious breast pain yesterday.
3. Stemmed phrase: “breast pain”
4. Sentence string: “I got serious breast pain yesterday”
5. Match “breast pain” to sentence string
6. “breast pain” is in string “I got serious breast pain yesterday”, consider the sentence contains phrase
7. Searching field

The keyword list is set up from different aspect, for each sub keyword list, we search different columns.

Search field of each sub-keyword list:

|  |  |  |
| --- | --- | --- |
| Sub-keyword list | Column of Foidev table | Column of Foitext table |
| Manufacture | MANUFACTURER\_D\_NAME | - |
| fill\_type | BRAND\_NAME, GENERIC\_NAME | FOI\_TEXT |
| surface\_type | BRAND\_NAME, GENERIC\_NAME | FOI\_TEXT |
| ALCL | - | FOI\_TEXT |
| side | - | FOI\_TEXT |
| biomarker | - | FOI\_TEXT |
| symptom | - | FOI\_TEXT |

1. Convert text to feature vector

For each keyword, if a record contains the keyword, it will be labeled as “1” in the related keyword column. For other keyword that do not appear in the record, label the record as “0” in the related column.

Example: Original records

|  |  |
| --- | --- |
| MDR\_REPORT\_KEY | FOI\_TEXT |
| 24568502 | WE HAVE A CONFIRMED CASE OF BI-ALCL IN A PT THAT RECEIVED BREAST RECONSTRUCTION AFTER SUBCUTANEOUS MASTECTOMY 2011 VIA SILICONE GEL IMPLANTS. IN THE CYTOLOGY OF PERIPROSTHETIC FLUID IN THE IMPLANT CAPSULE WE COULD IDENTIFY CD 30. |

After word-to-vector transformation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MDR\_REPORT\_KEY | Fill\_type | | | implantation\_indication | symptom | ALCL |
| SALINE | SILICONE | GEL | RECONSTRUCTION | FLUID | ALCL |
| 24568502 | 0 | 1 | 1 | 1 | 1 | 1 |

1. Multiple value solution
   1. Multiple same value

For a specific keyword, no matter which column we detect it, we will label the record as “1” in related keyword column.

Example:

We detect the keyword “SALINE” in both [BRAND\_NAME] and [GENERIC\_NAME] column, we will label the record as “1” in column “SALINE”

* 1. Conflicted value

Currently we define three pair of conflicted keywords:

* “LEFT” – “RIGHT” in side sub-list
* “SMOOTH” – “TEXTURED” in surface\_type sub-list
* “SALINE” – “SILICONE” in fill\_type sub-list

If we detect the conflicted keywords in the same record, we will label the recorded as “conflicted” for further processing.

**Step 4: Training classification model**

We need to answer the questions:

In what condition a record will be identified as ALCL case?

1. **Stop words sample**

Please see file [1\_Englist\_Stopwords\_List.txt]

1. **Generate new QC data and keyword frequency**

Please see file [2\_Quality\_Control\_Data\_0830.xlsx] and [3\_Keywords\_Word\_Frequency\_0830.xlsx]

1. **Conflicted record sample**

Please see file [4\_Conflicted\_Records.xlsx]

1. **Word cloud generator website**

Please see file [5\_Free\_Online\_Word\_Cloud\_Generator.txt]