From poor data
To training data
...

JeongMin Kwon

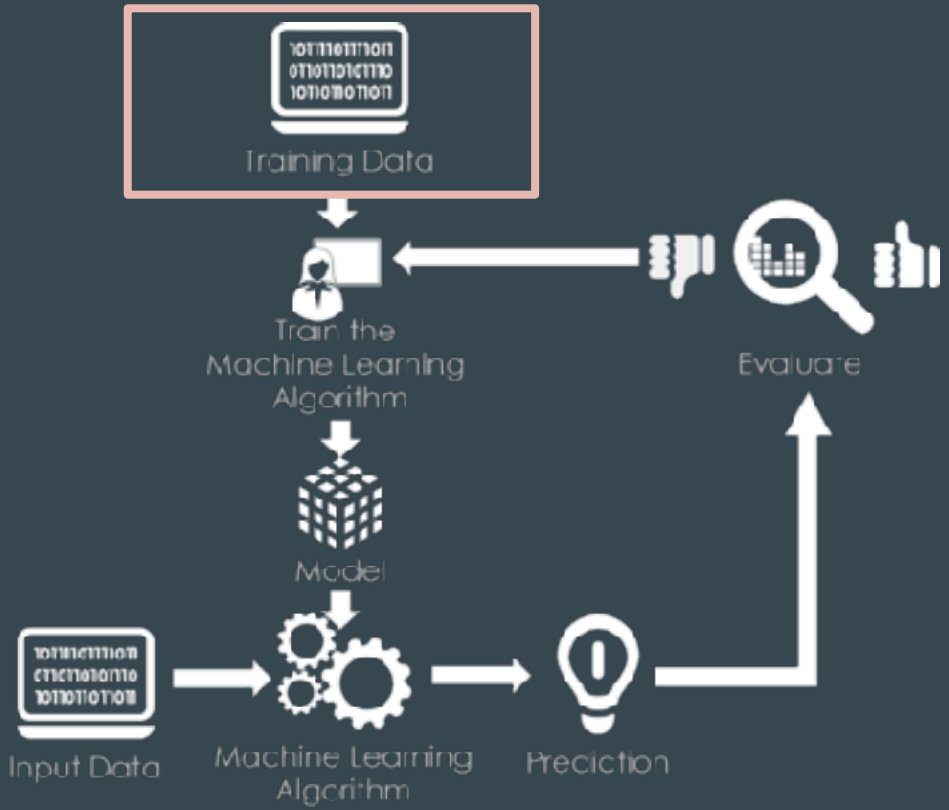
JeongMin Kwon

Job Title:
Data Scientist

Interests:
Data Problem Solving, R, Python,
translation,
and many more.

Contact:
cojette@gmail.com
/ @mazycat

Machine learning services with Data



Machine Learning Dataset Condition

- Clean
- Consistent
- Simple type
- Clear predictors and dependant variables
- Properly scaled and discretized

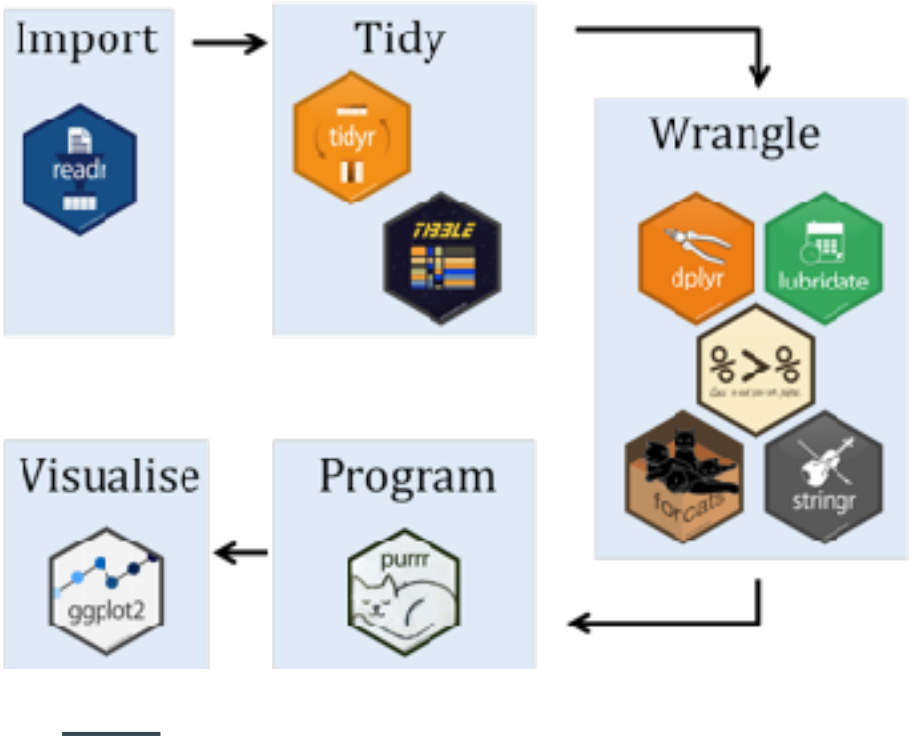
There are lots of poor data...

Tidy datasets are all alike
but every messy dataset is messy in its own way

- Raw and messy file (Raw mine for data)
- Data from various different sources and repositories
- Too large datasets
- Too various information or columns
- Different categories and criteria
- Meaningless text and unstructured data
- Lots of errata, omitted fields, various formats and missing values
- etc.

Tidyverse

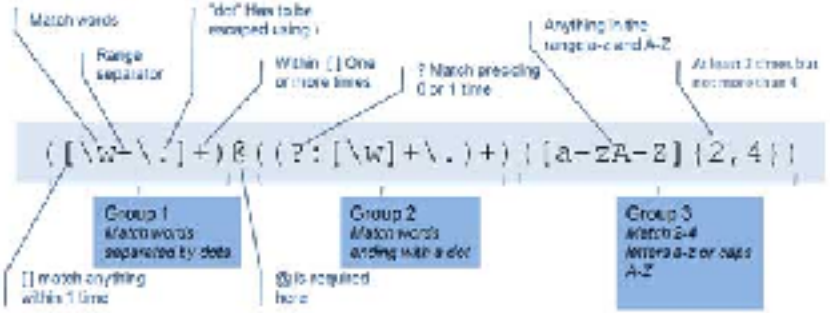
Basic Data Wrangling



Regular Expression

Basic Data Wrangling

- Text Pattern control
- Text Cleansing
- Default: extended regex
  - PCRE: perl=TRUE option



Outside of the data box

Change the problems

- Regression to Classification (vice versa)
- Numeric variables to Characteristic variables (vice versa)
- Word data(categorical) to Document(Text)
- And many more...

Purchase history data

But...Data?

- Fried Chicken (치킨)
  - 후라이드치킨
  - 크리스피치킨
  - 후라이드 치킨
  - 마일드치킨(Mild)
  - 오리지널치킨
  - 닭강정⭐
  - 써프라이드 - 매운맛
  - 소이갈릭스
  - 1. 치즐링
  - A. 순살후라이드
  - 마라치킨(麻辣)
  - ….

Restaurants and Users

Restaurants

- Each restaurant sells various dishes

Users

- Every person eats various food

Recommendation with TF-IDF

Word Cleansing

- Food name cleansing with regular expressions and rules

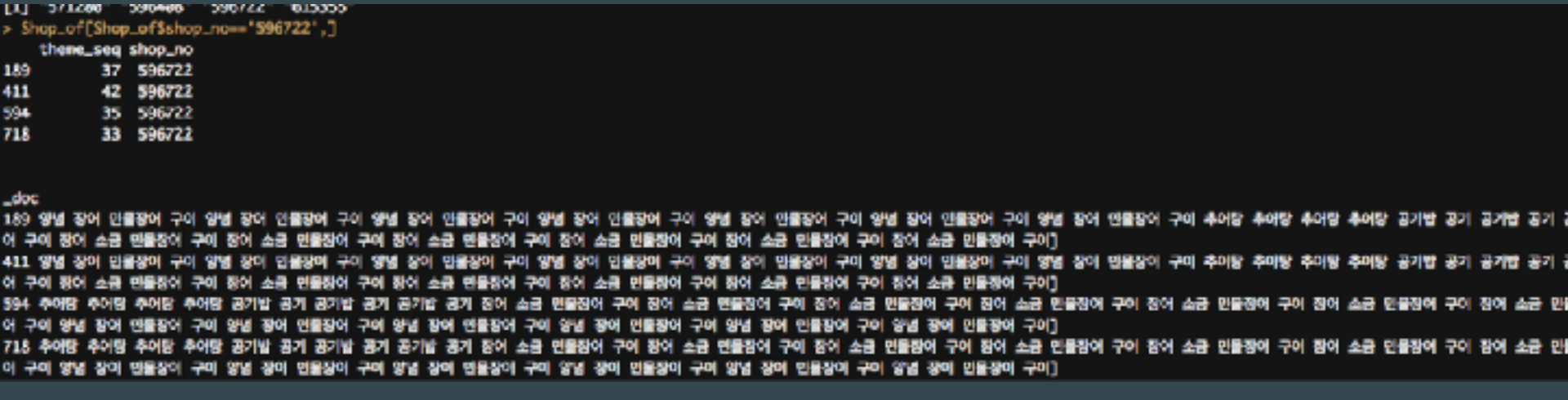
```
if(sum(grepl( "(^|[:space:]|[ｱ-ﾜ]|\\(|\\))f($|[:space:]|[:punct:]|\\))|(^|[:space:]|[ｱ-ﾜ]|\\(|\\))F($|[:space:]|[:punct:]|\\))", ret.menu))>0) {
    ret.menu <- str_replace_all(ret.menu, "(f|F)($|[:punct:]|[:space:])",  " ﾌﾗｲ ")
    if (flg == TRUE)
        ii <- ii + 1}
if(sum(grepl( "(^|[:space:]|[ｱ-ﾜ]|\\(|\\))p($|[:space:]|[:punct:]|\\))|(^|[:space:]|[ｱ-ﾜ]|\\(|\\))P($|[:space:]|[:punct:]|\\))", ret.menu))>0) {
    ret.menu <- str_replace_all(ret.menu, "(p|P)($|[:punct:]|[:space:])",   " ｽﾁｰﾑ ")
    if (flg == TRUE)
        ii <- ii + 1
}
```

```
if(sum(grepl("^A?([0-9]|[1-9][0-9]|[1-3][0-3][0-9]|[:alpha:])([:punct:]|\\-|\\+|\\-|/^)", ret.menu))>-1)
{
    ret.menu <- str_replace_all(ret.menu, "^A?([0-9]|[1-9][0-9]|[1-9][0-9][0-9]|[:alpha:])([:punct:])", "")
    print("22")
    if (flg == TRUE)
        ii <- ii + 1
}
```

Purchase data to Documents

Restaurant Documents

- Users and restaurants(Item) Documents with cleansed food names



Text Processing

Tidytext

- Text mining for word processing and sentiment analysis using 'dplyr', 'ggplot2', and other tidy tools.

Example

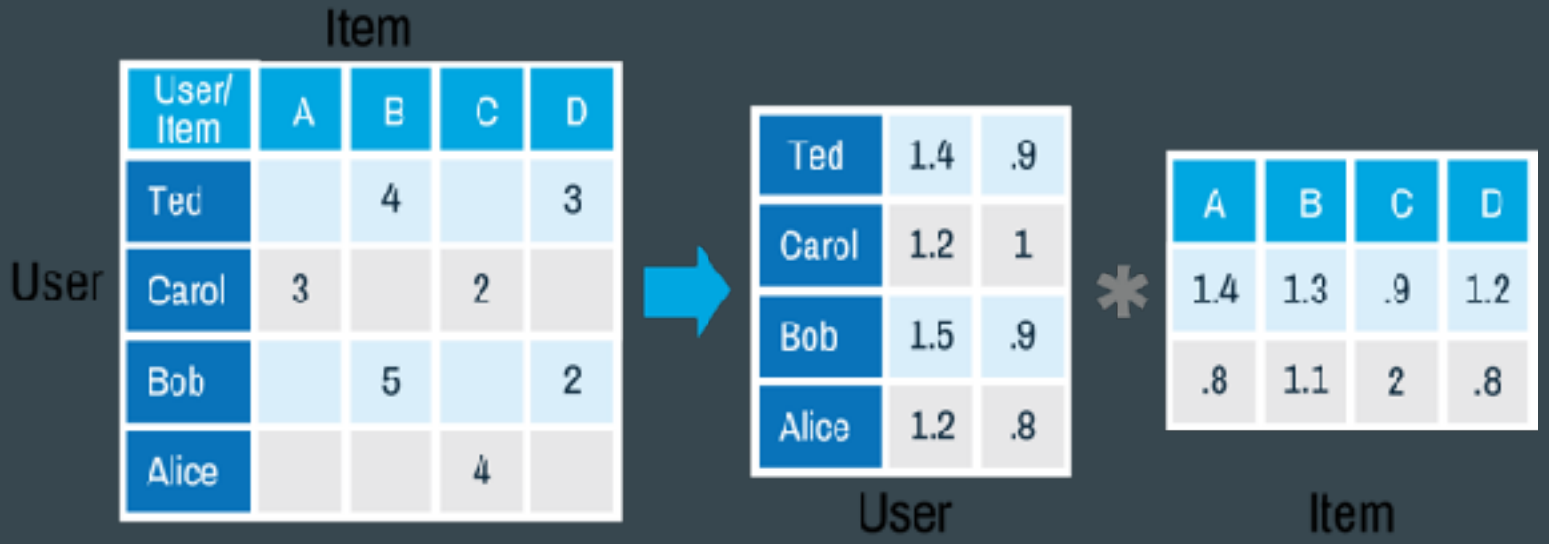
```
library(tidyverse)

shopTF <- shopDoc%>%unnest_tokens(word, shop_food_doc)
                   %>%count(theme_seq, shop_no, word)
                   %>%ungroup()
shopTF1 <  cbind(paste(shopTF$theme_seq, shopTF$shop_no),
                 shopTF[,c(3,4)])
colnames(shopTF1) <- c("theme_shop", "word", "n")
```

Recommendation with TF-IDF

ALS algorithms

- Users and restaurants(Item) matching



Summary

Machine Learning Data

There are far fewer clean data for machine learning, but we can made with data wrangling

Data Wrangling

Tidyverse + Regular Expression + Out of the box thinking

The Era of **Data Wrangling**
is coming….

Q & A

Thank you for listening

Contact: cojette@gmail.com

@mazycat