Case 1 - Obesity NLP Challenge

2021/10/20

林亦盛309551074 / 周君諦310551136 / 陳昱銘310554007

Outline

- Dataset Parsing
- Preprocessing
  - Text Recontruction
  - Remove irrelevant section
  - NLTK
  - TF-IDF
- Model
  - Random forest
  - NN
- Results
- Reference
- Team Member Contribution

Dataset Parsing

- Train (textual labels): 400 records, including 200 U and 200 Y
- Test (intuitive labels): 400 records, including 200 N and 200 Y
- Overlap Records:
  - Total records: 465 (335 overlap records)
  - Records of consistent label:
    - U -> N : 158
    - Y -> Y : 140
  - Records of inconsistent label (37 records)
    - U -> Y : 13
      - ID_716, ID_725, ID_728, ID_737, ID_740, ID_747,
      - ID_851, ID_855, ID_861, ID_869, ID_882, ID_884, ID_891
    - Y -> N : 24
      - ID_715, ID_726, ID_734, ID_739, ID_746, ID_750,
      - ID_854, ID_857, ID_868, ID_873, ID_876, ID_883, ID_890, ID_892, ID_897,
      - ID_904, ID_909, ID_915, ID_921, ID_929, ID_932, ID_935, ID_943, ID_945

Dataset Parsing

		Textual			
		Y	**U**	**Not included**	**Sum**
Intuitive	**Y**	140	13	47	200
	N	24	158	18	200
	Not included	36	29	--	65
	Sum	200	200	65	

Preprocessing - Reconstruction

```
PHYSICAL EXAMINATION: On admission , the patient was an
uncomfortable obese white female in moderate
distress. Vital signs were stable and her temperature was 98.6.
SKIN: Normal. HEENT: Normocephalic and atraumatic. CHEST:
Clear bilaterally. CARDIAC: Regular rate and rhythm with normal
S1 and S2 with a grade II/VI systolic ejection murmur. ABDOMEN:
Soft , non-tender , and non-distended except for in the right upper
quadrant where she had right upper quadrant pain and a positive
Murphy's but no peritoneal signs.
LABORATORY EXAMINATION: Significant for
hematocrit of 40.5 , and a normal set o
liver function tests with an amylase of
```

```
PHYSICAL EXAMINATION: On admission, the patient was an uncomfortable obese white female in
moderate distress. Vital signs were stable and her temperature was 98.6.
SKIN: Normal.
HEENT: Normocephalic and atraumatic.
CHEST: Clear bilaterally.
CARDIAC: Regular rate and rhythm with normal S1 and S2 with a grade II/VI systolic ejection
murmur.
ABDOMEN: Soft, non-tender, and non-distended except for in the right upper quadrant where s
he had right upper quadrant pain and a positive Murphy's but no peritoneal signs.
LABORATORY EXAMINATION: Significant for a white count of 7.2, hematocrit of 40.5, and a nor
mal set of liver function tests with an amylase of 63.
```

Preprocessing - Removing Irrelevant Sections

- Removed sections
  - 'Discharge Date'
  - 'Attending',
  - 'Reason for override',
  - 'Dictated By'
  - 'Batch',
  - 'T'
  - 'CC',
  - 'D',
  - 'ENTERED BY',
  - 'Service',
  - 'eScription document',
  - 'FAMILY HISTORY',
  - 'Previous override information',
  - 'ID'
  - 'Infectious Disease'
  - ...

Preprocessing - NLTK

- correct verb’s other express:
  - e.g. branching. branched, branches->branch
- correct plural:
  - e.g. apples->apple
- remove stopwords:
  - e.g. the, a, this

Feature Extraction - TF-IDF

- TF-IDF:

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

	note	renal	sbp	myocardi	gastric	hip	cpap	bipap	right	knee	...	chronic	hemodialysi	coronari
0	0.000000	0.000000	0.0	0.044391	0.0	0.0	0.000000	0.0	0.030887	0.0	...	0.000000	0.000000	0.215112
1	0.000000	0.085303	0.0	0.000000	0.0	0.0	0.000000	0.0	0.034433	0.0	...	0.000000	0.085218	0.039968
2	0.020170	0.095439	0.0	0.027684	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.000000	0.333707	0.022358
3	0.050491	0.074661	0.0	0.000000	0.0	0.0	0.000000	0.0	0.000000	0.0	...	0.014982	0.059670	0.013993
4	0.045709	0.000000	0.0	0.041826	0.0	0.0	0.000000	0.0	0.014551	0.0	...	0.018085	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Feature Extraction - TF-IDF

- Calculate TF-IDF score of different terms, and rank the feature importance by “mean” of each term.
- Compare the ranking by the difference of TF-IDF score in datasets labeled “Y” and labed “U”.

Important Terms by TF-IDF

	obes	chronic	right	note	pt	apnea	sleep	gastric	knee	morbid	daili	bleed	cpap	weight	renal	ms	continu	chf	copd	bipap
0	0.000000	0.000000	0.030887	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.066011	0.027966	0.000000	0.000000	0.0
1	0.000000	0.000000	0.034433	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.00000	0.085303	0.000000	0.000000	0.104849	0.000000	0.0
2	0.000000	0.000000	0.000000	0.020170	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.00000	0.095439	0.000000	0.087204	0.000000	0.000000	0.0
3	0.000000	0.014982	0.000000	0.050491	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.265409	0.000000	0.000000	0.00000	0.074661	0.000000	0.054575	0.000000	0.000000	0.0
4	0.000000	0.018085	0.014551	0.045709	0.000000	0.000000	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.031098	0.039525	0.000000	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

	instruct	aortic	sbp	valv	fall	electrocardiogram	po	descend	dc	cardiomyopathi	lovenox	myocardi	infarct	hemodialysi	mg	coronari	sp	arteri
0	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.062571	0.000000	0.000000	0.000000	0.044391	0.042229	0.000000	0.000000	0.215112	0.000000	0.136483
1	0.154965	0.065654	0.0	0.065654	0.000000	0.000000	0.119494	0.000000	0.079185	0.000000	0.000000	0.000000	0.000000	0.085218	0.085710	0.039968	0.048717	0.038038
2	0.000000	0.000000	0.0	0.000000	0.085943	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.027684	0.026336	0.333707	0.000000	0.022358	0.000000	0.021279
3	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.020917	0.024421	0.027722	0.022841	0.000000	0.000000	0.000000	0.059670	0.130031	0.013993	0.000000	0.039951
4	0.000000	0.000000	0.0	0.000000	0.000000	0.177942	0.050497	0.088433	0.000000	0.000000	0.000000	0.041826	0.039789	0.000000	0.120735	0.000000	0.000000	0.016074
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

40 terms for Final method

- tfidf-selection
  - ['obes', 'chronic', 'right', 'note', 'pt', 'apnea', 'sleep', 'gastric','knee', 'morbid', 'daili', 'bleed', 'cpap', 'weight', 'renal', 'ms', 'continu', 'chf', 'copd', 'bipap']
  - ['instruct', 'aortic', 'sbp', 'valv', 'fall', 'electrocardiogram', 'po', 'descend', 'dc', 'cardiomyopathi', 'lovenox', 'myocardi', 'infarct', 'hemodialysi', 'mg', 'coronari', 'sp', 'arteri', 'qd', 'hip']

Method 1 result

- data preprocessing
- keyword amounts as feature(['coronary', 'diabetic', 'diabetes', 'dyslipidemia', 'hypertension', 'hypothyroidism','hyperlipidemia', 'gout', 'chronic','nondistended'])
- consider keyword as “non-obese”,’non-distended’
- not consider text “obese” -> more accurate

Method 1 result(cont.)

use textual dataset(400)

- random forest -> f1 score: 0.62857(highest)
- nn model -> f1 score: 0.60000

use merged dataset(428)

- random forest ->f1 score: 0.60000
- nn model-> f1 score: 0.62857(highest)

Final method

- Data preprocessing (train_textual)
  - Remove section
  - NLTK preprocessing: remove stop words, vocabulary formity
  - TF-IDF
  - # feature: 40
- Model
  - Random forest: f1 score: 0.57142
  - NN: f1 score: 0.60000

Results

- Local train/test results
  - use textual dataset to split 7:3 for train/test

	Acc	f1 score
Train	0.8142856955528259	0.7984496124031009
Test	0.7166666388511658	0.6964285714285715

  - use original train/test split

-	Acc	f1 score
Train	0.8224999904632568	0.8065395095367848
Test	0.8525000214576721	0.8459530026109661

Reference

- Uzuner O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009 Jul-Aug.
- Mishra NK, Cummo DM, Arnzen JJ, Bonander J. A rule-based approach for identifying obesity and its comorbidities in medical discharge summaries. J Am Med Inform Assoc. 2009 Jul-Aug.
- Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. J Am Med Inform Assoc. 2009 Jul-Aug.
- Natural Language Toolkit (NLTK): https://www.nltk.org

Team Member Contribution

	Dataset Parsing	Preprocessing			Feature Extraction		Model/Analysis		Result Presentation	
	Dataset Parsing	Text Recontruct	Remove Irrelevant Sections	NLTK	TF-IDF	Manual Selection	Random Forest	Neural Network	Slides	Oral
林亦盛 309551074	V				V		V	V	V	V
周君諦 310551136	V	V	V					V	V	V
陳昱銘 310554007	V			V		V	V	V	V	V

Thank you !

The codes for this experiment are available at

https://github.com/tim310579/Digital-Medicine-Case-Presentation-1.git