



2023

CHINA

INTERCHANGE

BEIJING | 25-26 AUGUST



Automating aCRF-SDTM Mapping: NLP and Machine Learning Approach for Efficient Data Annotation

Presented by Lee Wan, Senior Data Science Analyst II, DS, MacroStat
(China) Clinical Research Co., Ltd., Shanghai, China



Meet the Speaker

Lee Wan

Title: Senior Data Science Analyst II

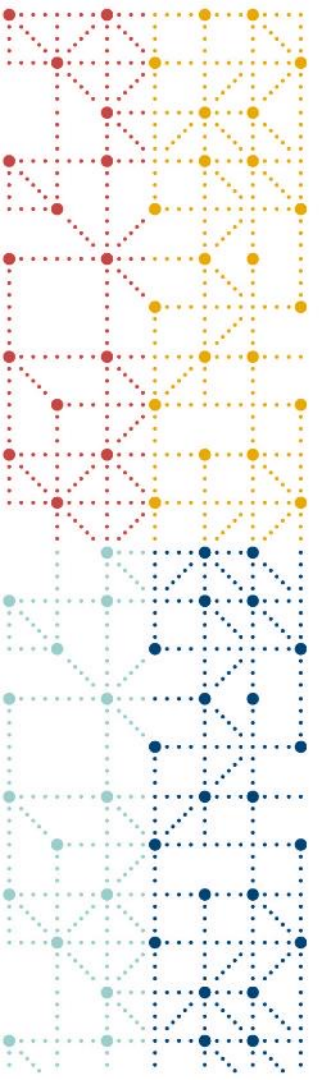
Organization: Tigermed

6 years of clinical statistical experience, with extensive programming experience in SAS, R, and Python.



Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*



Agenda

1. Background Introduction
2. Data Preparation
3. NLP
4. Machine Learning Model
5. Low-Cost Setup



Background Introduction



Background Introduction

aCRF到SDTM映射工作的重要性及目前面临的挑战:

- 重要性.
 - aCRF到SDTM映射是提交临床试验数据集必需的工作.
 - 映射工作直接影响了数据分析和结果报告的正确性.
 - 映射正确与否关系到最后研究结果的合规性和可靠性.
- 面临的挑战.
 - 映射工作目前主要依靠人工,非常费时费力.
 - 不同研究的CRF格式各异,映射难度大且重复性工作多.
 - 映射工作需要人工持续回顾核验,容易出现人为失误.



Background Introduction

NLP与机器学习技术简单介绍:

- 自然语言处理(Natural Language Processing, NLP).
 - 文本数字化—词向量.
- 机器学习.
 - 学习能力.
 - 数据驱动.
 - 模型构建.
 - 训练与优化



Background Introduction

准备工作:

- aCRF平台.
 - 统一映射格式.
 - 方便收集映射数据.
 - 简化操作流程.
- 研究设计说明(study design specification, SDS).
 - 包含CRF的各种详细信息, 比如变量名, 变量标签等等
 - 数据结构统一
 - CRF数据库的设计文档

Form: IC Form Name: 知情同意书 PageNumber: 7 Is Annotation: Yes

+ Field

7 / 92

实际大小

知情同意书 IC

DSDECOD = 签署知情同意书
DS (处置) DSCAT = 方案里程碑 DSSTERM = 签署知情同意书

知情同意书

No	知情同意书签署日期 (yyyy-MM-dd)	遵循方案版本号
1	RFICDTC DSSTDTC	PROVER in SUPPDM

Need Submit Yes No NA

Save

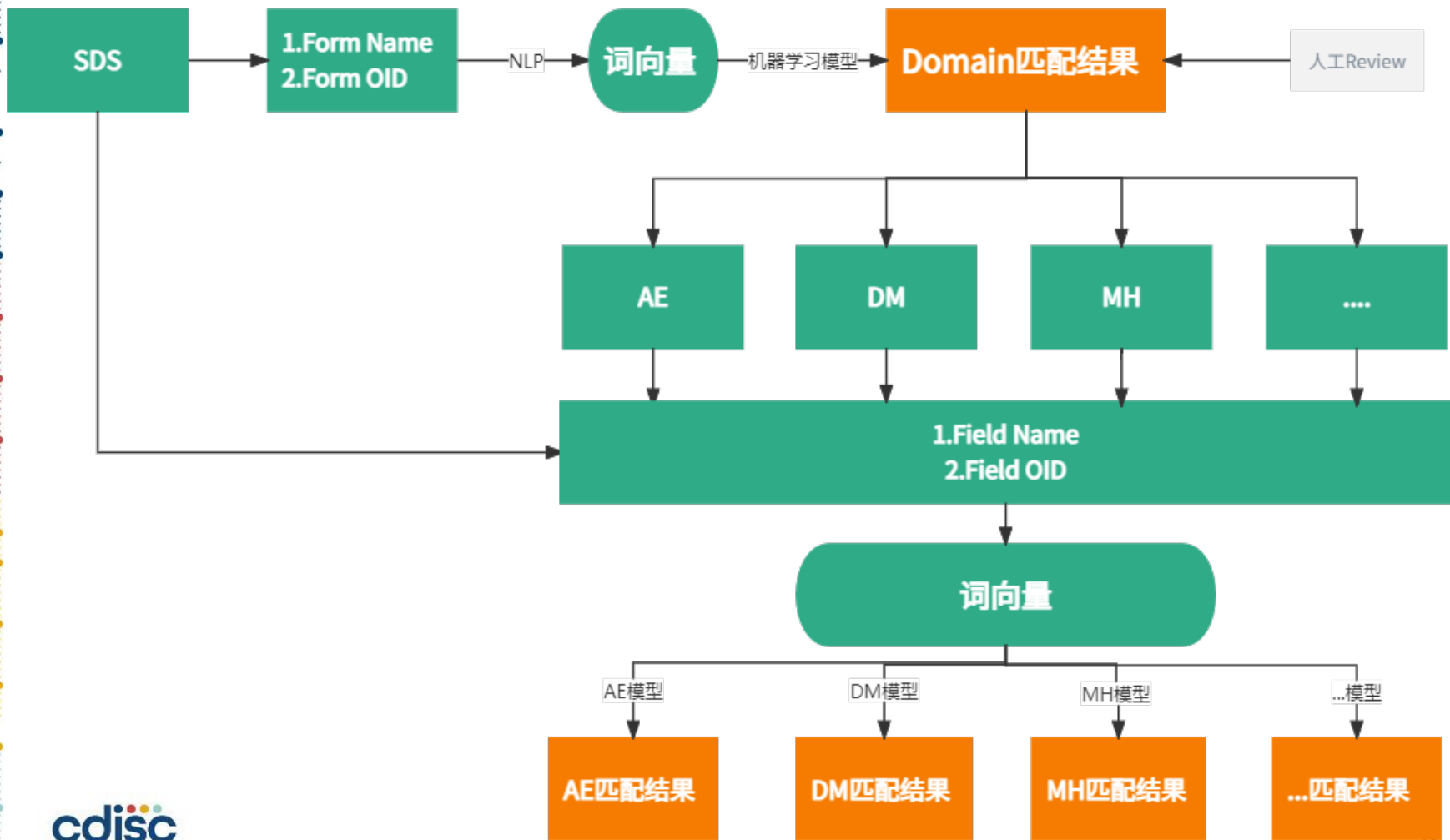
Domain

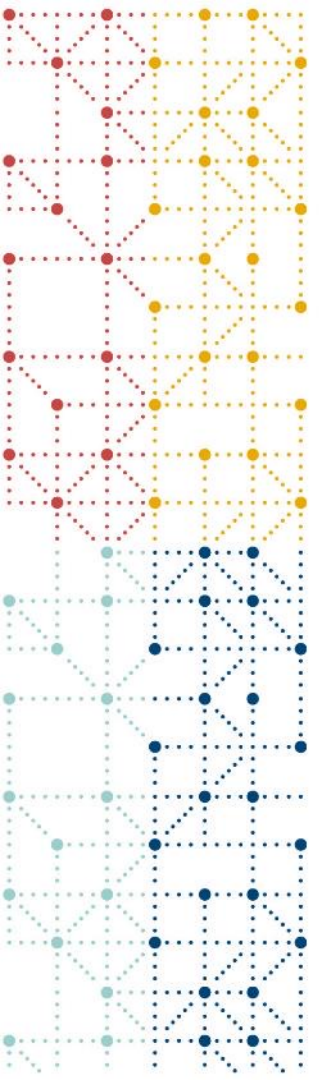
No	Domain	Description	Operation
1	DM	人口学	⊖
2	DS	处置	⊖

Assigned Value Variables

No	Domain	Variable	Value	Operation
1	DS	DSCAT	方案里程碑	⊖
2	DS	DSTERM	签署知情同意书	⊖
3	DS	DSDECOD	签署知情同意书	⊖

	A	B	C	D	E
	folderOID	moduleOID	FolderModuleName	ordinalPos	isVisible
1	SCR	SV	访视日期	0	1
2	SCR	IC	签署知情同意书	1	1
3	SCR	IE_1	入选/排除标准_1	2	1
4	SCR	DM	人口统计学	3	1
5	SCR	MH	既往及现病史	4	0
6	SCR	AH	过敏史	5	0
7	SCR	SU_BD	献血史	6	0
8	SCR	SU_B	吸烟史	7	0
9	SCR	SU_A	酒精滥用史	8	0
10	SCR	SU_C	药物滥用史	9	0
11	SCR	BMI	身高/体重/BMI	10	0
12	SCR	LB_X	胸部X线检查	11	0





Data Preparation

Domain Mapping

中文数据

Form OID	Form Name	Domain
AE	不良事件	AE
AH	过敏史	MH
BMI	身高/体重/BMI	VS

英文数据

Form OID	Form Name	Domain
AE	Adverse Events	AE
CNL	Tumor Results	TR
DA	Exposure	EX

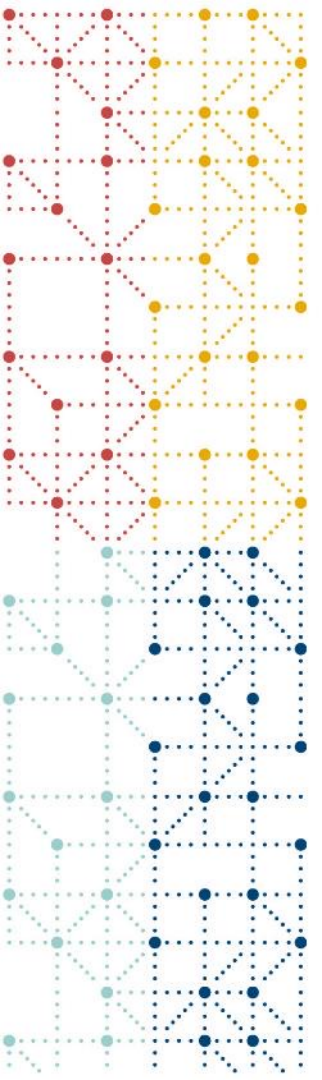
Variable Mapping

中文数据

Field OID	Field Name	Action
AESTDAT	开始日期	AESTDTC
AEYN	受试者给药后, 是否发生不良事件?	DROP
AEONGO	是否持续	SUPP

英文数据

Form OID	Form Name	Action
AESTDAT	Start Date	AESTDTC
AEOTHER	If Other Serious Situations, please specify	SUPP
AEYN	Has the subject experienced any new or changes to adverse events since signing the informed consent?	DROP



NLP

词袋 + Word2Vec + BERT

Field OID

AESTDAT

AEYN

AEONGO

- 非文本特征数据
- 词袋模型，以字母表中的字母作为词袋模型的词汇。

Field Name

开始日期

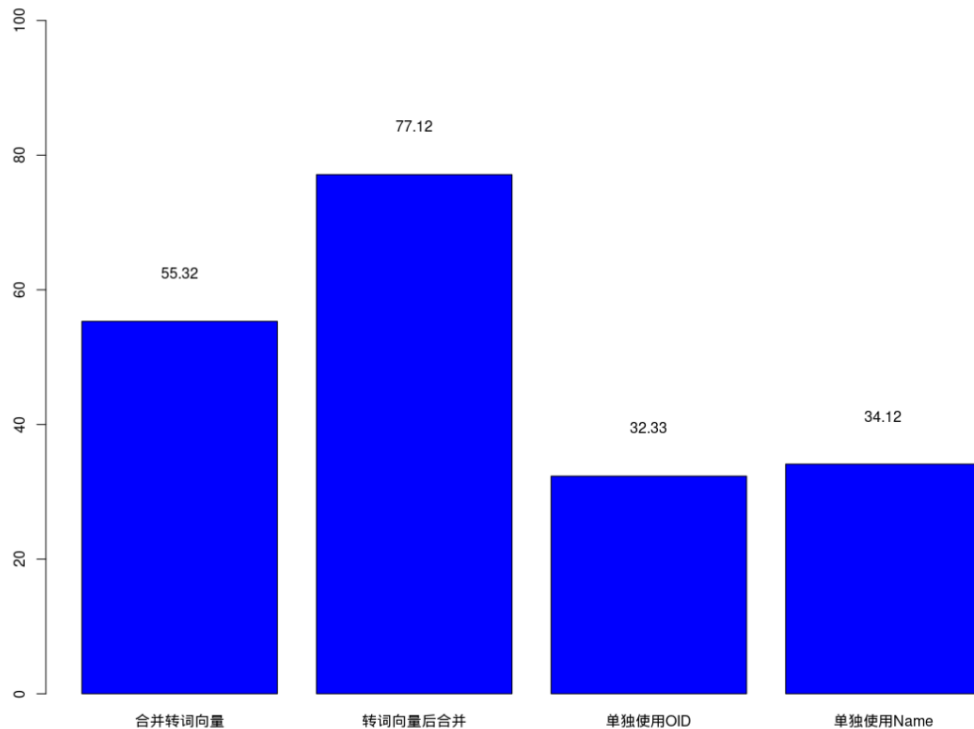
受试者给药后，是否发生不良事件？

是否持续

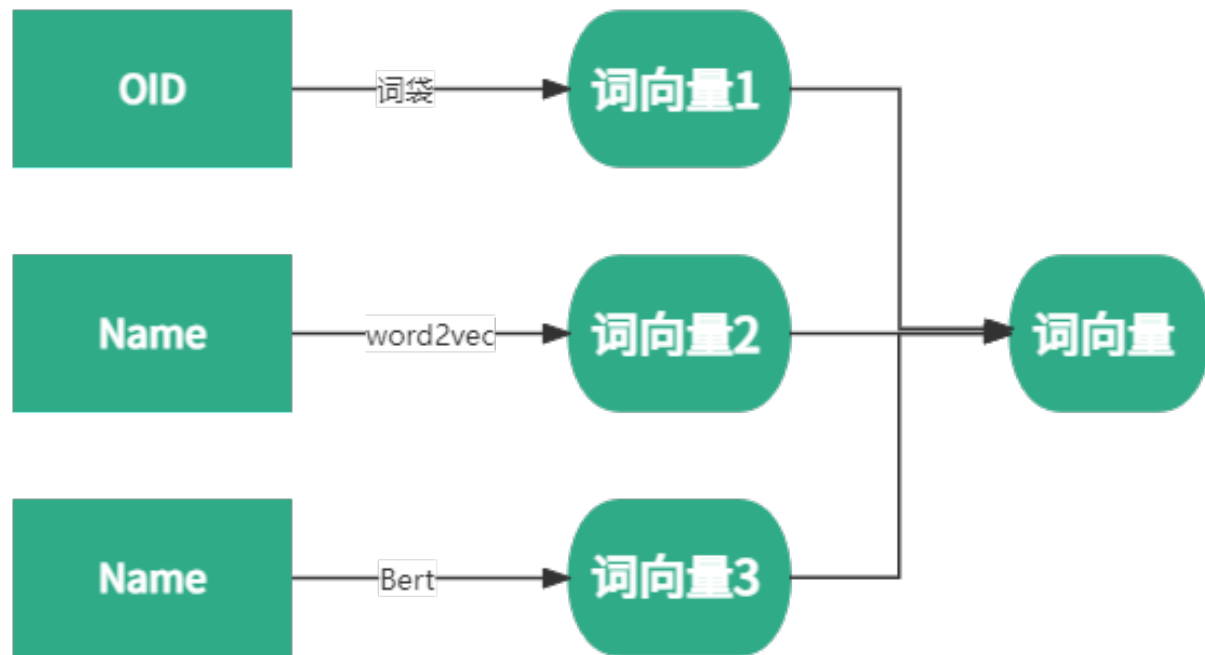
- 中文分词:使用jieba分词器对文本进行切分
- 英文分词:使用nltk分词器对文本进行切分.
- 使用Word2Vec对中英文文本进行词向量表示.
- 使用预训练BERT模型对中英文文本进行语义向量表示

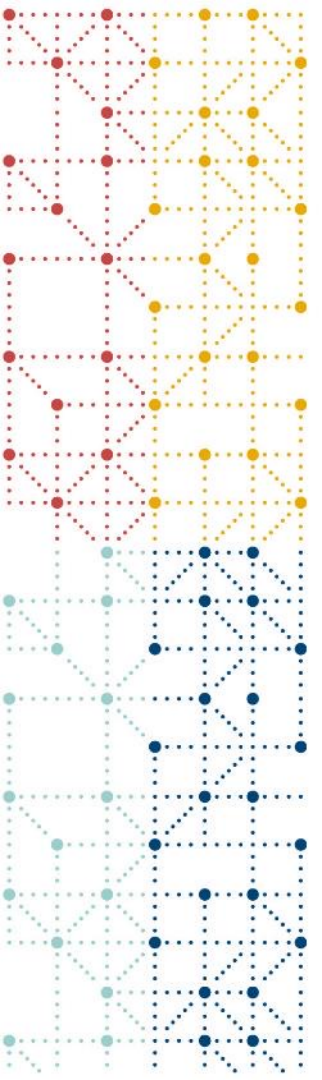
词向量测试

- 1、把OID和Name拼接起来再进行向量化
- 2、把OID和Name分别向量化再拼接起来
- 3、单独使用OID
- 4、单独使用Name



NLP





Machine Learning Model

分类模型

朴素贝叶斯

梯度提升树

支持向量机
SVM

神经网络

决策树

CatBoost

随机森林

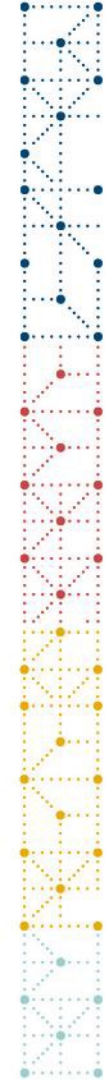
XGBoost

K近邻算法



投票融合



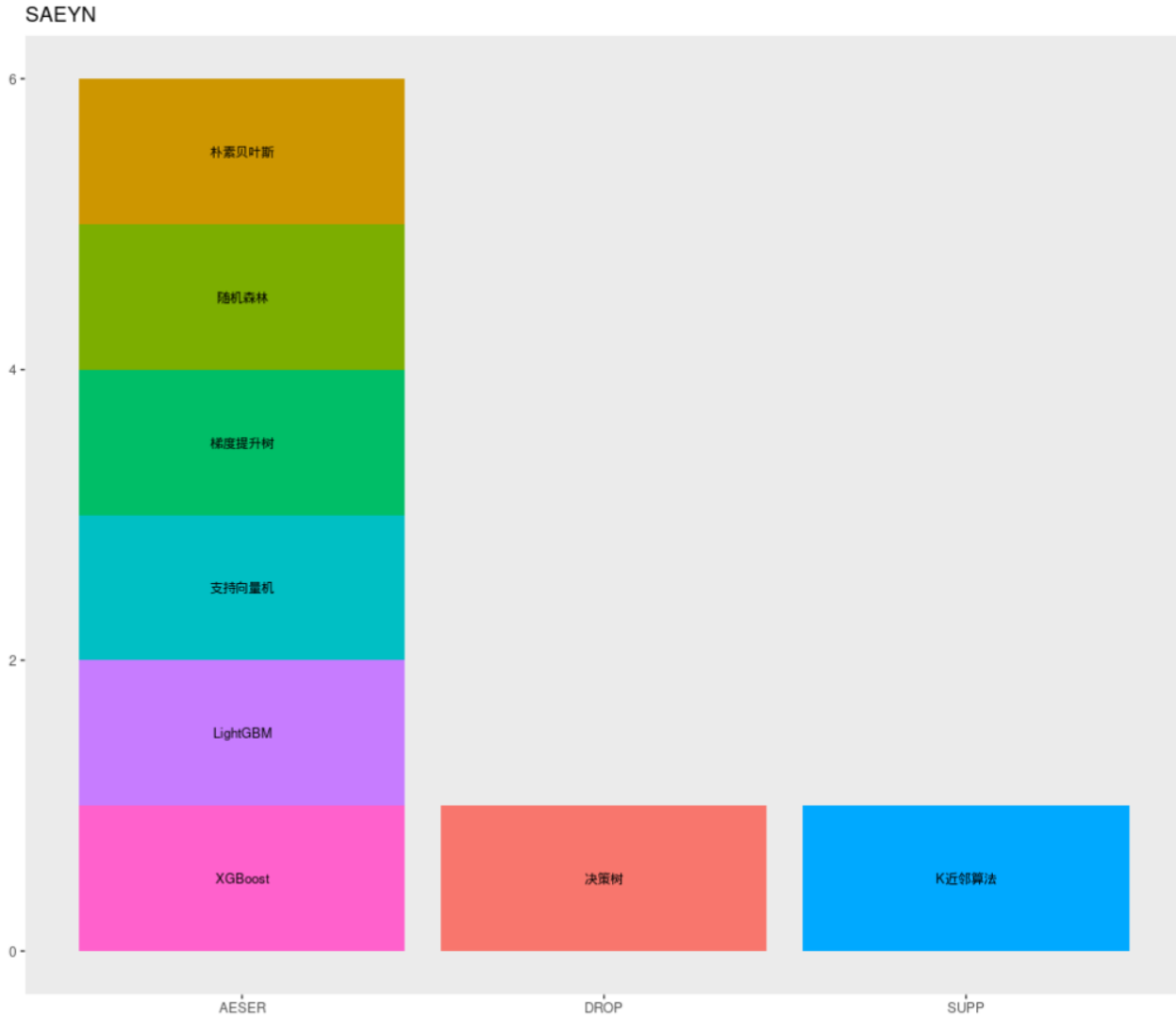


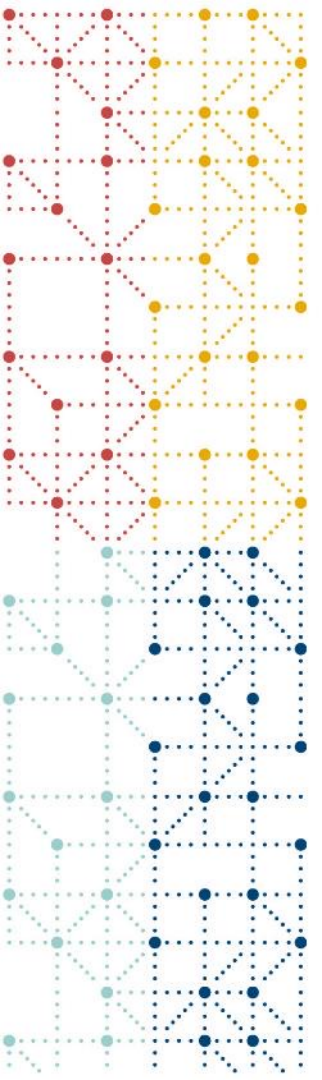
OID:SAEYN

NAME: Serious AE?

↓
AESER:6
DROP :1
SUPP :1

↓
SAEYN ==> AESER





Low-Cost Setup

W

我想让你担任机器学习和NLP方面的资深导师。我会写一些机器学习或NLP上的需求，你的工作就是用通俗易懂的术语来解释它们。这可能包括提供构建模型的分步说明、使用视觉效果演示各种技术，或建议在线资源以供进一步研究或提供相应的python代码。我的第一个建议请求是“我有一个长度小于等于12，由英文字母组成的类似变量名的数据集。我应该使用哪种方法让数据变成词向量来进行机器学习？”

View other drafts



你好!

我很乐意帮助你解决这个问题。

要将数据转换为词向量，以便进行机器学习，有几种方法。一种方法是使用单热编码。单热编码将每个单词映射到一个向量，其中向量的长度等于词汇表的大小。向量中的一个元素为1，表示该单词是所考虑的词。其他元素均为0。

另一种方法是使用词嵌入。词嵌入将每个单词映射到一个向量，其中向量的长度通常为100或更大。向量的元素代表单词的意义，而不是单词本身。

在你的例子中，你的数据集包含长度小于等于12，由英文字母组成的类似变量名。对于这种类型的数据集，使用词嵌入通常是更好的选择。这是因为单热编码无法捕捉单词之间的语义关系。例如，单热编码将单词“dog”和“cat”映射到两个不同的向量。然而，这两个单词之间存在语义关系，即它们都是动物。词嵌入能够捕捉这种语义关系，这是因为它们将单词映射到向量，其中向量的元素代表单词的意义。

有几种不同的方法可以生成词嵌入。一种方法是使用神经网络。神经网络是一种机器学习模型，可以从数据中学习。在这种情况下，神经网络将从数据中学习单词的意义。另一种方法是使用预训练的词嵌入。预训练的词嵌入是使用大型文本语料库训练的词嵌入。这些词嵌入可以免费获得，可以用在各种任务中。

我希望这有帮助！如果你还有其他问题，请告诉我。

让我知道如果你在实现过的知识。

你可以使用神经网络来学习字符嵌入。一种简单的方式是使用一个嵌入层 (Embedding Layer)，将每个字符的整数编码映射到一个固定维度的向量。这个嵌入层的权重将作为学习

```
1 import pandas as pd
2 from sklearn.feature_extraction.text import TfidfVectorizer
3 from sklearn.model_selection import train_test_split
4 from sklearn.naive_bayes import MultinomialNB
5
6 # 假设数据在dataframes df中,第一列是text1,第二列是text2,第三列是Label
7
8 # 对text1和text2分别进行词向量化
9 vectorizer1 = TfidfVectorizer()
10 vectorizer2 = TfidfVectorizer()
11 df["Name"] = df["Name"].apply(lambda x: " ".join(jieba.cut(x)))
12 df["OID"] = df["OID"].apply(lambda x: " ".join(nltk.word_tokenize(x)))
13
14 text1_vectors = vectorizer1.fit_transform(df["OID"])
15 text2_vectors = vectorizer2.fit_transform(df["Name"])
16
17 # 拼接词向量作为特征
18 X = np.hstack([text1_vectors.toarray(), text2_vectors.toarray()])
19 y = df["Class"]
20
21 # 划分训练集和测试集
22 X_train, X_test, y_train, y_test = train_test_split(
23     X, y, test_size=0.2, random_state=42
24 )
25
26 # 使用朴素贝叶斯进行分类
27 nb = MultinomialNB()
28 nb.fit(X_train, y_train)
29
30 # 评估测试集上的性能
31 print("Accuracy:", nb.score(X_test, y_test))
```

```
Accuracy: 0.5819672131147541
```

```
action.text import TfidfVectorizer
on import train_test_split
import MultinomialNB

中,第一列是text1,第二列是text2,第三列是Label
.xlsx")
向量化
izer()
izer()
1.fit_transform(df["OID"])
2.fit_transform(df["Name"])

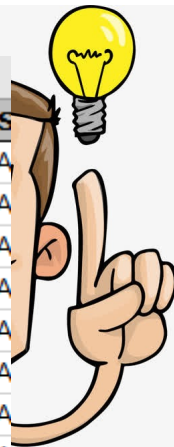
ors.toarray(), text2_vectors.toarray())

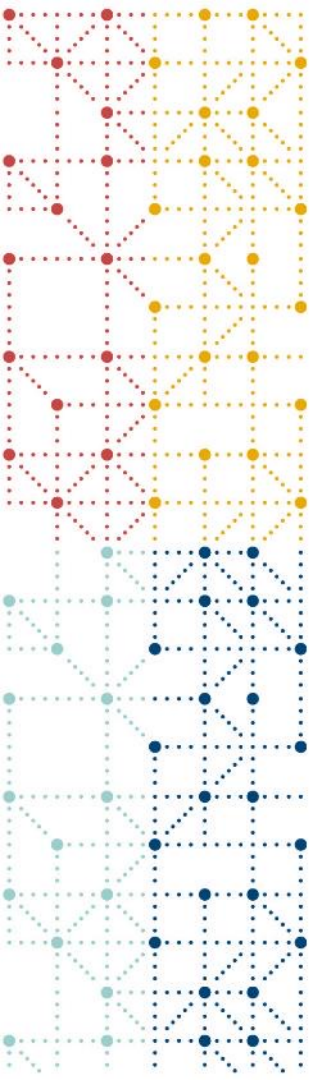
y_test = train_test_split(
random_state=42

e(X_test, y_test))
```


后续难点与挑战

A	B	C	D	S
FormOID	FieldNo	FieldName	ControlType	
AE	AEYN	Has the subject experienced a	水平单选框	A
AE	AELBL	Yes, record each adverse even	文本 标签	A
AE	AETERM	Adverse Event Term	长文本	A
AE	AESTDAT	Start Date	日期框	A
AE	AESTTIM	Start Time	日期框	A
AE	AEENDAT	End Date	日期框	A
AE	AEENTIM	End Time	日期框	A
AE	AEOUT	Outcome	下拉框	A
AE	AEIG	Grade	下拉框	A
AE	AEREL	Relationship to IMP	下拉框	A
AE	AEACN	Action taken with IMP	下拉框	A
AE	AEACNP	If Other, please specify	长文本	A
AE	AEACNO	Action Taken with AE	多选框	A
AE	AEACNOS	If Other, please specify	长文本	A





Thank You!

cdisc