



BUSINESS 2020

JULYED

## 探索文本分类

Joe

<https://www.julyedu.com/>



## CONTENTS

01



常见的文本分类模型

02



TextCNN简介

03



TextCNN实战

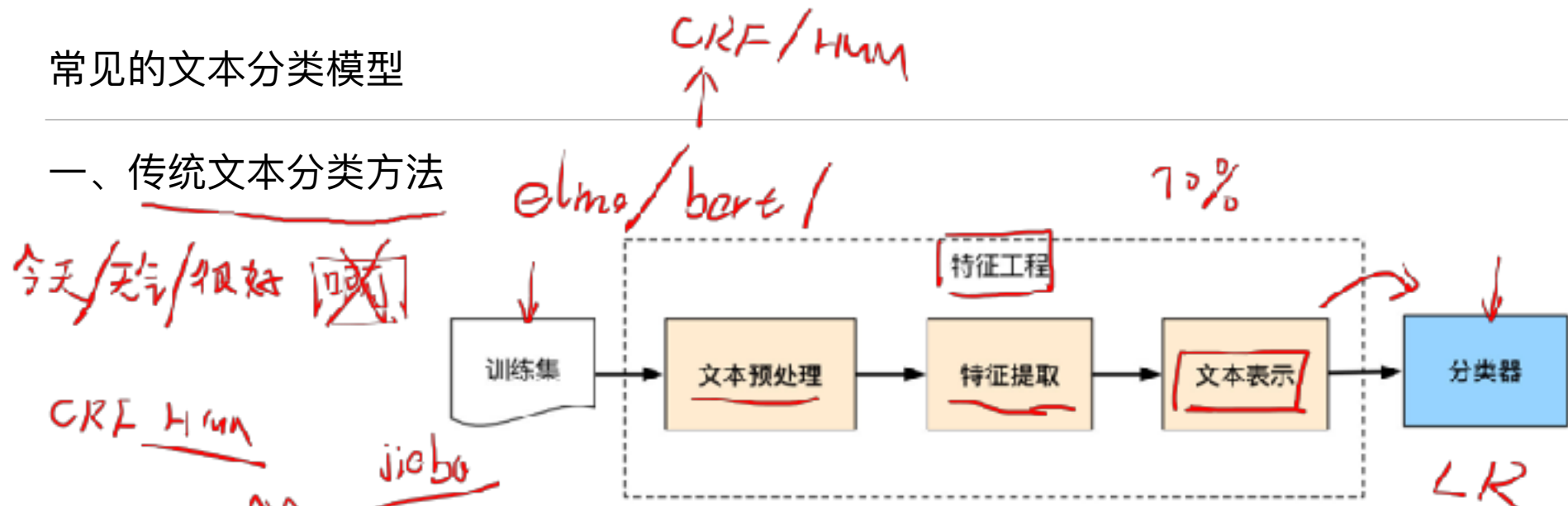


# 01

## 常见的文本分类模型

# 常见的文本分类模型

## 一、传统文本分类方法



文本预处理:

去除异常符号、分词、去除停用词

特征提取:

文本长度、词频等

文本表示:

Bag Of Words、TF-IDF、TextRank等

分类器:

朴素贝叶斯、LR、SVM等

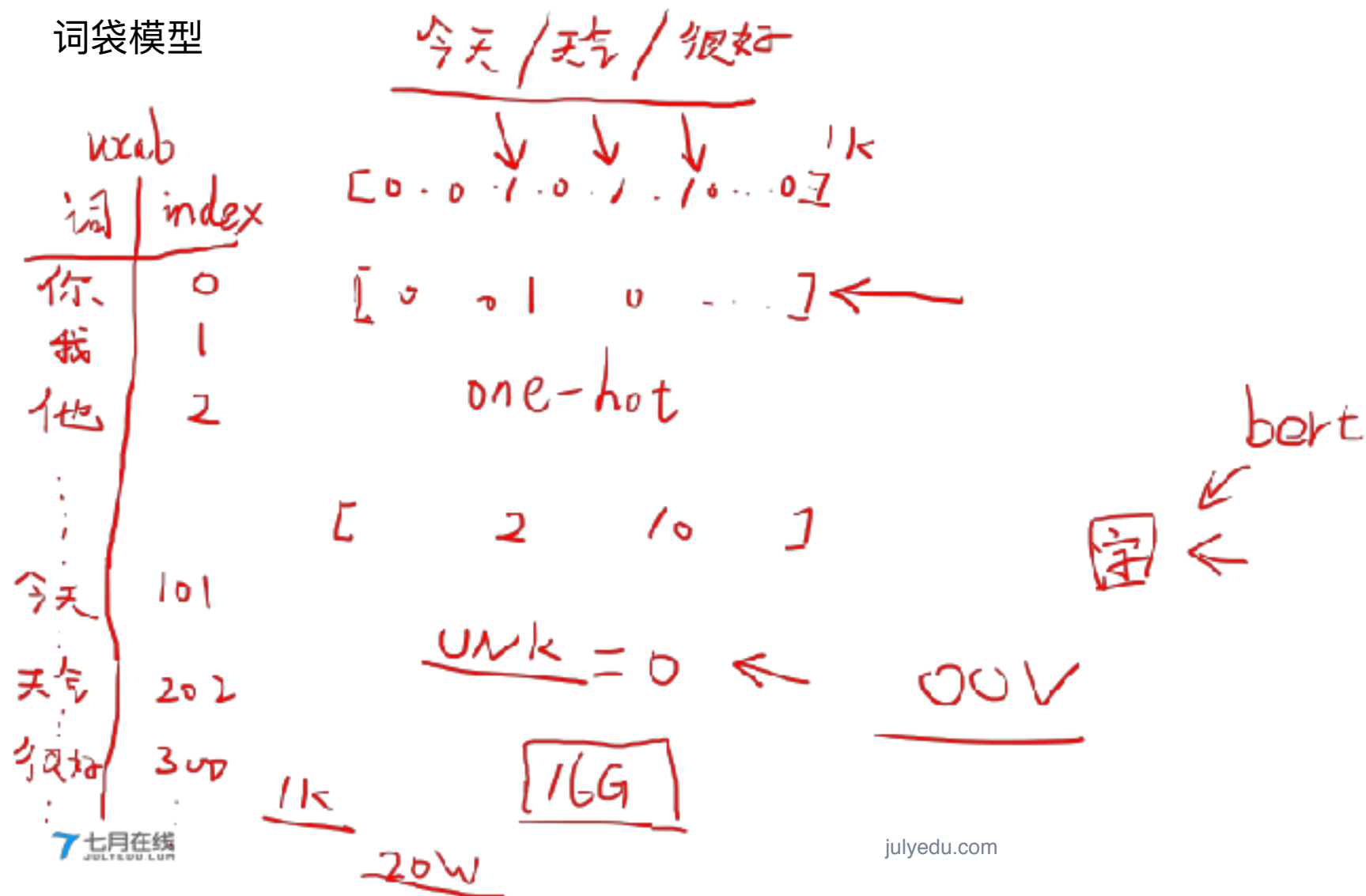
Handwritten formula for Naive Bayes classification:

$$P(\text{垃圾}|\text{信用卡}) = \frac{P(\text{信用卡}|\text{垃圾}) \cdot P(\text{信用卡})}{P(\text{信用卡})}$$

Handwritten notes around the formula include 'entropy', 'pagerank', 'n.', and '> 0.5'.

## 常见的文本分类模型

### 词袋模型



## 常见的文本分类模型

## TF-IDF

**TF:** 衡量一个词在文档中出现得有多频繁。

**IDF:** 衡量一个词有多重要。有些词出现的很多，但是作用不大。比如'is', 'the', 'and'之类的。为了平衡，我们把罕见的词的重要性 (weight) 搞高，把常见词的重要性搞低。

~~IDF = \log(\text{文档总数} / \text{含有}t\text{的文档总数}).~~

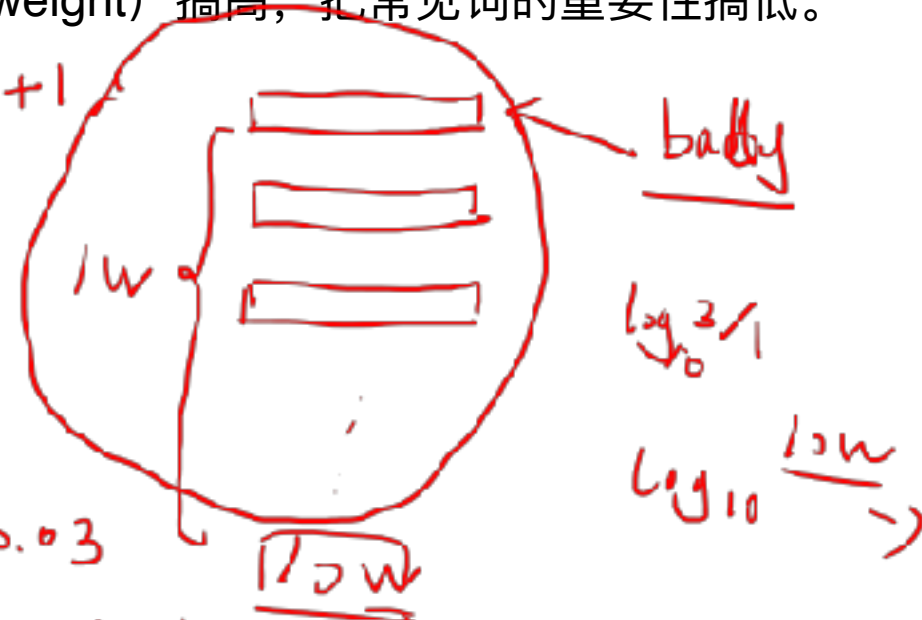
~~TF-IDF = TF \* IDF~~

$$\frac{10 \text{ W.}}{\downarrow} \quad \underline{9.3} =$$

100↑ 3次 TF:  $3/100 = 0.03$

100 10F  $\log 10000/100 = \log 100 = 2$

$$0.03 \times 2 = 0.06$$

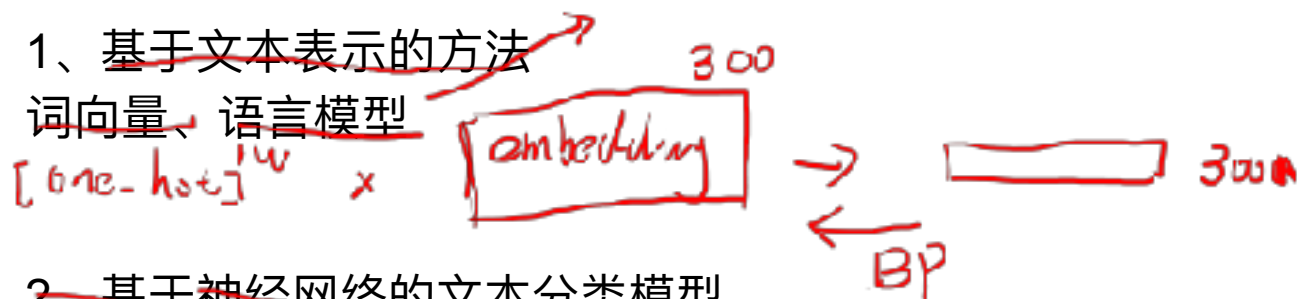


# 常见的文本分类模型

## 一、深度学习文本分类方法

### 1、基于文本表示的方法

词向量、语言模型



### 2、基于神经网络的文本分类模型

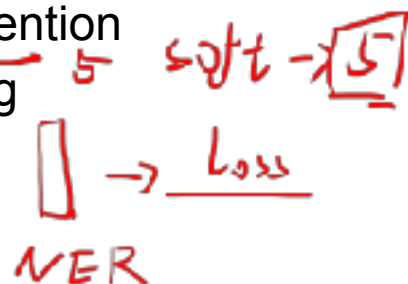
分类较为明显: fasttext

短文本: TextCNN

长文本: HAN

通吃: 精调的 Bi-LSTM+Attention

终极方案: bert+fine tuning





## 02

## TextCNN简介



## TextCNN简介

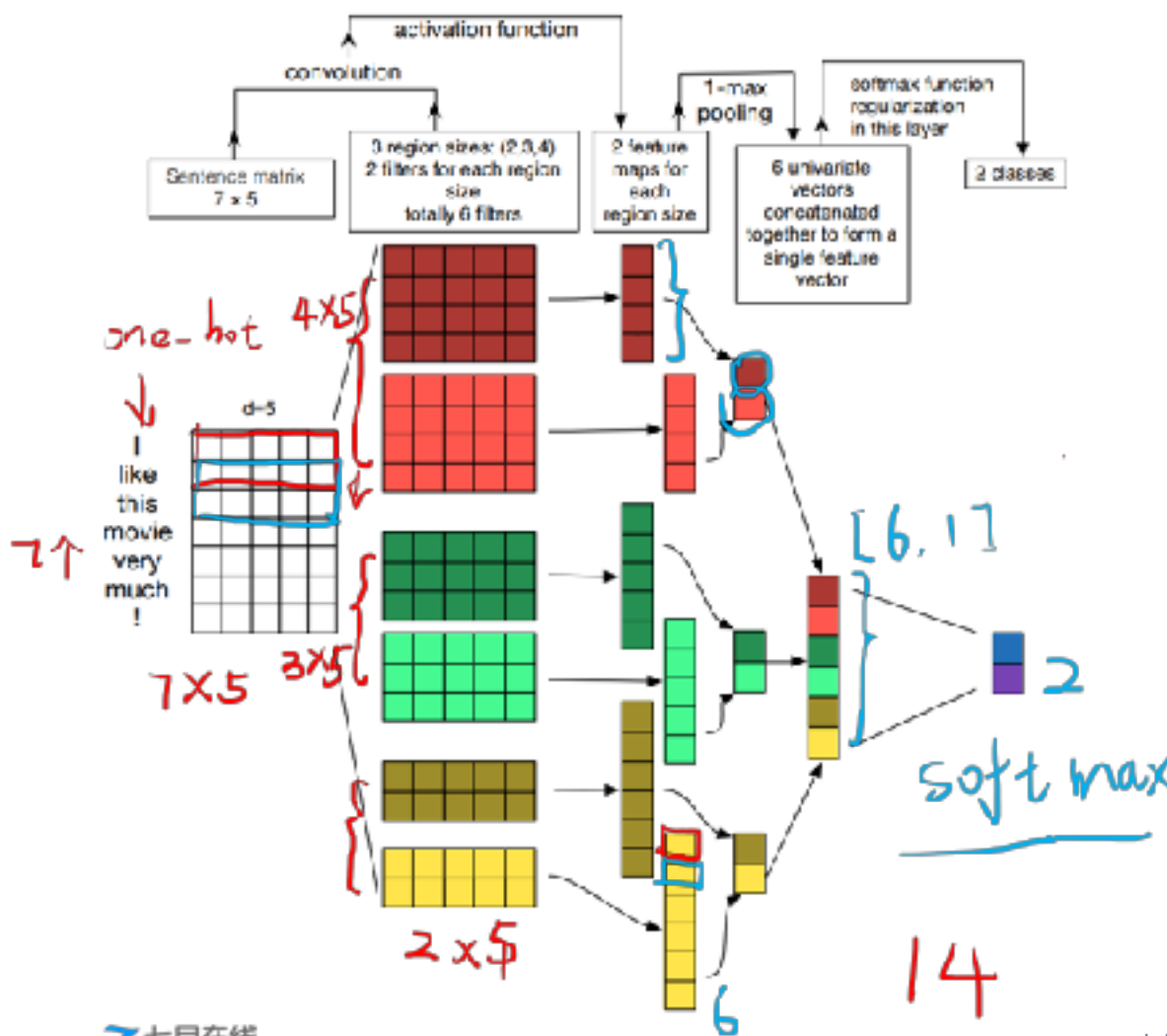
---

<https://arxiv.org/pdf/1408.5882.pdf>

<https://arxiv.org/pdf/1510.03820.pdf> 

TextCNN是将**卷积神经网络CNN**应用到**文本分类**任务，利用**多个不同size的kernel**来提取句子中的关键信息（类似于多窗口大小的ngram），从而能够更好地捕捉局部相关性。

## TextCNN简介



**Embedding:** 第一层是图中最左边的7乘5的句子矩阵，每行是词向量，维度=5，这个可以类比为图像中的原始像素点。

**Convolution:** 然后经过  $\text{kernel\_sizes}=(2,3,4)$  的一维卷积层，每个  $\text{kernel\_size}$  有两个输出 channel。

**MaxPolling:** 第三层是一个1-max pooling层，这样不同长度句子经过pooling层之后都能变成定长的表示。

**FullConnection and Softmax:** 最后接一层全连接的softmax层，输出每个类别的概率。

TextCNN最大优势网络结构简单，因此参数数目少，计算量少，训练速度快。

$$\frac{W-d}{s} + 1$$

Handwritten annotations for the formula:  $W$  (width),  $d$  (kernel size),  $s$  (stride), and  $\text{step}=1$ .



# 03

## TextCNN实战

## TextCNN实战

---

代码我们采用pytorch框架进行编写，详见py文件



微信扫一扫关注我们

# THANKS

---

掌握知识最终还是要靠自己去实践总结

---