文本分类任务介绍

鲲鹏创新实践课：鲲鹏应用数据分析与管理实战

前言

- 本章将主要介绍文本分类任务，文本分类指的是计算机通过算法对输入的文本按照一定的类目体系进行自动化归类的过程。在人工智能浪潮席卷全球的今天，文本分类技术已经被广泛地应用在文本审核、广告过滤、情感分析和反黄识别等NLP领域。

目标

- 学完本课程后，您将能够：
  - 理解文本分类的总体流程
  - 理解文本分类的常用预处理方法
  - 了解文本分类的评估方法

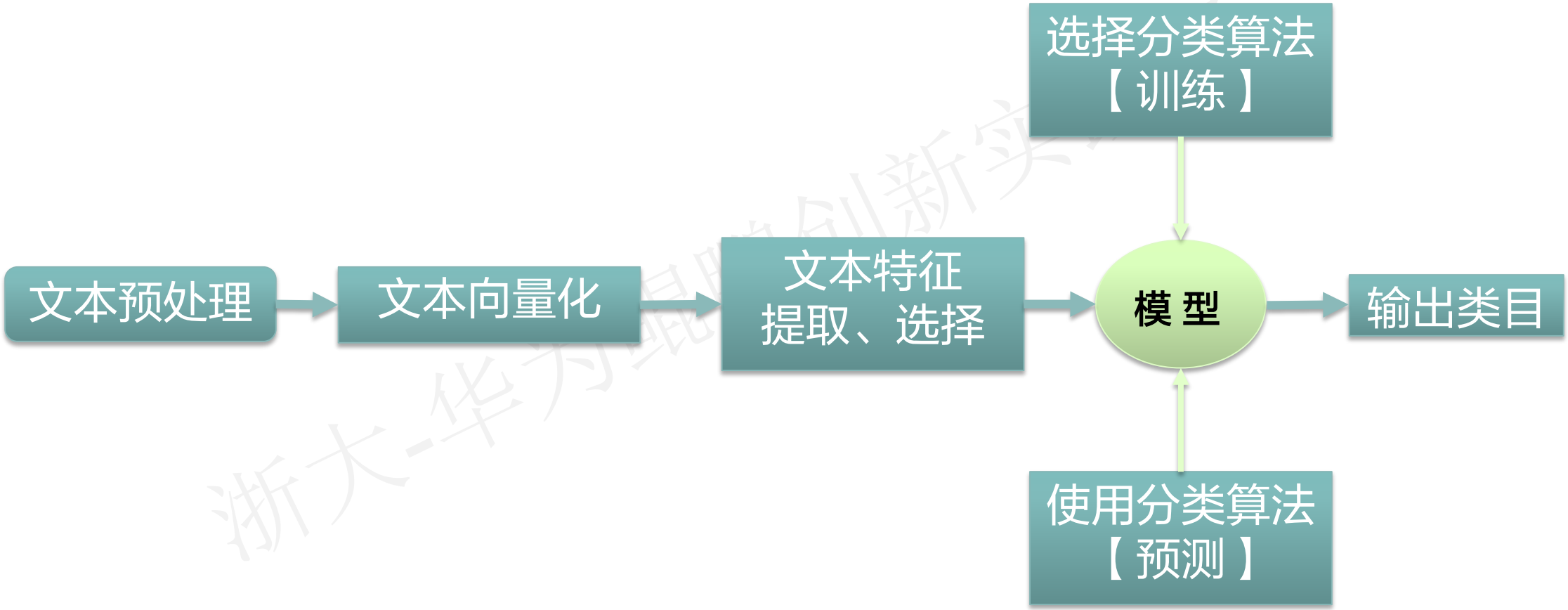
目录

1. **文本分类的总体流程**
2. 预处理介绍
3. 词向量介绍
4. 评估指标介绍

文本分类

- 文本分类（Text classification）算法是大规模处理文本数据的各种软件系统的核心。

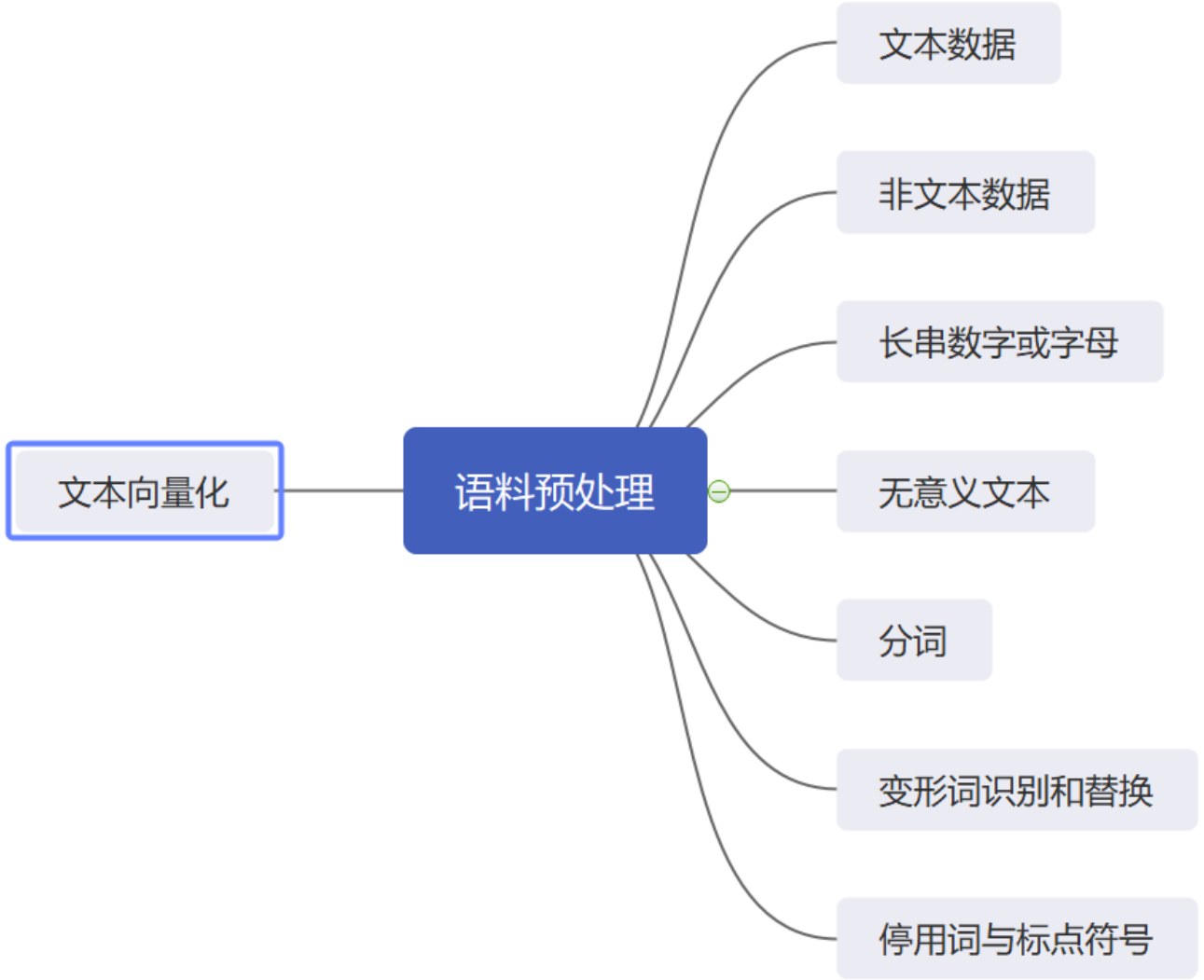
文本分类的总体流程



目录

1. 文本分类的总体流程
2. **预处理介绍**
3. 词向量介绍
4. 评估指标介绍

语料预处理



非文本数据

- 实际工业环境下使用到的文本数据很多时候都是来自爬虫，因此文本中通常会附带有HTML标签、URL地址等非文本内容，需要清除这些脏数据。

```
html='<p class="txt" node-type="feed_list_content" nick-name="人民日报">\n                                【<a href="
http://s.weibo.com/weibo?q=" target="_blank">#13个求职新方向#</a>！有你心动的职业吗<img src="//img.t.sina
js.cn/t4/appstyle/expression/ext/normal/aa/2018new_bingbujiandan_org.png" title="[并不简单]" alt="[并不简
单]" class="face">】近日，人社部、市场监管总局、统计局联合发布13个新职业，既有现在流行的人工智能、<em cla
ss="s-color-red">大</em><em class="s-color-red">数据</em>、云计算、物联网等工程技术人员，也有电子竞技员、
无人机驾驶员等新颖工种…你心动了吗？哪些学校开设了相关专业值得关注？戳图了解↓↓转给正在求学的TA！
      </p>
```

长串数字或字母

- 非特定情境下，文本中的长串数字通常代表手机号、车牌号、用户ID等内容，对分类结果无帮助，甚至增加分类难度，而且会使得词汇表的总词汇数增加很多，这时可以把它们去除或者替换成统一的标识，如 ”NUM”。
- 特定情境下，如微博或微信聊天数据，文本中含有大量表情符号，表情符号在文本中是以长串数字或字母形式出现，而这些表情数据在情感分析等场景下有巨大作用。

无意义文本

- Web文本中还有无意义的文本会影响模型的学习，如：
  - 广告
  - 版权信息
  - 个性签名

分词

- 传统中文自然语言处理最重要的一步就是分词，相比于英文有天然的空格符作为词与词之间的间隔标志，中文中词的提取必须通过HMM、CRF、RNN等序列标注算法来实现。

```
import jieba
print(list(jieba.cut('我爱杭州西湖')))
```

```
['我', '爱', '杭州', '西湖']
```

变形词识别和替换 (1)

- 在一些特殊的文本分类场景下，如广告识别、涉黄涉政审核等，会有大量的变形词出现，以躲避算法的检测，中文词的变形方法通常有：特殊符号替换、同音近型替换、简繁替换等。

想赚钱，请加微信《y123456》，改变命运，从现在开始

想赚钱，请加威信《y一23四5六》，改变命运，从现在开始

变形词识别和替换 (2)

- 变形词识别和替换方法：
  - 变形词映射表
  - 拼音首字母鉴别同音替换
  - 词向量对比变形词与原词的语义关联度
  - ……

停用词与标点符号

- 停用词指的是诸如代词、介词、连接词等不包含或包含极少语义的词，另外标点符号也可以被认为是一种停用词。通常情况下，**在文本中去掉这些停用词能够使模型更好地去拟合实际的语义特征，从而增加模型的泛化能力。**

文本向量化

- 文本向量化：将文本表示成一系列能够表达文本语义的向量。常用的向量化算法有：
  - one-hot
  - TF-IDF
  - word2vec
    - CBOW模型
    - Skip-gram模型
  - doc2vec/str2vec
    - DM（Distributed Memory）
    - DBOW（Distributed Bag of Words）

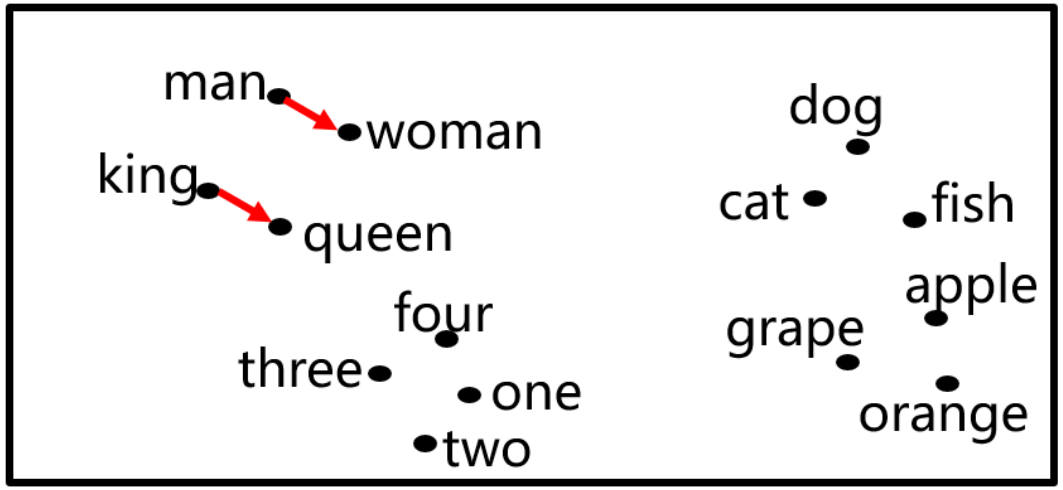
目录

1. 文本分类的总体流程
2. 预处理介绍
3. **词向量介绍**
4. 评估指标介绍

什么是词向量

- 自然语言是一套表达语义的复杂系统，在这套系统中，词是语义的基本单元。词向量是一种试图将**词表达为数值向量**的技术，并且该向量能够表达词的语义特征。

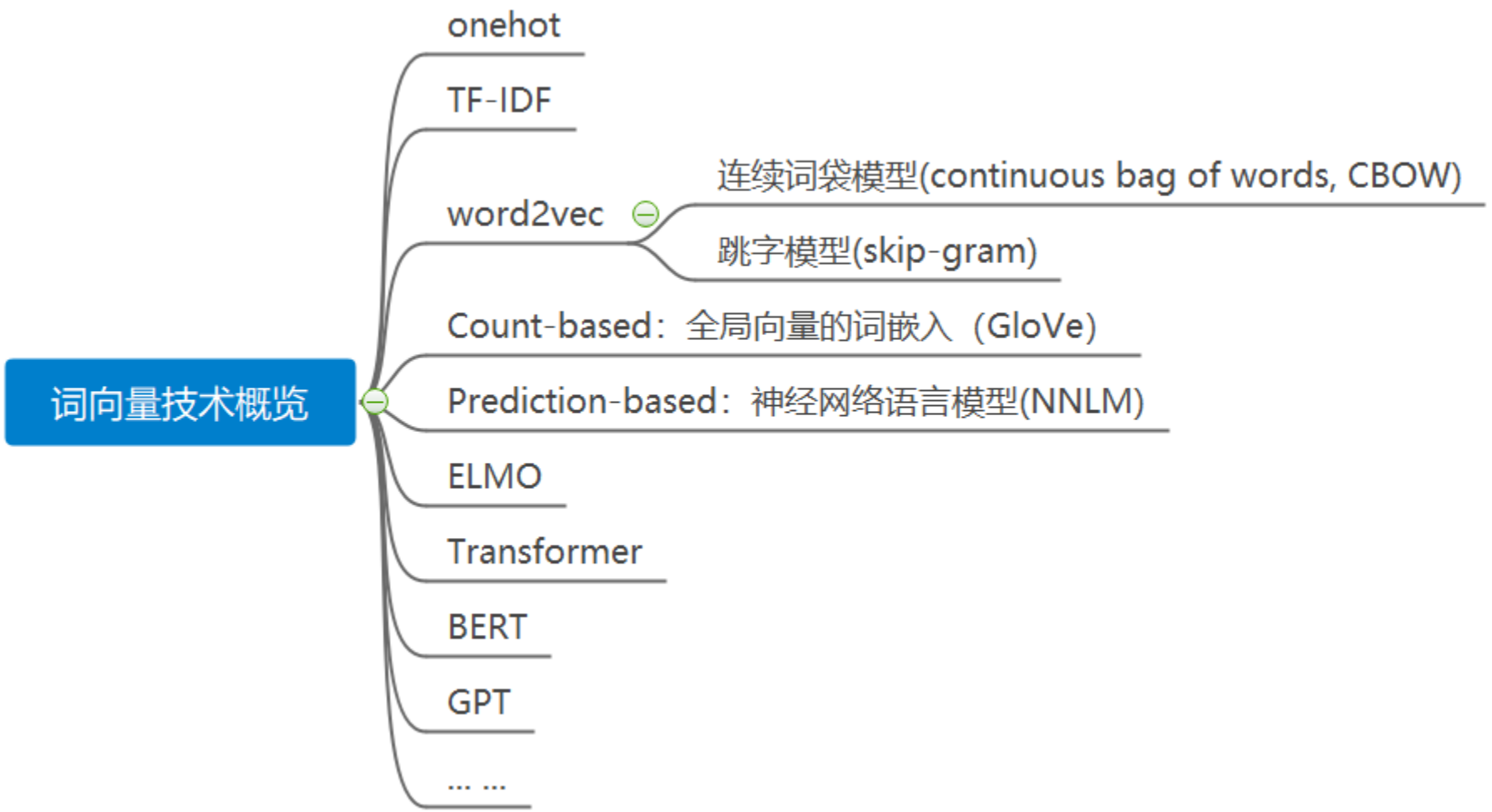
	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	−1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97



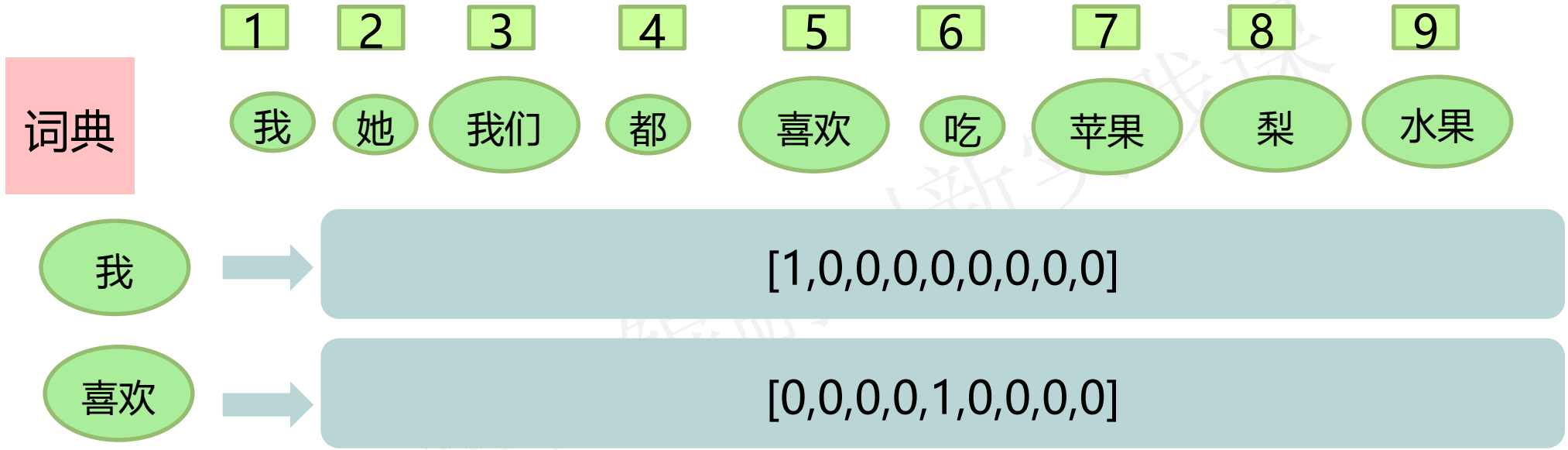
词向量的产生意义

- 将语言中的字符串表达成了数值向量，形式如: [0.12,0.102,0.031,0.26,...]；
- 基于统计的词向量能将词语映射到一个向量空间，语义越相近，它们的距离也越相近，这样的好处是词向量能够代表语义；
- 维度有限,比如200维，300维等，不至于产生维度爆炸的问题，这使得广泛应用词向量成为可能；
- 在使用机器学习、深度学习在处理自然语言时，能够进行迁移学习，将训练好的词向量作为embedding层的输入进行使用。
- 在大规模语料上预训练的词向量常常可以应用于下游自然语言处理任务中。
- 可以应用预训练的词向量求近义词和类比词。

词向量技术概览



One-hot编码



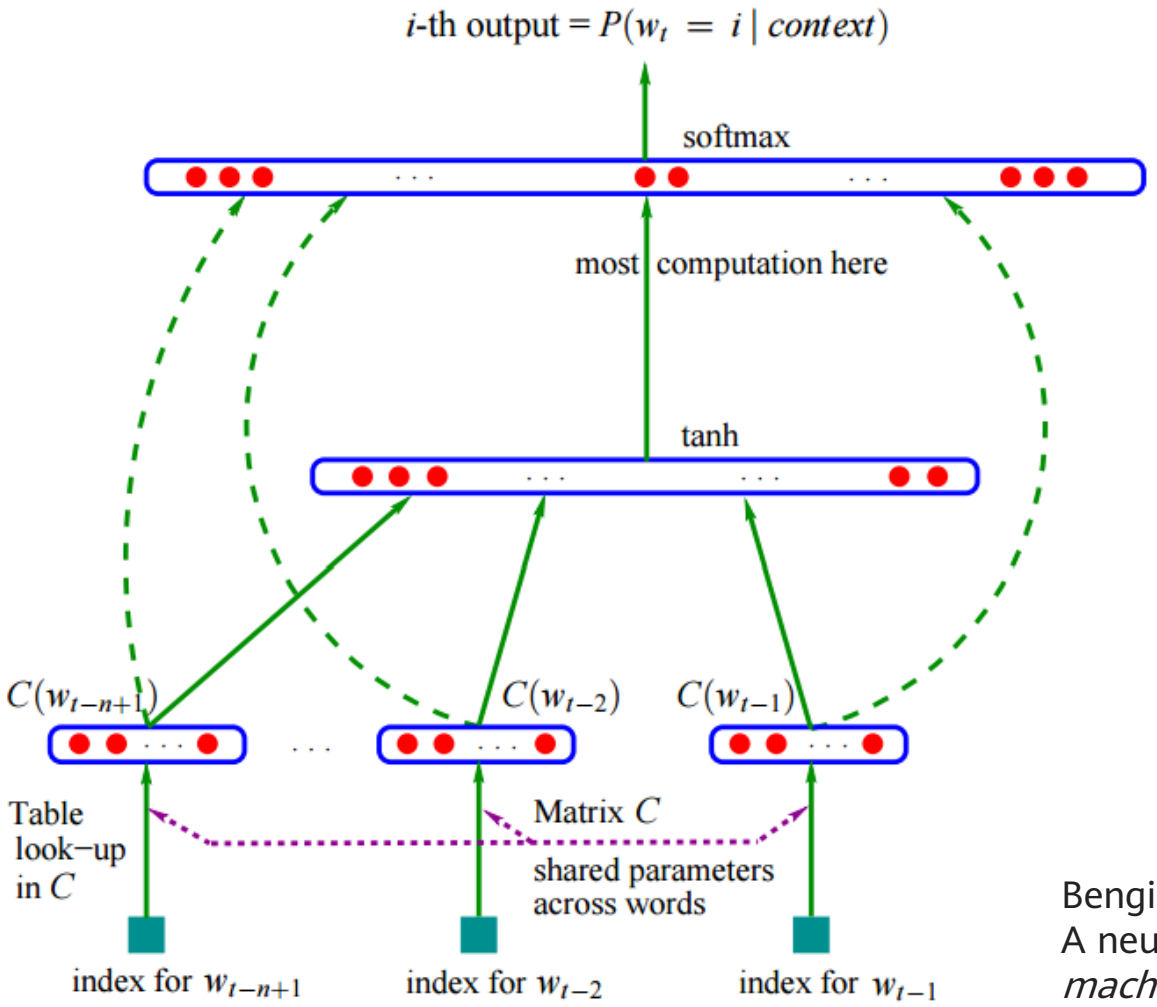
优缺点分析

- One-hot词向量解决了两个问题：
  - 词转化成了数字向量；
  - 句向量可以通过词向量相加得到。
- 局限：
  - 每个词不管出现多少次，都是唯一的位置为1，该位置并不能体现词的重要性；
  - 当词典非常大的时候，产生维数灾难；
  - 无法体现词与词之间的关联性；

神经网络语言模型

- 模型非常简陋，只有输入层、隐层和输出层；
- 模型从隐层到输出层经过softmax，预测词典中每个词可能是下一个词的概率，计算量非常大；
- 词向量只是其附属产品。

神经网络语言模型 – 示意



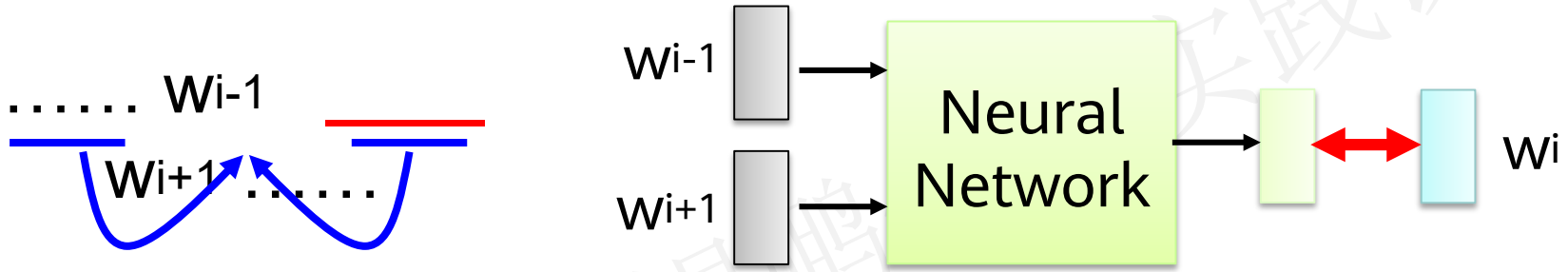
Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

互动问题

填空任务	造句任务	
我想喝一杯__解渴.	…牛奶…	…茶水…
你今天__看起来不错.	…心情…	….状态…
杭州是___的一个城市.	…省会…	….城市…
封闭的语义	封闭的上下文	

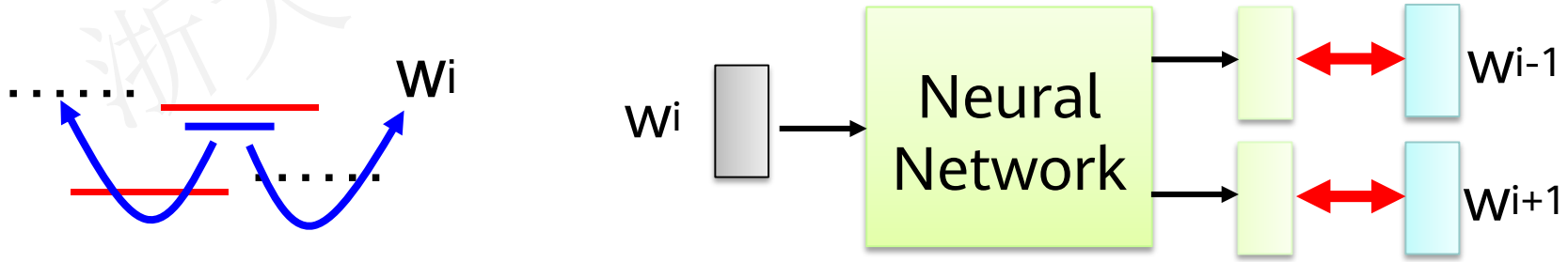
Word2vec – 实现机制

- Continuous bag of word (CBOW)



predicting the word given its context

- Skip-gram



predicting the context given a word

Word2vec - 获得词向量

```
model_path = "model/wordvec.model"
wordvec = word2vec.Word2Vec.load(model_path)
#获得"明亮"的词向量
wordvec.wv.get_vector("明亮")
```

```
array([ 1.07780874e+00, -1.79741406e+00, -1.28020555e-01, -2.41475165e-01,
       -1.23322749e+00,  8.50561917e-01, -1.09388351e+00,  3.64339948e-01,
       -1.43524313e+00,  6.58885241e-02,  5.67824244e-01, -1.76249480e+00,
       -1.44883171e-01,  1.00050099e-01,  1.30994678e+00,  1.08038366e+00,
       -7.35839546e-01,  6.19860113e-01,  6.32545471e-01,  2.95912504e-01,
       -1.63466275e+00,  1.82607925e+00, -1.61556453e-01,  4.54814583e-01,
        4.60491091e-01,  7.15260267e-01,  9.63576674e-01, -8.86829734e-01,
       -7.89200142e-02, -1.27947986e+00, -2.02853775e+00, -2.04116035e+00,
       -5.25810540e-01,  3.37982178e-01, -8.18537712e-01, -5.93916953e-01,
        1.30555242e-01, -4.94256467e-01,  7.52102673e-01, -6.12779796e-01,
```

一个200维的词向量

Word2vec - 优点

避免产生维数灾难，词向量的维度大小是人为定义的，词向量的维度可以是50维，100维，200维等。

词义相近的词，词向量的距离也是相近的。

词向量的应用 - 文本分类

- 词向量作为输入层数据进行文本分类,步骤如下:
  - 加载语料数据;
  - 分词去除停用词;
  - 利用Word2vec训练词向量;
  - 词向量作为输入数据进行分类模型的训练。

目录

1. 文本分类的总体流程
2. 预处理介绍
3. 词向量介绍
4. **评估指标介绍**

二分类结果的混淆矩阵

- 评价分类器性能的指标一般是分类准确率（accuracy）。
- 对于二分类问题：
  - 常用的评价指标还有“查准率（precision）”与“查全率（recall）”。
- 二分类结果的混淆矩阵(confusion matrix)：
  - P：正元组
  - N：负元组
  - 真正例（true positive）
  - 假正例（false positive）
  - 真反例（true negative）
  - 假反例（false negative）

真实情况	预测结果		合计
	正例	反例	
正例	TP（真正例）	FN（假反例）	P
反例	FP（假正例）	TN（真反例）	N

思考题

1. （算一算）在二分类问题中，accuracy如何计算？请列出计算公式。

真实情况	预测结果		合计
	正例	反例	
正例	TP（真正例）	FN（假反例）	P
反例	FP（假正例）	TN（真反例）	N

性能度量之查准率、查全率与F1

- 查准率（precision）= $\frac{TP}{TP+FP}$
- 查全率（recall）= $\frac{TP}{TP+FN}$
- F1度量 = $\frac{2\,precision \times recall}{precision + recall}$

$$= \frac{2\ TP}{\text{样例总数} + TP - TN} = \frac{2\ TP}{2TP + \mathrm{FP} + \mathrm{FN}}$$

真实情况	预测结果		合计
	正例	反例	
正例	TP（真正例）	FN（假反例）	P
反例	FP（假正例）	TN（真反例）	N

思考题

1. （简答）在垃圾邮件分类问题和流感模型预测这两种场景情况下，precision和recall这两个指标哪一个更重要呢?

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

precision

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

recall

本章总结

- 本章主要介绍了文本分类任务中，普遍涉及的共性问题——语料预处理、文本向量化技术，以及有哪些文本分类任务的评估指标。


谢 谢
www.huawei.com