Microsoft

广东职业教育信息化研究会

GLOBAL AI BOOTCAMP

Orgonized by
GlobolAIBootCamp.com

MVP Microsoft® Most Valuable Professional

Powere by
MVP Meet Up China

广州站

AI BOOTCAMP CHINA

2018/12/15

Microsoft 广东职业教育信息化研究会 GLOBAL AI BOOTCAMP Orgonized by GlobolAIBootCamp.com MVP Microsoft Most Valuable Professional Powere by MVP Meet Up China

基于FastText进行文本分析

微服务知多少

主讲人：颜圣杰

MVP Microsoft Most Valuable Professional

广州 AI BOOTCAMP CHINA 2018/12/15

CONTENTS

广州 AI BOOTCAMP CHINA 2018/12/15

1. AI&&ML&&DL简介

简要介绍AI、ML和DL，并介绍他们的区别

2. NLP&&FastText

简要介绍NLP和FastText

3. FastText文本分类应用

介绍如何基于FastText进行文本分析

4. 应用场景

分析FastText适合处理哪一类问题，以及可以应用在哪些场景

01

AI&ML&DL简介

简要介绍：人工智能（Artificial Intelligence, AI）、机器学习(Machine Learning, ML) 和深度学习(Deep Learning, DL)。

AI&&ML&&DL

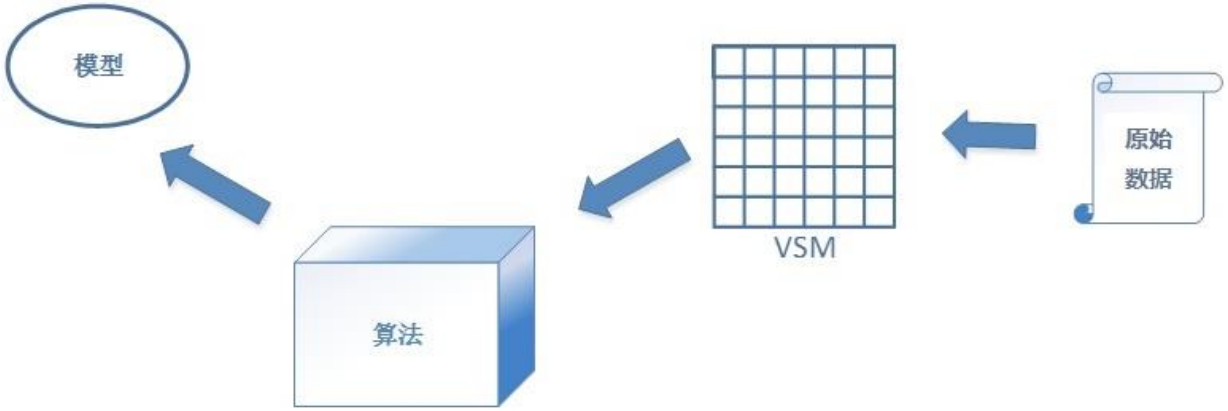
机器如何学习

让计算机程序（机器），不是通过人类直接指定的规则，而是通过自身运行，习得（学习）事物的规律和事物间的关联。

计算机能够处理的，只有数值和运算。

要让一段程序了解客观世界变化万千的事物，则必须将这些事物**数值化**，将事物的变化和不同事物之间的关联**转化为运算**。

机器学习三要素



模型 = 数据 + 算法

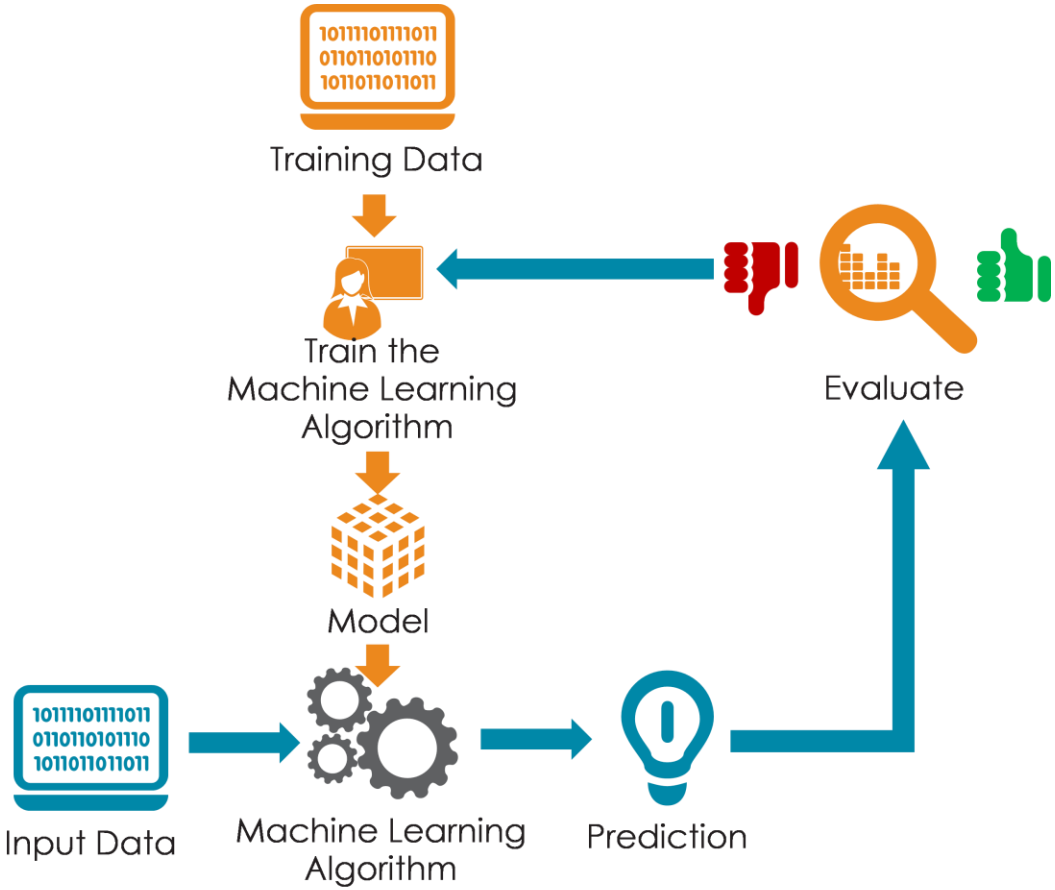
有监督学习

通过标注数据进行学习的方法，就叫做有监督学习或直接叫监督学习（Supervised Learning）。

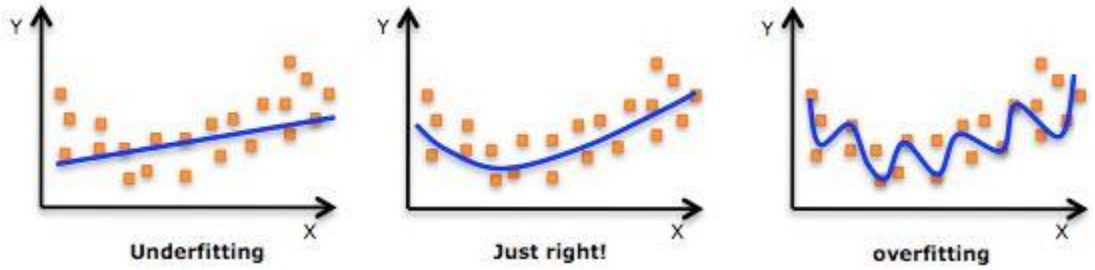
无监督学习

用于学习的数据只有样本，没有标签，这种通过无标注数据进行学习的方法，就叫做无监督学习（Unsupervised Learning）

获取模型的过程



模型的好坏



欠拟合：如果一个模型，在训练集上的预测结果就不佳，指标偏低

过拟合：如果在训练集上指标很好，而在验证/测试集上指标偏低（泛化能力差）

模型调优

数据

- 大量的高质量训练数据，是提高模型质量的最有效手段。

算法

- 超参数的调整

模型

- 对于不同的问题，更换不同的模型进行尝试

02

NLP

NLP (Natural Language Processing)，自然语言处理作为人工智能（AI）的一个子领域。其有很广的应用领域，比如文本分类、信息检索、机器翻译等等。而FastText在文本分类上十分高效

NLP的基础知识

NLP

用计算机对自然语言的形音义等信息进行处理，即对字词句篇章的 输入、输出、识别、分析、理解、生成等的操作和加工

自然语言理解

计算机能够理解自然语言文本的意义

自然语言生成

计算机能够以自然语言文本来表达给定的意图

NLP应用场景

机器翻译、文本摘要、文本分类、文本校对、信息抽取、语音合成、语音识别等

NLP基础技术

词法分析

词性标注和词义标注

句法分析

判断句子的句法结构和成分，明确各成分的相互关系

语义分析

根据句子的句法结构和句子中每个实词的词义推导出来能够反映这个句子意义的某种形式化表示

语用分析

人对语言的具体运用，是对自然语言的深层理解。

篇章分析

对段落和整篇文章进行理解和分析

NLP的难点

语法级

语法不完备；复杂多变甚至矛盾的规则

语义级

模糊；多指；多义；歧义；

语用级

语境认知；上下文关联与整理；深度信息挖掘；信息补充与重构

交互级

修辞手法；言外之意；以一反三；

研究方向

应用方向	应用目标	重要性/成熟度	重要性原因
文本纠错	纠正处理一些文本错误	中/低	对话机器人通常的输入错误很容易被发现和更改。语音类的输入错误通常会被语音识别模块纠正。只有少数专有词需要额外纠正。
中文分词	将中文文本按空格分割成可分析的词汇单元型文本	高/高	中文处理必须有的步骤。
词性标注	标注每个词的词性	中/高	词性标注可以用于辅助文本分类、事件抽取等任务
命名实体识别	识别文本中指定的特殊类别文本	中/高	实体识别有助于提取专有类别词，对事件抽取也有帮助
句法分析	分析句子的形式语言结构或者依存关系	高/高	分析整体句法结构，后续处理严重依赖于此分析的准确性
文本分类	将文本根据其意义或作用进行分类	高/中	可以视为由上述方法辅助需要完成的核心任务，这决定了对话机器人对整个问题的判断
事件抽取	提取一个能够表达的事件要素	高/低	需要将对话人类方的语言信息重分有效地提取
时序分析	分析处理时间标志	低/高	在长对话过程中需要提取正确的时序
指代消解	完成正确的本体映射	中/高	长对话中存在多次出现的指代需要进行有效
模板填充	完成自然语言到某类模型参数的准确转化	中/低	将重要信息重整为可用的参数填入指定目标模板。对非指令型的系统意义不大。
回答生成	根据持有信息自动生成有效回答	中/低	需要应对数据库外回答时较为有用。对非开放域系统意义不大

FastText简介

Library for efficient text classification and representation learning

FastText是一个开源，免费，轻量级的库，用于文本分类和词向量学习。其可在标准的通用硬件上运行，产生的模型可压缩以适应移动设备。

https://fasttext.cc/

FastText的特点

Model	AG	Sogou	DBP	Yelp P.	Yelp F.	Yah. A.	Amz. F.	Amz. P.
BoW (Zhang et al., 2015)	88.8	92.9	96.6	92.2	58.0	68.9	54.6	90.4
ngrams (Zhang et al., 2015)	92.0	97.1	98.6	95.6	56.3	68.5	54.3	92.0
ngrams TFIDF (Zhang et al., 2015)	92.4	97.2	98.7	95.4	54.8	68.5	52.4	91.5
char-CNN (Zhang and LeCun, 2015)	87.2	95.1	98.3	94.7	62.0	71.2	59.5	94.5
VDCNN (Conneau et al., 2016)	91.3	96.8	98.7	95.7	64.7	73.4	63.0	95.7
`fastText`, $h = 10$	91.5	93.9	98.1	93.8	60.4	72.0	55.8	91.2
`fastText`, $h = 10$, bigram	92.5	96.8	98.6	95.7	63.9	72.3	60.2	94.6

Table 1: Test accuracy [%] on sentiment datasets. `FastText` has been run with the same parameters for all the datasets. It has 10 hidden units and we evaluate it with and without bigrams. For VDCNN and char-CNN, we show the best reported numbers without data augmentation.

	Zhang and LeCun (2015)		Conneau et al. (2016)			`fastText`
	small char-CNN*	big char-CNN*	depth=9	depth=17	depth=29	$h = 10$, bigram
AG	1h	3h	8h	12h20	17h	3s
Sogou	-	-	8h30	13h40	18h40	36s
DBpedia	2h	5h	9h	14h50	20h	8s
Yelp P.	-	-	9h20	14h30	23h00	15s
Yelp F.	-	-	9h40	15h	1d	18s
Yah. A.	8h	1d	20h	1d7h	1d17h	27s
Amz. F.	2d	5d	2d7h	3d15h	5d20h	33s
Amz. P.	2d	5d	2d7h	3d16h	5d20h	52s

Table 2: Training time on sentiment analysis datasets compared to char-CNN and VDCNN. We report the overall training time, except for char-CNN where we report the time per epoch. * Training time for a single epoch.

- 专注于文本分类
- 支持多语言表达
- 适合大型数据，训练速度快

原理

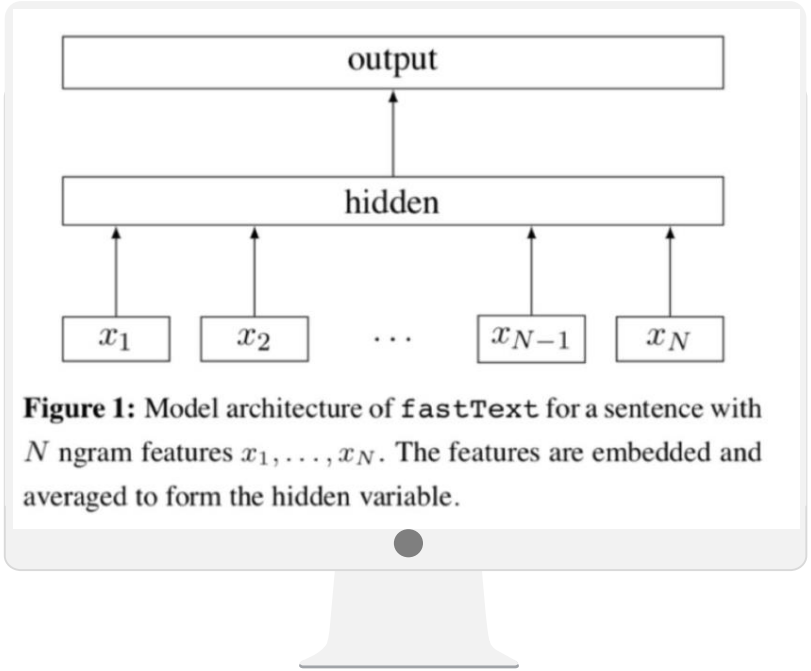


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

输入

输入一个词的序列（一段文本或者一句话)。序列中的词和词组组成特征向量。

映射到中间层

特征向量通过线性变换映射到中间层。

输出

输出这个词序列属于不同类别的概率。

API简介

```
>> ./fasttext
usage: fasttext <command> <args>
The commands supported by fasttext are:
  supervised              train a supervised classifier
  quantize                quantize a model to reduce the memory usage
  test                    evaluate a supervised classifier
  predict                 predict most likely labels
  predict-prob            predict most likely labels with probabilities
  skipgram                train a skipgram model
  cbow                    train a cbow model
  print-word-vectors      print word vectors given a trained model
  print-sentence-vectors  print sentence vectors given a trained model
  nn                      query for nearest neighbors
  analogies               query for analogies
```

文本分类
有监督学习

词向量学习
无监督学习

03

FastText文本分类

通过有监督学习介绍如何使用FastText进行文本分类

使用fasttext进行文本分析的步骤

分析步骤

01 **语料获取**

已有语料；网上下载、爬取；

02 **语料预处理**

语料清洗、脱敏、去重、分词、去停用词等

03 **特征工程**

打标签

04 **模型训练**

根据不同的模型，选择不同的训练方法

05 **模型调优**

从数据、算法和模型角度进行评估调优

06 **上线应用**

部署应用

语料预处理：分词

未来，NLP 将跟其他人工智能技术一道深刻地改变人类的生活。当然前途光明、道路曲折是亘古不变的道理，为了实现这个美好的未来，我们需要大胆创新、严谨求实、扎实进取。讲求研究和应用并举，普及与提高同步。我们期待着与业界同仁一道努力，共同走进 NLP 下一个辉煌的十年。

未来 ， NLP 将 跟 其他 人工智能 技术 一道 深刻 地 改变 人类 的 生活 。 当然 前途 光明 、 道路 曲折 是 亘古 不变 的 道理 ， 为了 实现 这个 美好 的 未来 ， 我们 需要 大胆 创新 、 严谨 求实 、 扎实 进取 。 讲求 研究 和 应用 并举 ， 普及 与 提高 同步 。 我们 期待 着 与 业界 同仁 一道 努力 ， 共同 走进 NLP 下 一个 辉煌 的 十年 。

语料预处理：去停用词

未来， NLP 将 跟 其他 人工智能 技术 一道 深刻 地 改变 人类 的 生活。 当然 前途 光明、 道路 曲折 是 亘古 不变 的 道理， 为了 实现 这个 美好 的 未来， 我们 需要 大胆 创新、 严谨 求实、 扎实 进取。 讲求 研究 和 应用 并举， 普及 与 提高 同步。 我们 期待 着 与 业界 同仁 一道 努力， 共同 走进 NLP 下 一个 辉煌 的 十年。

未来， NLP 人工智能 技术 一道 深刻 改变 人类 生活。 前途 光明、 道路 曲折 亘古 不变 道理， 实现 美好 未来， 需要 大胆 创新、 严谨 求实、 扎实 进取。 讲求 研究 应用 并举， 普及 提高 同步。 期待 业界 同仁 一道 努力， 共同 走进 NLP 一个 辉煌 十年。

特征工程：打标签

未来， NLP 人工智能 技术 一道 深刻 改变 人类 生活 。 前途 光明 、 道路 曲折 亘古 不变 道理， 实现 美好 未来， 需要 大胆 创新 、 严谨 求实 、 扎实 进取 。 讲求 研究 应用 并举， 普及 提高 同步 。 期待 业界 同仁 一道 努力， 共同 走进 NLP 一个 辉煌 十年 。 __label__science

训练

```
import fastText
# 使用默认参数训练模型
classifier = fastText.train_supervised(input="corpus/news_fasttext_train.txt")

result = classifier.test("corpus/news_fasttext_train.txt")#测试模型质量
print(result)
classifier.quantize()#模型压缩
classifier.save_model("news.model.bin")#保存模型
```

```
Read 65M words
Number of words:  603048
Number of labels: 13
Progress: 100.0% words/sec/thread:  265041 lr:  0.000000 loss:  0.414554 ETA:   0h 0m
(121155, 0.9446494160373076, 0.9446494160373076)
```

指标评价

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)

准确率，缩写表示用 P。准确率是针对我们预测结果而言的，它表示的是预测为正的样例中有多少是真正的正样例。定义公式如下：

$$P = \frac{TP}{TP+FP}$$

召回率，缩写表示用 R。召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确。定义公式如下：

$$R = \frac{TP}{TP+FN}$$

F1 衡量，表达出对查准率/查全率的不同偏好。定义公式如下：

$$F1 = \frac{2 * P * R}{P + R}$$

模型调优

```
fastText.train_supervised(input, lr=0.1, dim=100, ws=5,
epoch=5, minCount=1, minCountLabel=0, minn=0, maxn=0, neg=5,
wordNgrams=1,loss='softmax', bucket=2000000, thread=3, lrUpdateRate=100,
t=0.0001, label='__label__', verbose=2, pretrainedVectors='') # 参数列表
```

```
# 使用预训练的词向量，使用ngrams=2，lr=0.5，训练50轮
classifier = fastText.train_supervised(
input="corpus/news_fasttext_train.txt", dim=300, wordNgrams=2, lr=0.5,
epoch=50, pretrainedVectors="cc.zh.300.vec")
```

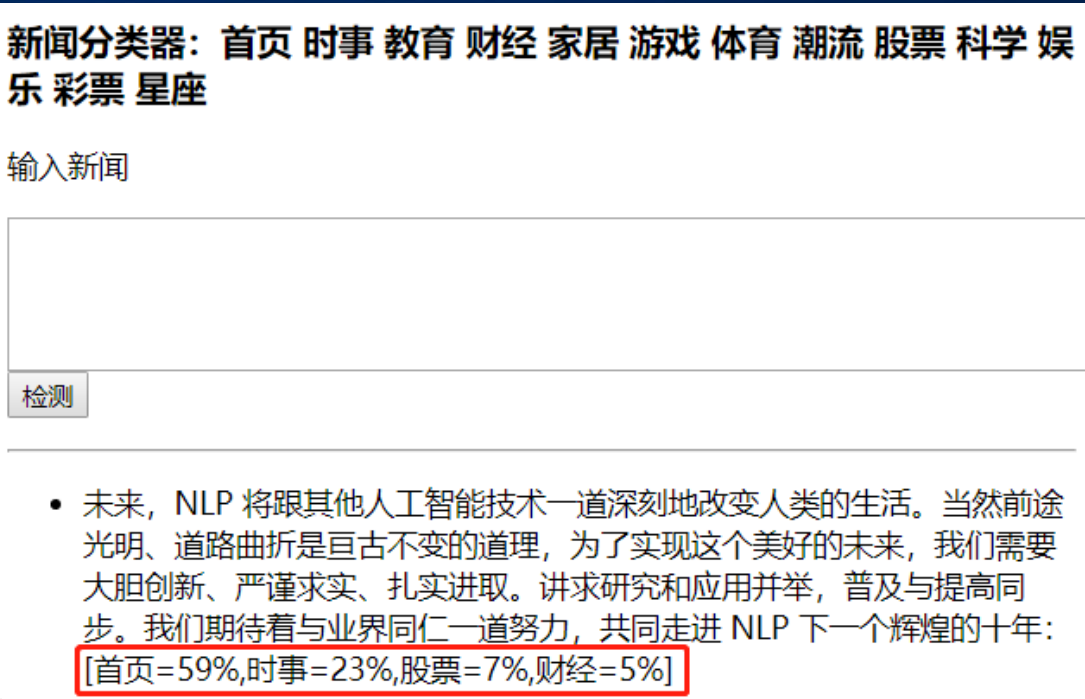
```
Read 65M words
Number of words:  603048
Number of labels: 13Progress: 100.0% words/sec/thread:   85034 lr:  0.000000 loss:  0.054984 ETA:   0h 0m
(121155, 0.9988362015599851, 0.9988362015599851)
```

预测

```
import fastText
import jieba

classifier = fastText.load_model("news.model.bin")
news = "内地与香港签署货贸协议：明年起原产香港的货物进口内地将零关税 "
cutNews = jieba.cut(news) # 进行分词处理
testNews = removeStopwords(cutNews) # 进行分词处理
result = classifier.predict(“ ”.join(testNews)，k=4)
print(result)
```

Demo展示



https://github.com/sheng-jie/fasttext.demo

词向量学习（无监督学习）

```
def train_unsupervised(input, model="skipgram", lr=0.05, dim=100,
ws=5, epoch=5, minCount=5, minCountLabel=0, minn=3, maxn=6, neg=5,
wordNgrams=1, loss="ns", bucket=2000000,
thread=multiprocessing.cpu_count() -1, lrUpdateRate=100, t=1e-4,
label="__label__", verbose=2, pretrainedVectors="")
```

```
classifier.get_word_vector("中国")
array([ 5.47645316e-02, -2.96028946e-02,  3.31672430e-01, -2.25172248e-02,\n       -1.69864759
  [0:300] : [0.05476453, -0.029602895, 0.33167243, -0.022517225, -0.16986476, 0.20184939, -0.1521955,
  dtype: dtype('float32')
  max: 0.94604397
  min: -0.8332312
  shape: (300,)
  size: 300
  __internals__: {'T': array([ 5.47645316e-...e=float32), 'base': None, 'ctypes': <numpy.core._intern
```

小结

语料获取

数据质量决定了模型的质量，语料库要平均

数据预处理

分词、去停用词，剔除语料噪声

打标签

训练数据和测试数据中的标签要打散，不要顺序排列

模型调优

超参数的调整，借助预训练词向量库提升模型性能

FastText适用于分类类别非常大而且数据集足够多的情况

当分类类别比较小或者数据集比较少的话，很容易过拟合

04 应用场景

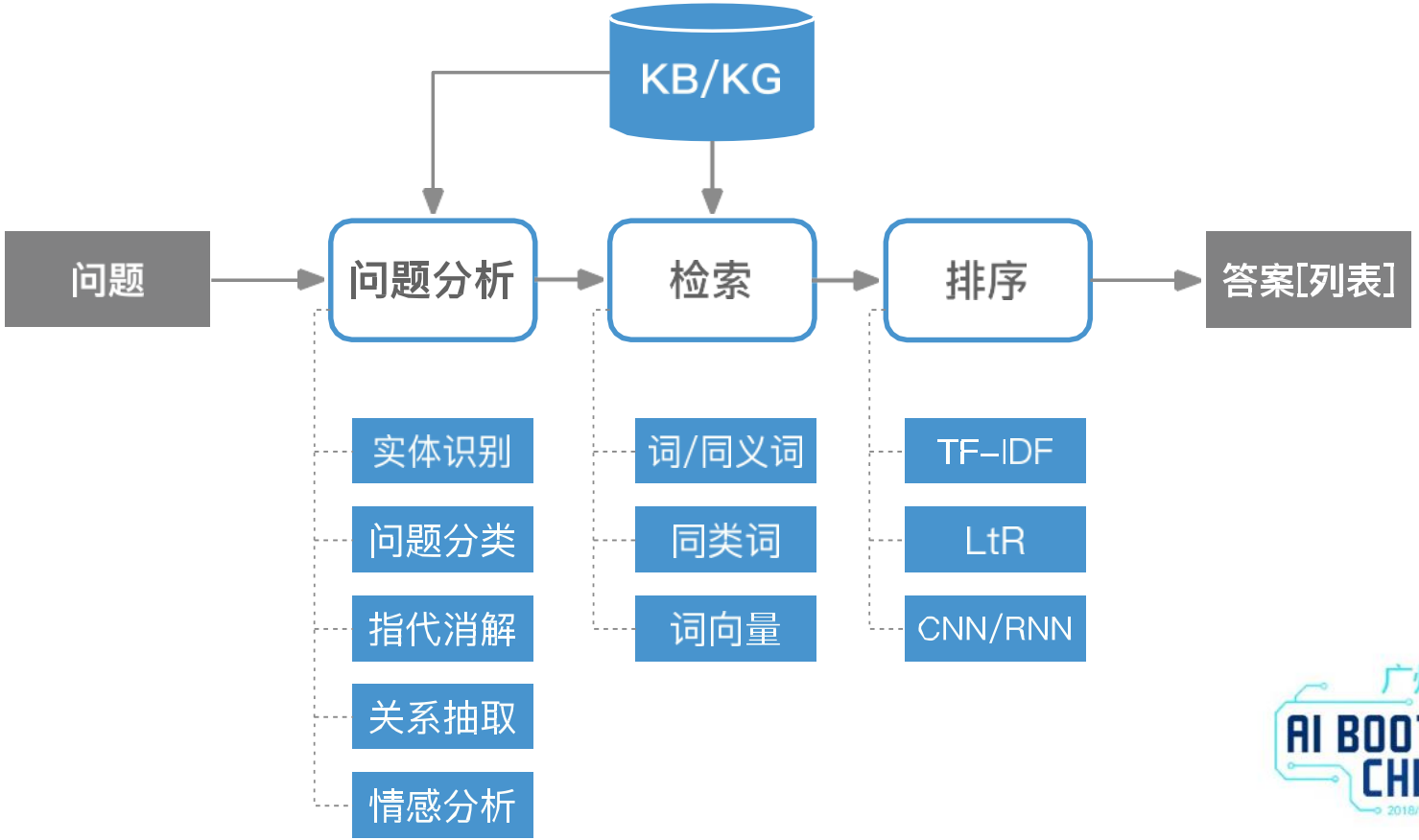
分析FastText适合处理哪一类问题，以及可以应用在哪些场景

新闻的分类和推荐

快递下单地址智能填充

智能机器人

问题检索系统



其他应用场景

垃圾邮件分类

对邮件文本进行识别分类

文章识别

对朋友圈的文章进行学习，包括广告，活动，招聘和公告

智能客服

基于现有知识库，构建企业智能客服.

其他场景

涉及到文本分类和词向量学习的，都可以考虑使用fasttext

思考题：闲聊检测

某公司老总，发现员工上班期间经常使用微信闲聊，刷朋友圈，工作效率低下。为了提升工作效率，把控工作质量，选择了某企业聊天应用解决这一问题。虽然将聊天工具从微信切换到企业聊天应用上，但依旧不能解决工作时间的闲聊问题。现希望能通过检测聊天内容，进行闲聊检测。当检测到闲聊时，予以提示，以督促员工专注工作沟通。

Thank you!

分享者：颜圣杰

Microsoft

广东职业教育信息化研究会

GLOBAL AI BOOTCAMP

Orgonized by
GlobolAIBootCamp.com

MVP Microsoft® Most Valuable Professional

Powere by
MVP Meet Up China

广州站

AI BOOTCAMP CHINA

2018/12/15