自然语言处理的定义

阿里云

- 自然语言：指人类使用的在社会生活中自然形成的语言；
- 自然语言处理：指计算机识别、理解、计算分析、生成自然语言的过程。
- 包含**自然语言理解**和**自然语言生成**两部分的两大研究方向。

自然语言理解

所有支持机器理解文本内容的方法模型或任务的总称，是推荐、问答、搜索等系统的必备模块。

自然语言生成

将非语言格式的数据转换成人类可以理解的语言格式，是翻译、写作等系统的必备模块。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

自然语言处理的发展趋势

阿里云

01 智能人机交互

- 不同语言、不同领域下的人机交互提升;
- 多语言交互从不同语言理解上升到不同文化的理解。

02 多模态融合

- 视频、图像、文本、语音等模态的全面融合;
- 在对话系统产品中应用效果显著。

03 解决方案建设

- 每种场景领域都有特定的需求及其相应的场景数据;
- 模型结合场景数据进行训练能够更好地满足场景需求。

04 形成行业知识库

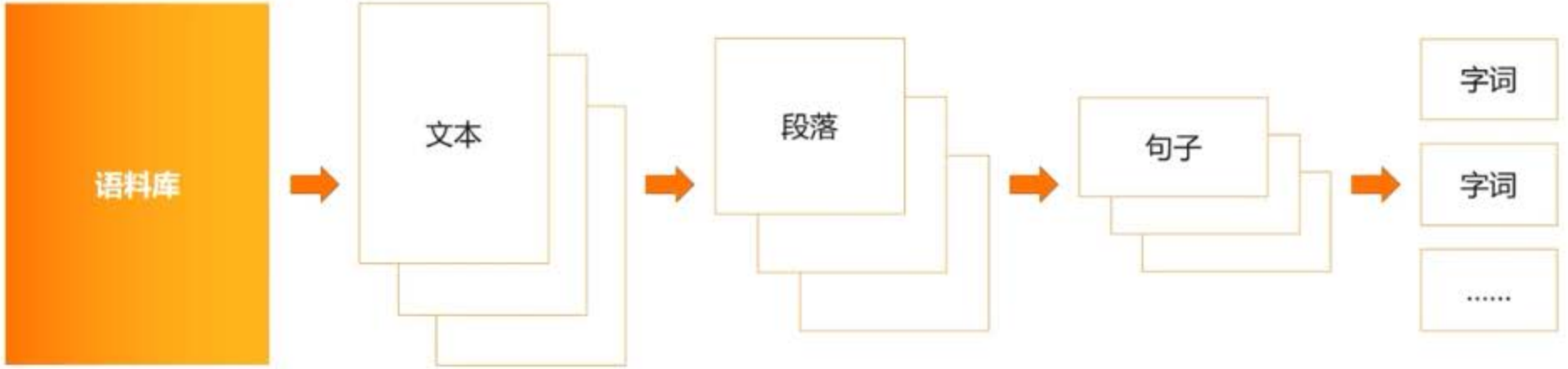
- 结合已有的知识和技术提高非结构化数据理解能力;
- 基于过去已知知识进行推理，理解行业事件知识。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

自然语言处理的语料库

阿里云

- 大量已知语料数据的集合，在自然语言处理模型的定型过程中起到基准的作用；
- 在语言的实际使用中真实出现过的语言材料，通常经过整理，**具有既定格式与标记**；
- 若要获得最小单位的字词，需要由外而内一层多层剥开。



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

常见的语料库

阿里云

中文语料库

国家语委现代汉语语料库

- 提供在线检索，约1亿字符，标注语料5千万。

古代汉语语料库

- 提供了分词、词性标注软件、字频统计等软件。

分词库

- 包含非常多的各行业词汇。

英文语料库

布朗语料库

- 第一个在计算语言学处理中使用的通用英语语料库；
- 代表通用英语样本，采样自小说，新闻和宗教文本。

停用词库

- 包含了 来自 11 种不同语言（包括英语）的 2400 个停用词。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

现代汉语与英语的区别

【主要区别】

- 汉语中，名词、动词、形容词的语法功能是相互交错的；
- 一类词往往充当多种句子成分，一种成分往往也不是某一类词专有的。

【举例】

- 名词作谓语：八月一日建军节。
- 名词作状语：我明天下午很忙。

现代汉语的一些特点

阿里云

语音

没有辅助音

- 音节开头或结尾，没有两个或三个辅音连在一起的现象；
- 举例：next spring。

元音占优势

- 音节中可以没有辅音，但不能没有元音；
- 举例：啊，噢。

词汇

构词法简单

- 常用词根复合法构词；
- 举例：国家，雪白。

双音节词占优势

- 现代汉语中，频率最高的1万个词中，单音节词占24%，双音节词占63%；
- 举例：桌子，冬天，教师。

语法

常用语序和虚词

- 语序和虚词是表达语法意义的主要手段；
- 举例：我和老师，我的老师。

结构简单

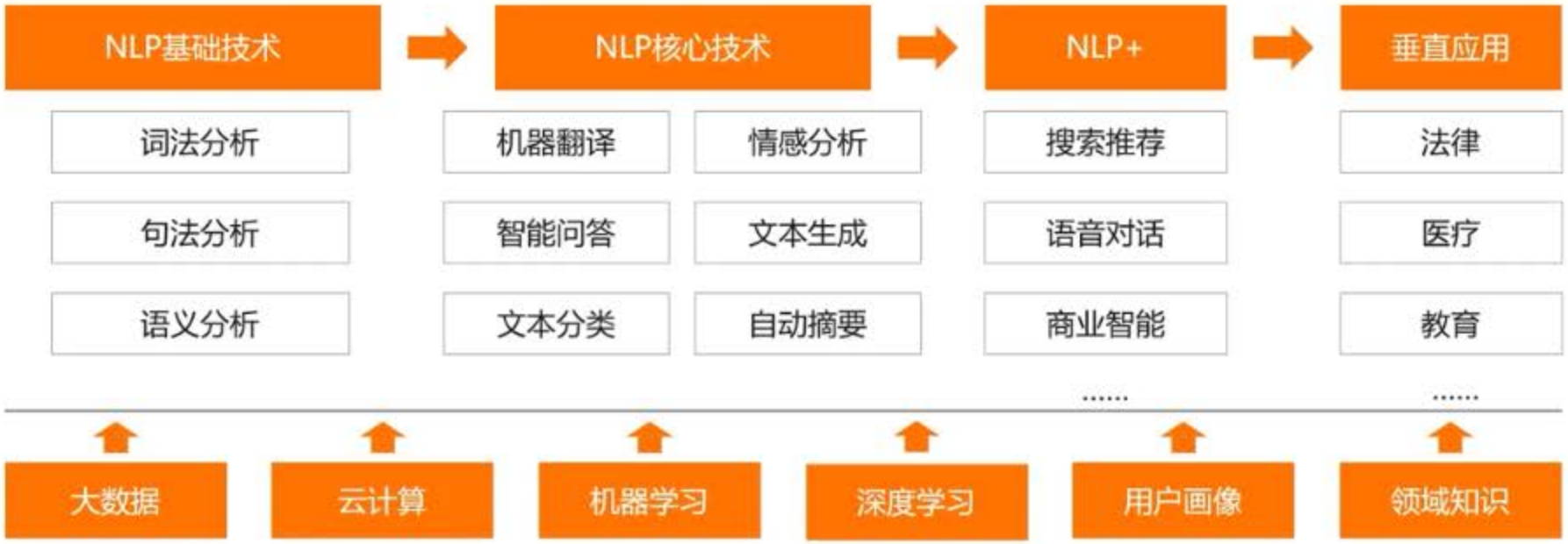
- 词和句子的结构原则基本一致。

词类丰富

- 除名词、动词、数词外，还有语气词、量词等词汇。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

自然语言处理的技术体系



自然语言处理的基础技术

阿里云

词法分析

主要作用

将句子、段落、文章这种长文本分解为以字词为单位的数据结构，方便后续的处理分析工作。

所需知识

- 分词；
- 词性标注。

句法分析

主要作用

识别句子所包含的字词组合以及这些组合之间的关系，构成句法结构。

所需知识

- 关键词提取；
- 命名实体识别；
- 语法分析。

语义分析

主要作用

分析语法结构的同时，分析单词、词组、句子、段落所包含的意义。

所需知识

- 文本向量化。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

分词的概念

阿里云

- 将句子、段落、文章等长文本分解为以字词为单位的数据结构；
- 常见的方法包括最大匹配分词算法和最短路径分词算法。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

分词的难点

- **界定中文词汇**：
  - 分词工具效果的评判标准；
- **分词歧义问题**：
  - 交集型切分歧义；
  - 组合型切分歧义；
- **分词切分粒度问题**：
  - 粗粒度切分；
  - 细粒度切分；
  - 搜索切分；
- **未登录词问题**。

分词歧义示例

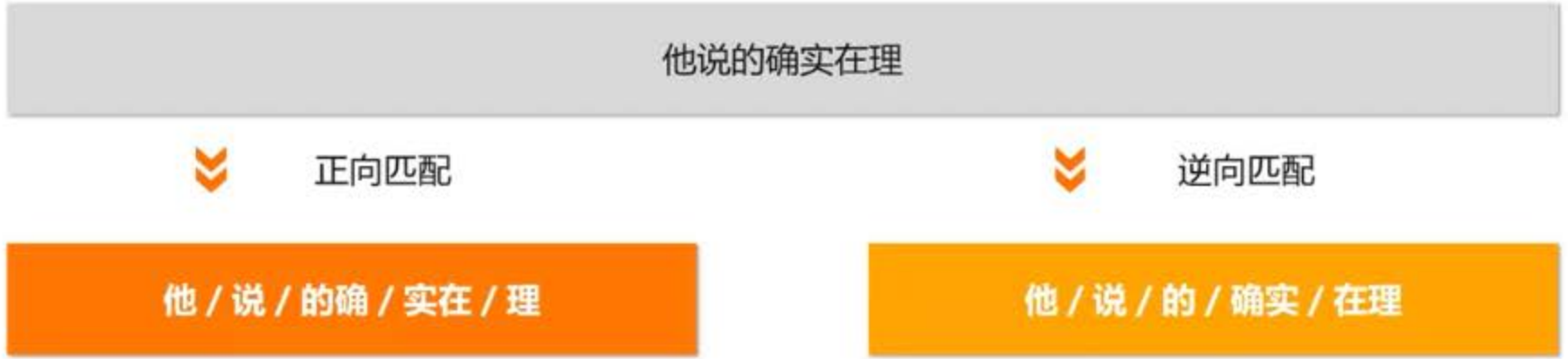
文本	错误分词	正确分词	错误类型
南京市长江大桥	南京/市长/江大桥	南京市/长江大桥	交集型切分歧义
研究生命的起源	研究生/命/的/起源	研究/生命/的/起源	交集型切分歧义
结婚的和尚未结婚的	结婚/的/和尚/未/结婚/的	结婚/的/和/尚未/结婚/的	交集型切分歧义
化妆和服装	化妆/和服/装	化妆/和/服装	交集型切分歧义
从马上跳下来	从/马上/跳/下来	从/马/上/跳/下来	组合型歧义切分

分词的实现方法（1/2）

阿里云

最大匹配分词算法：

- 以词典为依据，取词典中最长词长度作为第一次取字数量的长度；
- 在词典中进行匹配，然后逐字递减，在对应的词典中进行查找；
- 根据匹配的方向不同，分为**正向匹配**和**逆向匹配。**



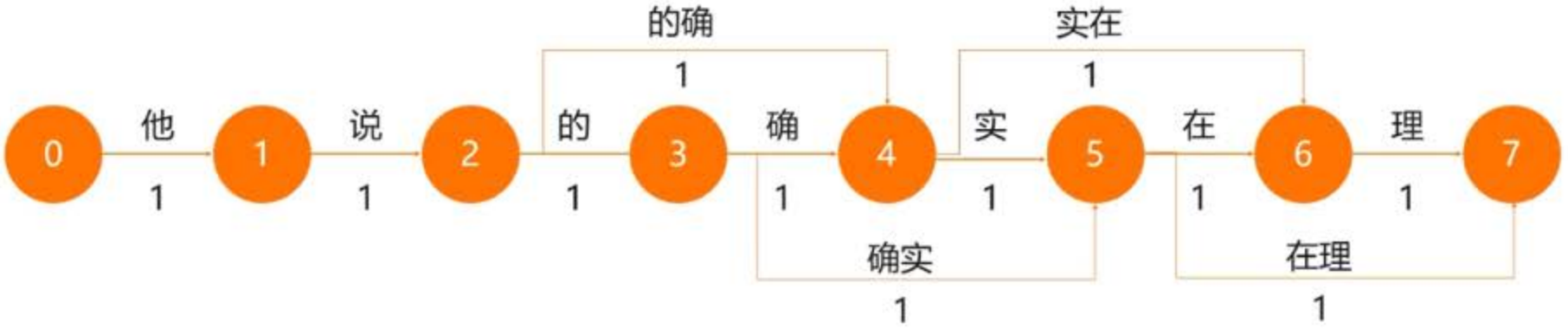
阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

分词的实现方法（2/2）

阿里云

最短路径分词算法：

- 首先将文本中的所有词匹配出来，构成词图，寻找从起始点到终点的最短路径；
- 词图中每个词的权重都是相等的，因此每条边的权重都为1；
- 两点之间的最短路径也包含了路径上其他顶点间的最短路径。



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

词性标注的概念

阿里云

- 词性是指词的语法分类，又称词类；
- 词性标注是在给定句子中判断每个词的语法范畴，确定其词性并加以标注的过程；
- 中文词性分类：名词、动词、形容词、副词、代词、介词、连词、数词、量词、助词、感叹词、拟声词；

自然语言处理是有趣的科学。

自然语言处理[名词] 是[动词] 有趣[形容词] 的[助词] 科学[名词] 。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

词性标注的标注规范

阿里云

- 词性标注需要有一定的标注规范，如先将词分为名词、动词、形容词等，然后用 “n”、“v”、“adj”等来进行表示。

自然语言处理是有趣的科学。

自然语言处理[n]是[v]有趣[a]的[d]科学[n]。

标记	词性	标记	词性
a	形容词	d	副词
n	名词	p	介词
m	数词	q	量词
c	连词	v	动词

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

关键词提取的概念

阿里云

- 关键词即文本中一些“重要的”词，通过这些重要的词可以理解文本中心思想；
- 关键词提取的质量，质量体现在关键词提取的准确性、全面性和代表性；
- 关键词提取的评价指标为词的权重。

ID	原始文本	关键词
1	裤子的质量好，穿起来很好看，非常满意	裤子，质量，好看，满意
2	这本书无论是质量还是内容都非常棒，内容全面，介绍顺序循序渐进，图文并茂，可以作为高中生的入门课程	质量，内容，棒，高中生，入门课程

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

关键词提取的实现方法（1/2）

阿里云

- 关键词提取的实现包括两个步骤，第一步是获取文本的候选词，第二步则是对候选词进行打分；
- 输出的关键词是候选词中得分比较高的。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

关键词提取的实现方法（2/2）

阿里云

- 关键词提取算法一般分为**有监督**和**无监督**两类。

有监督的关键词提取

- 首先构建一个的词表，然后判断文档与词表中词的匹配程度，以类似打标签的方式，达到关键词提取的效果。

无监督的关键词提取

- 对数据的要求低，既不需要一张人工生成，维护的词表，也不需要人工标注语料辅助训练。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

命名实体识别的概念

阿里云

- 识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等；
- 通常把对这些词的识别在词汇形态处理任务中独立处理。

小明将参加阿里巴巴主办的天池大数据竞赛

实体-人名　　实体-机构名称　　实体-活动名称

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

命名实体识别的标注方法

阿里云

- **三大类**：实体类、时间类和数字类；
- **七小类**：人名、机构名、地名、时间、日期、货币和百分比；
- 常用BIOES-四位序列标注法；

我 [O] 是 [O] 一 [S] 名 [O] 大 [B] 学 [I] 生 [E] , [O] 我 [O] 爱 [O] 中 [B] 国 [B] 。 [O]

类型	说明
B	Begin，代表实体片段的开始
I	Intermediate，代表实体片段的中间
O	Other，代表字符不为任何实体
E	End，代表实体片段的结束
S	Single，代表实体片段为单个字

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

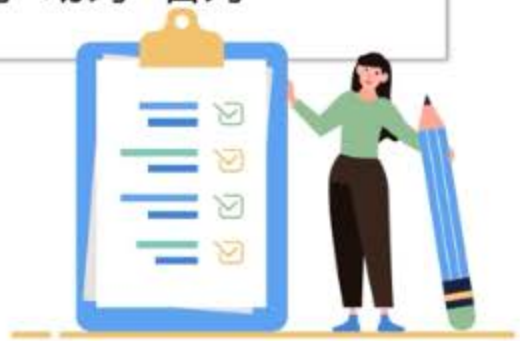
语法分析的概念

阿里云

- 判断输出的字符串是否属于某种语言；
- 消除输入句子中词法和结构等方面的歧义；
- 分析输入句子的内部结构，如成分构成、上下文关系。

外面摆着花。
外面演着戏。

- **词序列相同**：处所短语+动词+助词+名词



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

语法分析的难点

阿里云

语法分析主要有以下两个障碍：歧义和搜索空间。

歧义

1.山上的水宝贵,我们把它留给晚上来的人喝；

2.这个人连小明都不认识。

.....

搜索空间

名词	花卉	老虎	.....
动词	踢	跑	.....
形容词	快乐	温暖	.....

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

语法分析的实现方法

- 基于规则的方法是语法分析中的常用方法；
- 以“小明在快乐地学习”为例：

基于规则的方法

- 由人工来组织语法规则，建立语法知识库；
- 通过条件约束和检查来**实现语法结构歧义的消除**；
- 能够较好的**处理句子歧义和超语法现象**；
- 会存在语法规则覆盖有限、系统可迁移等缺陷。

文本向量化的概念

- 自然语言处理前，需要**将文本表示成计算机可识别的数值形式**；
- 一个语言模型来构建关于输入和输出之间的映射关系；
- 离散式词向量和分布式词向量是文本向量化中的常用方法。

文本向量化的实现方法（1/2）

阿里云

离散式词向量：

- 常用**One-Hot 编码，**每一个词特征都被表示成一个很长的向量；
- 其长度等于词表大小，当前词对应位置为1，其他位置为0；
- 无法衡量不同词之间的相似关系，无法突出词之间重要性的区别。

我喜欢学习NLP

我 | 喜欢 | 学习 | NLP

	我	喜欢	学习	NLP
我	1	0	0	0
喜欢	0	1	0	0
学习	0	0	1	0
NLP	0	0	0	1

文本向量化的实现方法（2/2）

阿里云

分布式词向量：

- 将词转化成一种分布式表示，即将词表示成一个定长的连续的稠密向量。

离散式词向量				
我	1	0	0	0
喜欢	0	1	0	0
学习	0	0	1	0
NLP	0	0	0	1

- 数据过于稀疏，难以捕捉文本含义；
- 不能表示单词相似度，数据信息量较少。

分布式词向量				
我	0.1	0.2	0.4	0.2
喜欢	0.2	0.3	0.7	0.1
学习	0.5	0.9	0.1	0.3
NLP	0.2	0.3	0.6	0.2

- 词之间**存在“距离”的概念，可表示相关性**；
- 词向量能够**包含更多信息**。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本分类技术与实现方法介绍

阿里云

- 能够对文本按照一定的分类标准进行自动分类标记；
- 机器自动化标注的文本数据具有一致性、高质量等特点；
- 利用待分类数据的特征与类别进行匹配，选择最优的匹配结果作为分类结果。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本分类技术的应用场景

阿里云

文本分类从给定的标签集合中自动地给文本打标签，应用广泛。

邮件属性分类

1.判断是否为垃圾邮件；
2.检测邮件类别。

广告内容审核

1.判断是否为广告；
2.检测是否存在灌水评论；
3.判断是否违规内容。

新闻分类推送

1.判断新闻类别；
2.按类别推送新闻。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

情感分析技术与实现方法介绍

阿里云

- 互联网用户生产的带有主观性的文本有助于制定决策;
- 对带有情感色彩的主观性文本进行分析、处理、归纳和推理;
- **实现方法:**
  - 由预标记词汇组成的字典，使用词法分析器将输入文本转换为单词序列;
  - 将每一个新的单词与字典中的词汇进行匹配，根据匹配结果提高或降低文本得分。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

情感分析技术的应用场景

阿里云

目前，情感分析在实际生产场景中得到越来越多的应用。

电子商务

判断产品反馈的情感倾向进行产品优化。

舆情分析

感知舆情情感倾向以维护公司品牌。

市场分析

分析产品与服务市场反响，制定营销策略。

用户维护

提取用户意见确认产品功能需求。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本纠错技术介绍

阿里云

- 文本纠错就是将文本中有错误的地方进行纠正;
- 错误类型包含错别字、缺失字、冗余字、词语搭配错误和语法错误等;
- 能够评估和权衡相关因素，比人类更快、更准确地识别。

自然语言处理技术是人工只能的一个重要分支。

自然语言处理技术是人工智能的一个重要分支。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本纠错技术应用的实现方法

阿里云

- 文本纠错通常包含两个步骤，第一步是**错误检测**，第二步是**错误纠正**；
- 从字粒度和词粒度两方面来检测文本错误；遍历所有的疑似错误位置，使用音似、形似等相关字词替换错误位置的字词。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本纠错技术的应用场景

阿里云

文本纠错是针对文本拼写错误进行检测与纠正的一项工作，具有丰富的应用场景。

写作辅助

写作时自动检查并提示错别字情况。

公文纠错

提供字词、标点、专名、数值内容纠错。

搜索纠错

自动纠正搜索查询并提示用户。

对话纠错

自动修正语音识别转文本过程中的错别字。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

问答系统技术介绍

阿里云

- 一个能回答任意自然语言形式问题的自动化系统；
- 能够对于一个指定问题，能够得到简短、精确的答案。

1.分析问题 → 如何去分析问题

2.检索答案 → 如何根据分析结果缩小答案存在的范围

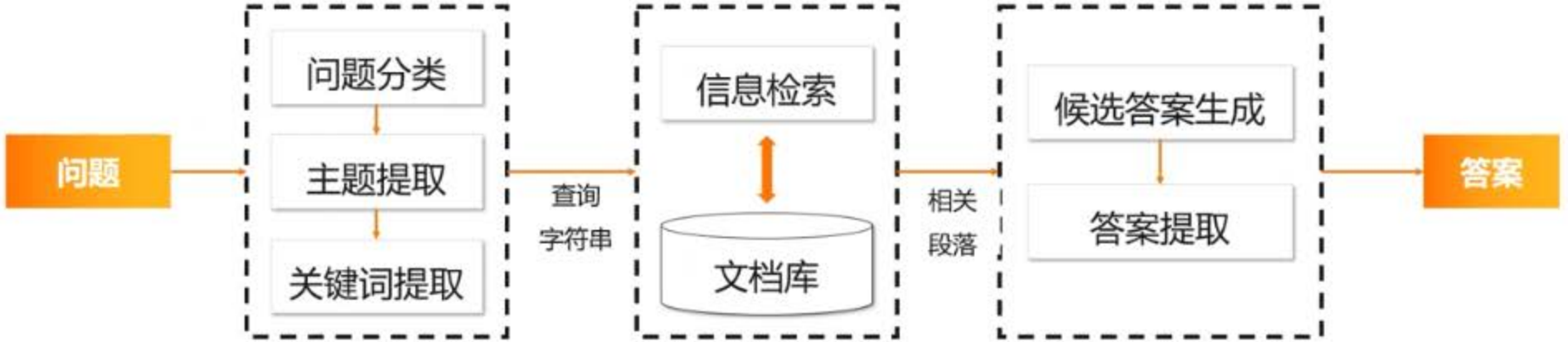
3.提取答案 → 如何从可能存在答案的信息块中抽取答案

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

问答系统技术的实现方法

阿里云

- 问答系统通常采用基于自由文本的方式实现；
- 属于开放域问答系统，能够回答一些答案存在于文档集合中的问题。



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

问答系统技术的应用场景

阿里云

智能客服

以用户需求为导向的智能问答已经逐步形成一种完善的信息服务系统，满足客户的服务需求。

快速检索

基于信息检索的问题回答，目的是通过在网页或文档集合中查找短文本段来满足用户的即问即答的需求。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本标签生成技术与实现方法介绍

阿里云

定义

- 生成的标签在一定程度上能够体现文本内涵；
- 是文本检索、文档比较、摘要生成、文档分类和聚类等文本挖掘研究的基础性工作；

实现方法

- 采用计算权重的方式从候选集合中得到文本标签；
- 主要包括词性、词频、逆向文档频率、相对词频、词长等。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本标签生成技术的应用场景

个性化推荐

通过对文章的标签计算，结合用户画像，精准的对用户进行个性化推荐。

主题聚合

根据生成的文本标签，聚合相同主题标签的文本，让用户能及时查阅同一主题下的所有的文本。

文本摘要生成技术与实现方法介绍

阿里云

- 定义：自动生成含原文本中重要信息的新文本内容；
- 目标：通过机器自动输出简洁、流畅、保留关键信息的摘要。

阿里云河北省教育脱贫云计算初级认证工程师培训是阿里云基于与河北教育厅签署的关于河北省贫困大学生云计算人才培养合作协议，为河北省建档立卡的贫困大学生提供免费的云计算培训及初级工程师认证考试。该活动获得了河北省教育厅及众多高校的大力支持……

培养一个人，帮助一个家，云计算技术赋能梦想与未来

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

文本摘要生成的应用场景

阿里云

文本摘要生成有非常多的应用场景，如自动报告生成、新闻标题生成、搜索结果预览等。

自动报告生成

为用户提供个性化定制生成报告服务。

新闻标题生成

快速抽取核心内容摘要，节约阅读时间成本。

搜索结果预览

提供每条搜索结果的内容预览缩图，帮助用户更快地找到所需的信息。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

智能创作技术介绍

阿里云

- 智能创作技术可以分为人工智能自动写作和人工智能辅助写作两类；
- 具有作品制作高效、具有强大潜能、内容客观、节省人力成本等。

人工智能自动写作

人工智能算法自主完成写作任务，不需要人工干预。

人工智能辅助写作

在人类写作的全流程中提供辅助功能，帮助完成写作任务。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

智能创作技术的应用场景

阿里云

智能写诗

根据关键词和格律自动生成诗词，具有多体裁、多风格、人机交互创作模式等特点。

智能春联

早在2018年就已经通过网络春晚的舞台推出，可通过关键词输入的方式生成春联，包括上联、下联和横批。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

人工智能辅助写作的应用方法

阿里云

写什么？

- 选题困难；
- 不了解读者需求。

如何写？

- 没有思路；
- 缺乏资料；
- 处理效率低。

如何写好？

- 图文丰富；
- 字词纠错；
- 文章排版。

- 热门话题推荐。

- 素材推荐；
- 随材归纳；
- 内容提示。

- 智能纠错；
- 智能配图；
- 自动排版。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER