非技术也能看懂的NLP入门科普

- **NLP全景图**

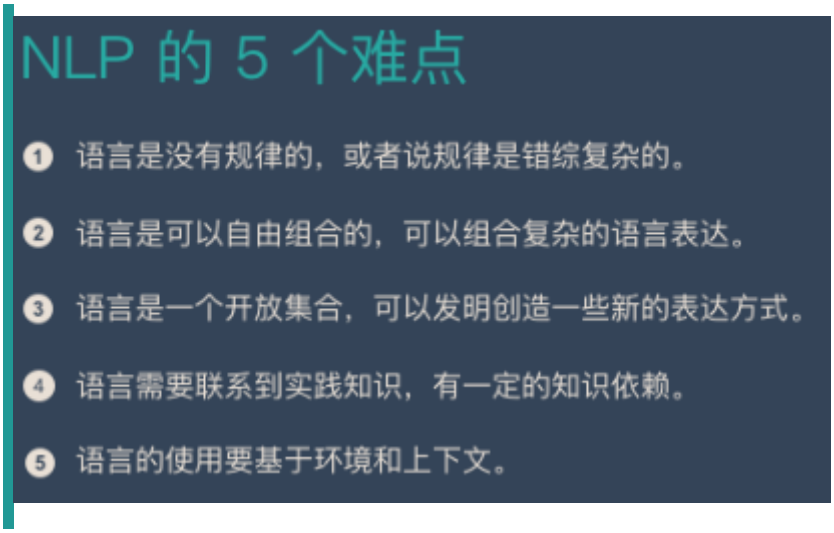
- > 在人工智能出现之前，机器智能处理结构化的数据（例如 Excel 里的数据）。但是网络中大部分的数据都是非结构化的，例如：文章、图片、音频、视频..
  > 为了能够分析和利用这些文本信息，我们就需要利用 NLP 技术，让机器理解这些文本信息，并加以利用。

  NLP的必要性：机器智能可以处理结构化数据，但非结构化数据不行。数据类型的扩展和信息的增多。

- 

  NLP的两个核心任务：NLU处理和NLG生成。

- **NLP的五个难点：1）0无规律。2）自由组合。3）发明创造。4）联系实际。5）基于上下文。**



NLP的五个难点：1）0无规律。2）自由组合。3）发明创造。4）联系实际。5）基于上下文。

- 

NLP的四个应用：1）情感分析。2）聊天机器人。3）语音识别。4）机器翻译。

- 

NLP预处理的四个步骤：1）分词。2）词性标注。3）实体识别。4）去除停用词。

- **自然语言处理 - Natural language processing | NLP**

- **NLP 为什么重要?**

- **什么是自然语言处理 - NLP**

> NLP 就是人类和机器之间沟通的桥梁！

NLP：人与机器沟通的语言。

- **NLP 的2大核心任务**

- **自然语言理解 - NLU|NLI**

> 自然语言理解就是希望机器像人一样，具备正常人的语言理解能力，由
> 上有很多难点(下面详细说明)，所以 NLU 是至今还远不如人类的表现。

自然语言理解NLU：机器像人一样能够语言理解。

> **自然语言理解的5个难点：**
>
> 1. 语言的多样性
> 2. 语言的歧义性
> 3. 语言的鲁棒性
> 4. 语言的知识依赖
> 5. 语言的上下文

NLU的五个难点

- **自然语言生成 - NLG**

> NLG 是为了跨越人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式，如文章、报告等。

自然语言生成NLG：机器将数据转化为人类可以理解的语言。

- **NLP 的5个难点**

- **NLP 的4个典型应用**

- **NLP 的 2 种途径、3 个核心步骤**

- 1. 中文文分词 - Chinese Word Segmentation
2. 词性标注 - Parts of Speech3. 命名实体识别 - NER4. 去除停用用词

中文语料预处理的4个步骤：1）中文分词。2）词性标注。3）命名实体识别。4）去除停用词。

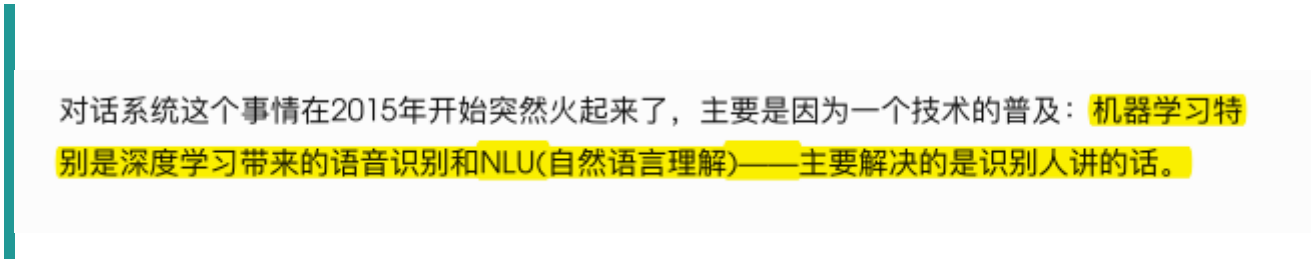
- **总结**

- 自自然语言言处理理（NLP）就是在机器器语言言和人人类语言言之间沟通的桥梁梁，以实现人人机交流的目目的。NLP的2 个核心心任务：1. 自自然语言言理理解 - NLU2. 自自然语言言生生成 - NLG NLP 的5个难点：1. 语言言是没有规律律的，或者说规律律是错综复杂的。2. 语言言是可以自自由组合的，可以组合复杂的语言言表达。3. 语言言是一一个开放集合，我们可以任意的发明创造一一些新的表达方方式。4. 语言言需要联系到实践知识，有一一定的知识依赖。5. 语言言的使用用要基于环境和上下文文。NLP 的4个典型应用用：1. 情感分析2. 聊天机器器人人3. 语音音识别4. 机器器翻译 NLP 的6个实现步骤：1. 分词-tokenization2. 次干干提取-stemming
3. 词形还原-lemmatization4. 词性标注-pos tags5. 命名实体识别-ner6. 分块-chunking

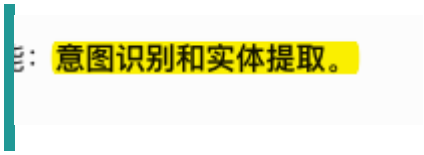
NLP总结

- **自然语言理解 - NLU | NLI**

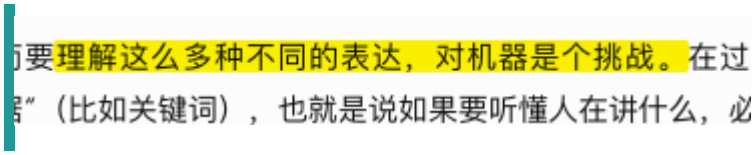
- **什么是自然语言理解(NLU)?**

- 

  NLU的发展得益于：语音识别。

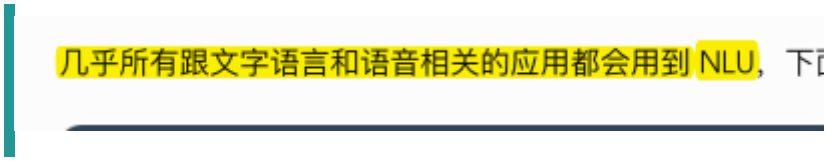
- 

  NLU的关键技能：意图识别和实体提取。

- 

  机器理解自然语言的难点eg.:
  1）同一意图，表达方式多样。比如订机票的意图，可以表达为①有去上海的航班吗，②看看航班，③查下机票。
  2）过于依赖关键词，甚至会理解成相反意图。比如订机票，退订机票包含这个词。

- **自然语言理解（NLU）的应用**

- 

  NLU的主要应用领域：文字语言。语音相关。例如机器翻译，客服，智能音箱等。

- **机器翻译**

机器翻译难点：一词多义。例如Apple。

- **机器客服**

机器客服难点：多轮对话的理解。

- **智能音箱**

智能音箱难点：意图理解。例如冷了调高空调温度。

- ## 自然语言理解（NLU）的难点

- **难点1：语言的多样性**

多样性：不同表达，但是同一个意思。

- **难点2：语言的歧义性**

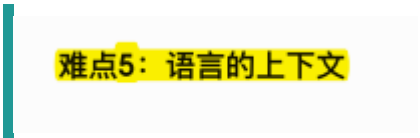
歧义性：一个词被理解成多个意图。

- **难点3：语言的鲁棒性**

鲁棒性：文字会多字少字错字等等，特别是语音转文本。

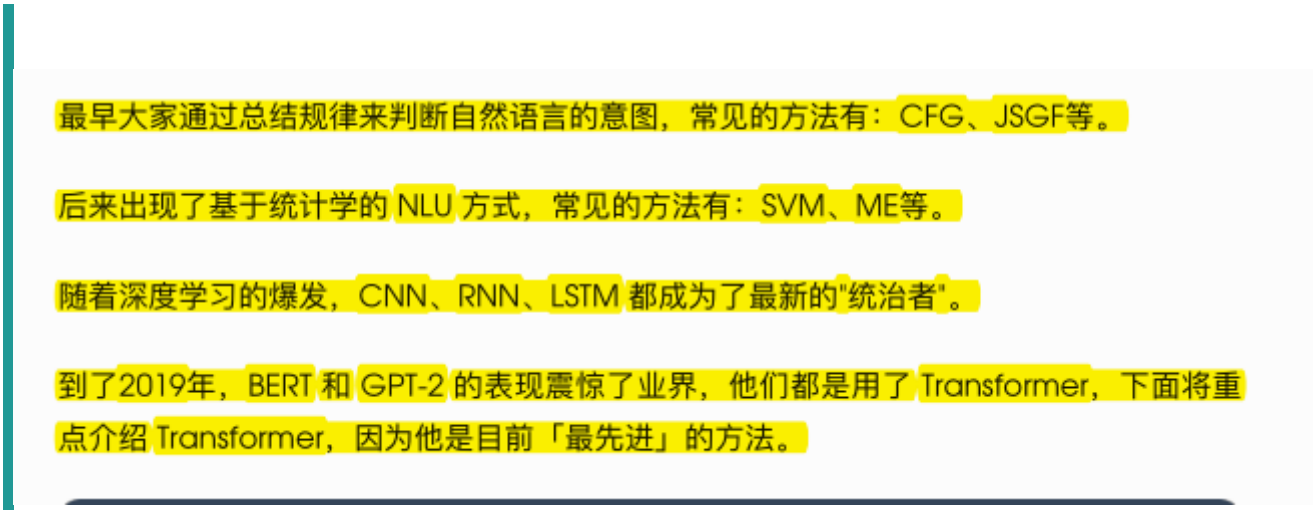
- **难点4：语言的知识依赖**

知识依赖：比如Apple，大鸭梨，7天。

- 

上下文：对话或者长文章，一个词的意思会不一样。

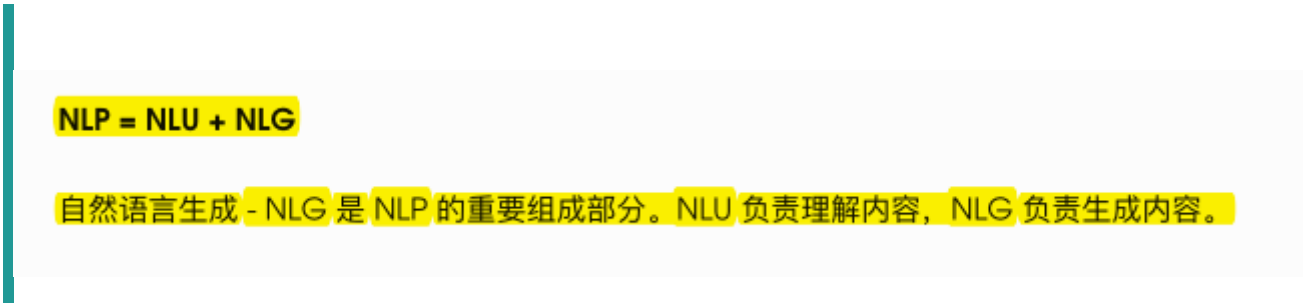
- **NLU 的实现方式**

- 

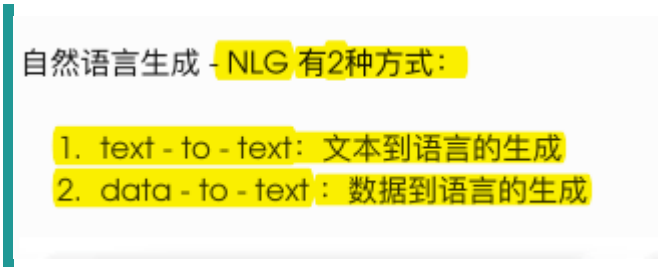
NLU的发展：1）规则。2）统计。3）深度学习。

- **自然语言生成 - NLG**

- **什么是 NLG?**

- 

NLP=NLU+NLG。

- 

NLG的两种方式：1）从文本到文本。2）从数据到文本。

- **NLG 的3个 Level**

- 

NLG的三个LEVEL：1）数据合并。2）模板抽取。3）理解意图。

- **NLG 的6个步骤**

- 第一一步：内容确定 - Content Determination作为第一一步，NLG 系统需要决定哪些信息应该包含在正在构建的文文本中，哪些不不应该包含。通常数据中包含的信息比比最终传达的信息要多。 第二二步：文文本结构 - Text Structuring确定需要传达哪些信息后，NLG 系统需要合理理的组织文文本的顺序。例例如在报道一一场篮球比比赛时，会优先表达「什什么时间」「什什么地点」「哪2支支球队」，然后再表达「比比赛的概况」，最后表达「比比赛的结局」。 第三步：句句子子聚合 - Sentence Aggregation不不是每一一条信息都需要一一个独立立的句句子子来表达，将多个信息合并到一一个句句子子里里里表达可能会更更加流畅，也更更易易于阅读。 第四步：语法化 - Lexicalisation当每一一句句的内容确定下来后，就可以将这些信息组织成自自然语言言了了。这个步骤会在各种信息之间加一一些连接词，看起来更更像是一一个完整的句句子子。
第五步：参考表达式生生成 - Referring Expression Generation|REG这个步骤跟语法化很相似，都是选择一一些

> 单词和短语来构成一一个完整的句句子子。不不过他跟语法化的本质区别在于“REG需要识别出内容的领域，然后使用用该领域（而而不不是其他领域）的词汇”。 第六步：语言言实现 - Linguistic Realisation最后，当所有相关的单词和短语都已经确定时，需要将它们组合起来形成一一个结构良好的完整句句子子。

NLG的6个步骤：1）内容确定。2）文本结构。3）句子聚合。4）语法化。5）参考表达式生成。6）语言实现。

- **NLG 的3种典型应用**

NLG的3个应用：1）自动写新闻。2）聊天机器人。3）生成BI报告。

- **总结**

- 自自然语言言生生成 - NLG 是 NLP 的重要组成部分，他的主要目目的是降低人人类和机器器之间的沟通鸿沟，将非非语言言格式的数据转换成人人类可以理理解的语言言格式。NLG 的3个level:1. 简单的数据合并2. 模块化的 NLG3. 高高级 NLG NLG 的6个步骤：1. 内容确定 - Content Determination2. 文文本结构 - Text Structuring3. 句句子子聚合 - Sentence Aggregation4. 语法化 - Lexicalisation5. 参考表达式生生成 - Referring

> Expression Generation|REG
> 6. 语言言实现 - Linguistic Realisation NLG 应用用的3个目目的：1. 能够大大规模的产生生个性化内容2. 帮助人人类洞洞察数据，让数据更更容易易理理解3. 加速内容生生产 NLG 的3个典型应用用：1. 自自动写新闻2. 聊天机器器人人3. BI 的解读和报告生生成

NLG总结

- **分词 - Tokenization**
- **什么是分词?**
- **分**

> 分词就是将句句子子、段落、文文章这种长长文文本，分解为以字词为单位的数据结构，方方便便后续的处理理分析工工作

分词：将文章、段落、句子等长文本，分解到单词的粒度

- **为什么要分词?**

- > 将复杂问题转化为数学问题

分词原因1）复杂问题转化为数学问题。

- > 词是一一个比比较合适的粒度

分词原因2）词是表达完整含义的最小单位。

- 深度学习时代，部分任务中也可以「分字」

分词原因3）关键词提取，命名实体识别必需

- **中英文分词的3个典型区别**

- **在**

分词方方式不不同，中文文更更难

中英文分词区别1）中文无空格等分隔符，一词多义情况多

- 英文文单词有多种形态

中英文分词区别2）英文单词的变形很多，需要词形还原/词干提取

- 中文文分词需要考虑粒度问题

中英文分词区别3）中文需要根据不同场景选择不同粒度

- **中文分词的3大难点**

- 没有统一一的标准

中文分词难点1）无统一标注

- 歧义词如何切分

中文分词难点2）歧义消解

- 新词的识别

中文分词难点3）新词识别

- **3种典型的分词方法**

- 常见见的分词器器都是使用用机器器学习算法和词典相结合，一一方方面面能够提高高分词准确率，另一一方方面面能够改善领域适应性

常用分词：机器学习方法与词典结合。

- **中文分词工具**

- **英文分词工具**

- **总结**

- 分词就是将句句子子、段落、文文章这种长长文文本，分解为以字词为单位的数据结构，方方便便后续的处理理分析工工作。分词的原因：1. 将复杂问题转化为数学问题2. 词是一一个比比较合适的粒度3. 深度学习时代，部分任务中也可以「分字」中英文文分词的3个典型区别：1. 分词方方式不不同，中文文更更难2. 英文文单词有多种形态，需要词性还原和词干干提取3. 中文文分词需要考虑粒度问题中文文分词的3大大难点1. 没有统一一的标准2. 歧义词如何切分3. 新词的识别3个典型的分词方方式：1. 基于词典匹配2. 基于统计3. 基于深度学习

分词总结

- **词干提取Stemming和词形还原Lemmatisation**
- **词干提取和词形还原在 NLP 中在什么位置?**
- **什么是词干提取和词形还原?**
- > 词干干提取是去除单词的前后缀得到词根的过程。

  词干提取：去除单词的前后缀。例如复数、进行时、过去式等。

- > 词形还原是基于词典，将单词的复杂形态转变成最基础的形态。

  词性还原：将单词的复杂形态转为最基础状态。例如be动词等。

- > 词干干提取和词形还原的目目的就是将长长相不不同，但是含义相同的词统一一起来，这样方方便便后续的处理理和分析

  词干提取和词性还原的目的：统一相同含义的词。

- **词干提取和词形还原的 4 个相似点**
- **词干提取和词形还原的 5 个不同点**

- 在原理理上，词干干提取主要是采用用“缩减”的方方法，将词转换为词干干，如将“cats”处理理为“cat”，将“effective”处理理为“effect”。而而词形还原主要采用用“转变”的方方法，将词转变为其原形

区别1）原理。词干提取是缩减，词形还原是转变。

- 在实现方方法上，虽然词干干提取和词形还原实现的主流方方法类似，但二二者在具体实现上各有侧重。词干干提取的实现方方法主要利利用用规则变化进行行行词缀的去除和缩减，从而而达到词的简化效果。词形还原则相对较复杂，有复杂的形态变化，单纯依据规则无无法很好地完成。其更更依赖于词典，进行行行词形变化和原形的映射，生生成词典中的有效词。

区别2）实现方法。词干提取利用规则，词形还原是词典。

- 在结果上，词干干提取和词形还原也有部分区别。词干干提取的结果可能并不不是完整的、具有意义的词，而而只是词的一一部分，如“revival”词干干提取的结果为“reviv”，“ailiner”词干干提取的结果为“airlin”。而而经词形还原处理理后获得的结果是具有一一定意义的、完整的词，一一般为词典中的有效词。

区别3）结果。词干提取的结果可能不完整、不具有意义，词形还原是词典里的词。

- **3 种主流的词干提取算法**

- **词形还原的实践方法**

- **总结**

- 词干干提取和词形还原都是将长长相不不同，但是含义相同的词统一一起来，这样方方便便后续的处理理和分析。他们是英文文语料料预处理理中的一一个环节。词干干提取和词形还原的 4 个相似点：1. 目目标一一致2. 部分结果一一致
3. 主流实现方方式类似4. 应用用领域相似词干干提取和词形还原的 5 个不不同点：1. 原理理上不不同2. 词形还原更更加复杂3. 具体实现方方式的侧重点不不同4. 呈现结果有区别5. 应用用领域上，侧重点不不完全一一致3 种词干干提取的主流算法：1. Porter2. Snowball3. Lancaster英文文的词形还原可以直接使用用 Python 中的 NLTK 库，它包含英语单词的词汇数据库。

词干提取和词性还原总结

- **词性标注 – Part of speech**

- **什么是词性标注?**

- 词性指以词的特点作为划分词类的根据。词类是一一个语言言学术语，是一一种语言言中词的语法分类，是以语法特征（包括句句法功能和形态变化）为主要依据、兼顾词汇意义对词进行行行划分的结果。

词性：以语法特征为主要依据，兼顾词汇意义对词进行分类的结果。

- 从组合和聚合关系来说，一一个词类是指：在一一个语言言中，众多具有相同句句法功能、能在同样的组合位置中出现的词，聚合在一一起形成的范畴。词类是最普遍的语法的聚合。词类划分具有层次性。

词类：聚合相同语法和组合位置的词形成的。具有层次性。

- **中文词性标注的难点**

> 汉语是一种缺乏词形态变化的语言，词的类别不能像印欧语那样，直接从词的形态变化上来判别。
>
> 常用词兼类现象严重。《现代汉语八百词》收取的常用词中，兼类词所占的比例高达22.5%，而且发现越是常用的词，不同的用法越多。由于兼类使用程度高，兼类现象涉及汉语中大部分词类，因而造成在汉语文本中词类歧义排除的任务量巨大。
>
> 研究者主观原因造成的困难。语言学界在词性划分的目的、标准等问题上还存在分歧。目前还没有一个统的被广泛认可汉语词类划分标准，词类划分的粒度和标记符号都不统一。词类划分标准和标记符号集的差异，以及分词规范的含混性，给中文信息处理带来了极大的困难。

中文词性标注的难点：1）汉语词汇没有形态变化。2）常用词的用法很多，歧义消除很复杂。3）目前暂无汉语划分标准。

- **词性标注4种常见方法**
- **词性标注工具推荐**
- **命名实体识别 – Named-entity recognition | NER**
- **什么是命名实体识别?**

> 命名实体识别（Named Entity Recognition，简称NER），又又称作“专名识别”，是指识别文文本中具有特定意义的实体

命名实体识别：识别文本中具有特定意义的边界和类别

- **命名实体识别的发展历史**

- **4类常见的实现方式**

- > 宗成庆老老师在统计自自然语言言处理理一一书粗略略的将这些基于机器器学习的命名实体识别方方法划分为以下几几类：有监督的学习方方法：这一一类方方法需要利利用用大大规模的已标注语料料对模型进行行行参数训练。目目前常用用的模型或方方法包括隐马马尔可夫模型、语言言模型、最大大熵模型、支支持向量量机、决策树和条件随机场等。值得一一提的是，基于条件随机场的方方法是命名实体识别中最成功的方方法。半监督的学习方方法：这一一类方方法利利用用标注的小小数据集（种子子数据）自自举学习。无无监督的学习方方法：这一一类方方法利利用用词汇资源（如WordNet）等进行行行上下文文聚类。混合方方法：几几种模型相结合或利利用用统计方方法和人人工工总结的知识库。

  命名实体识别方法：1）有监督。2）半监督。3）无监督。4）混合方法。

- **NER 的相关数据集**

- **相关工具推荐**