附件三：

推荐资源一：文本理解系统 TexSmart

针对业界现有的 NLP 工具在深层次文本理解方面的不足，腾讯 AI Lab 开放自然语言理解系统 TexSmart，用于对中文和英文两种文本进行词法、句法和语义分析。除了支持分词、词性标注、命名实体识别、句法分析、语义角色标注等常见功能外，TexSmart 还提供细粒度命名实体识别、语义联想、深度语义表达等特色功能。腾讯 AI Lab 此次开源，是依托公司数据源的优势，对自身基础 AI 能力的一次展示，可为自然语言文本提供更多语义层次的结构化分析与处理，推动学术研究和工业应用环境下 NLP 任务效果的提升（更多详情请访问：https://mp.weixin.qq.com/s/pLdKTgXogtITR2BvpwEvmg）。

为什么使用腾讯 AI Lab 开源的 TexSmart 文本理解系统？

1. **更细粒度命名实体识别：**支持上千种实体类型，类型之间具有层级结构，可为下游的 NLP 应用提供更为丰富的语义信息。目前多数公开的文本理解工具只支持人、地点、机构等几种或者十几种粗粒度的实体类型；TexSmart 能够识别包括人、地点、机构、产品、商标、作品、时间、数值、生物、食物、药品、病症、学科、语言、天体、器官、事件、活动等上千种实体类型。在常见的人、地点、机构等大类中，能够识别出常见的细粒度子类型，如演员、政治人物、运动员、国家、城市、公司、大学、金融机构等。如下图所示：

2. **增强的语义理解功能：**

语义联想：语义联想的功能是对句子中的实体给出与其相关的一个实体列表。语义联想是增强理解实体语义的一种方式，在工业界应用广泛，比如搜索、推荐。

特定类型实体的深度语义表达：针对时间、数量等特定类型的实体，TexSmart 能够分析它们潜在的结构化表达，以便进一步推导出这些实体的精准语义。深度语义理解对某些类型的 NLP 应用至关重要。比如，在智能对话中，某用户于 2020 年 4 月 20 日向对话系统发出请求，“帮我预定一张后天下午四点去北京的机票”。智能对话系统不但需要知道"后天下午四点"是一个时间实体，还需要知道这个实体的语义是"2020 年 4 月 22 日 16 点"。目前大多数公开的 NLP 工具不提供这样的深度语义表达功能，需要应用层自己去实现。

3. **为多维度应用需求而设计：**学术界和工业界不同的应用场景对速度、精度和时效性的要求有所不同；此外，速度和精度通常很难兼得。因此，在设计 TexSmart 时我们全面考虑了这三方面的需求。首先，针对某一功能（比如命名实体识别）TexSmart 实现了多种不同速度和精度的算法与模型，供上层应用按需选择，以便满足不同应用场景下的多样化需求。其次，TexSmart 的构建利用了大规模的无结构化数据以及无监督或弱监督方法。一方面这些无结构化数据覆盖大量时效性很强的词和实体（比如上文中的"流浪地球"，再比如新的疾病"新冠肺炎"）；另一方面无监督或弱监督方法的采用使得该系统可以以较低的代价进行更新，从而保证它具有较好的时效性。

	现有工具	TexSmart
实体粒度	人、地点、机构等 十几种粗粒度实体类型	千种不同粒度的实体，包括 **作品、时间、数值**等粗粒度， **演员、政治人物、运动员**等细粒度
语义联想	不支持	实现**语义联想** 如：流浪地球 -> 战狼二、上海堡垒等
实体语义表达	不支持	**特定类型实体的语义表达** 如：（20日）后天下午 -> 22日下午

如何使用 Tencent ML-Images

请访问 https://texsmart.qq.com，阅读具体的使用说明，使用本系统。

推荐资源二：基于深度图神经网络的自监督分子图预训练模型

AI 药物研发领域存在两个棘手问题，其一是带标注信息（label）的药物小分子数据不足； 其二是已有模型的迁移/泛化能力不足，在某个小分子数据集上训练的模型往往很难泛化到另一个数据集上。

为了解决在药物研发中的上述问题，腾讯 AI Lab 机器学习团队开发了大规模自监督分子图预训练模型 GX。作为业界首个开源的基于深度图神经网络大规模的分子图预训练模型，研究人员可以快速将 GX 作为基础组件应用到需要对小分子进行编码的药物研发相关项目中，进而推动药物研发相关应用性能提升，例如分子属性预测，虚拟筛选等任务。 为了方便研究人员使用 GX 预训练模型，我们将发布不同大小的模型实例。研究人员可以根据不同算力需求灵活选择对应大小的预训练模型进行使用。对于非机器学习相关领域研究者，我们也将发布预先计算好的两百万常见分子的指纹数据，方便直接查询使用。

为什么使用 GX 预训练模型？

超大规模分子信息编码： GX 使用了千万级的无标签分子数据进行训练，编码了丰富的分子结构信息。灵活的使用方式：GX 可以根据需要单独生成分子指纹，也可以作为一个有效的分子编码器融合到现有模型中。

灵活的使用方式：GX 可以根据需要单独生成给定分子的分子指纹，也可以作为一个有效的分子编码器融合到现有模型中。

如何使用 GX 预训练模型？

分子指纹：通过自监督任务的构造，GX 编码了丰富的分子结构信息。直接使用 GX 可以输出原子特征向量以及分子指纹作为下游任务的输入。

分子编码器：因为 GX 是基于深度图神经网络模的模型，可以将其作为基础编码组件融合到现有的模型中进行 Fine-tune，以获取更好的性能。

资源发布时间：

2020 年 5 月 30 日前发布分子指纹部分；同年 6 月中旬发布预训练模型和相关代码。

资源下载地址

资源链接和使用文档，请参阅：https://ai.tencent.com/ailab/ml/gnnpretrain.html。