模板：A类企业命题

## 命题方向

人工智能

## 题目类别

计算类

## 题目名称

金融领域公司实体消歧系统构建

## 背景说明

### 1、整体背景

互联网时代，信息爆炸。面对海量资讯，我们希望前沿AI技术能将文本和海量实体（公司、人名等）信息关联起来，提升用户阅读流畅度、实现精准化内容推荐等。智能资讯处理，不仅给金融提供智能化服务，还能为金融业务提供更多创新空间。

### 2、公司背景

恒生电子股份有限公司（简称恒生公司）于1995年成立于杭州，2003年在上海证券交易所主板上市（代码600570），是中国领先的金融软件和网络服务供应商。恒生聚焦于财富资产管理，致力于为证券、期货、基金、信托、保险、银行、交易所、私募等机构提供整体的解决方案和服务，为个人投资者提供财富管理工具。 目前，恒生公司拥有约8000人的高素质专业队伍，其中研发工程人员约占80%，在规划咨询、软件开发实施、技术服务、系统集成及系统维护服务等方面，具备强劲实力和竞争优势。 多年来，恒生以技术服务为核心，凭借多年金融IT建设经验，以及对互联网的深刻洞察和理解，用优质的产品与服务，驱动金融机构创新发展。

### 3、业务背景

文本资讯是公司实体信息传播的主要媒介，精确定位发生新闻的公司实体直接决定如何开展下游金融工作。在金融资讯中，公司实体（公司实体数以千万计）多以领域简称的形式出现，从而引发歧义。例如，苹果是一个美国上市公司，也是一种水果。实体消歧旨在消除资讯处理过程中的实体多义问题，提纯文本信息。

歧义消除一般通过结合实体知识实现。近年来，人工智能技术飞速发展，使得众多问题的解决成为可能。我们希望能应用前沿人工智能方法，解决智能资讯中的实体歧义问题。

## 任务描述

本任务为金融领域中文文本的公司实体消歧系统构建。实体消歧（实体链接）的本质在于一个单词可能有多个意思，需通过上下文和知识库相关知识确定它所表达的确切含义。公司全称是确定性的，但在资讯、研报、问答等文本中，公司实体一般以简称的形式出现。同一个公司简称可能指向多个实体，例如,“恒生”可能指向恒生银行、恒生指数、恒生电子股份有限公司等。再比如，苹果是一个美国上市公司，也是一种水果，也是一种期货，还是大家熟知的手机iphone。

实体消歧系统构建可包括知识库构建（维护实体知识，可包括目标实体的属性、文本描述等）、实体识别（识别命名实体，确定待消歧对象及其类别信息）和实体链接（与知识库中的对应实体进行关联）三个子任务。如有实体识别模块，实体识别模型的训练可使用额外的训练资源。

传统实体消歧任务主要是基于知识库的长文本，知识库较全，长文本具有更丰富的上下文信息以辅助实体消歧。本任务是基于垂直域（公司实体）消歧数据的实体消歧系统构建存在更大挑战：

基于标注训练数据中生成知识库，且知识库需与下游消歧模型呼应

短文本上下文语境不丰富，知识库和语义理解需更加准确

与英文相比，中文由于自身的特点，短文本实体消歧更具挑战

## 数据描述

本任务的标注数据包括训练集、开发集和测试集。前期将发布训练集和开发集，其中，知识库构建主要基于训练集，模型训练及选择基于训练集和开发集。测试集将在后期公布，用于对参赛队伍基于训练集和开发集构建的消歧系统进行评测。

**训练集和开发集**数据形式为，

text\_id：测试数据集编号

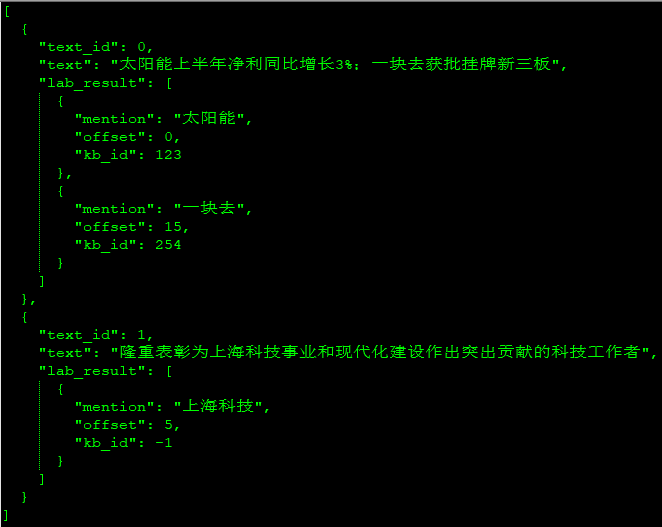
text：测试数据集编号对应的句子文本

mention\_result：该句子的消歧结果（list,每个元素为一个mention的消歧结果）

mention：实体提及（待消歧的实体）

offset：该实体提及在句子中的起始位置（句首的位置为0）

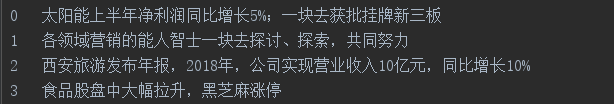
kb\_id：消歧结果对应在实体表中的编号（-1表示非领域词）



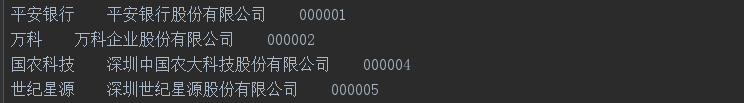
**输入**：text

**输出**：(mention, offset, kb\_id), 表示待消歧实体mention（通过offset定位）消歧之后在实体库中的id

**测试数据**只提供**可能包含待消歧实体出现的文本及其文本编号**。例如，“各领域营销的能人智士一块去探讨、探索，共同努力”，该文本包含待消歧实体“一块去”，其对应编号1；文本编号3对应“食品股盘中大幅拉升，黑芝麻涨停”，其中“黑芝麻”需要消歧。具体形式如下，



每一行数据包括文本编号和文本，用\t进行分隔。测试过程中，对于单条测试文本，选手基于我们提供的待消歧实体列表对该文本中的待消歧实体进行消歧，消歧结果格式见**测试结果提交见《任务提交》章节。**待消歧实体列表形式如下：



分别对应待消歧实体、消歧后所关心的实体和消歧所后关心实体的编号。

## 评价指标

实体消歧系统输出结果包括文本中的特定命名实体及其消歧结果，模型评测通过对比模型输出和人工标注结果，计算准确率（Precision），召回率（Recall）和F1值（F1 score）。F1值越高，模型排名越靠前。

给定文本输入（用Query表示）， Q中N个实体mentions、位置及其链接到是知识库的实体id人工标注结果为:。相应地，对于模型输出结果为:。则有，

**精准率**：

**召回率：**

**F1值：**

**最终解决方案评分划分**

|  |  |  |  |
| --- | --- | --- | --- |
| **参考评分表** | | **合计分值** | **备注**  **说明** |
| 模型表现 | F1值 | 30分 |  |
| 算法创新 | 算法独特、新颖，创新元素多，具有技术含量，可以有效地解决此类问题。 | 30分 |  |
| 技术实现 | 技术路线清晰明确、技术工具成熟可靠；技术方案可行性高，项目完成度好；代码可读性强；技术资源及经济成本控制合理，与项目需求匹配恰当。 | 20分 |  |
| 项目展示 | 提交文档结构清晰合理、逻辑顺畅、文笔简练。 | 20分 |  |
| 总分 | | **100分** |  |

## 任务提交

待测试集（未标注，只有句子及编号信息）发布之后，参赛队伍提交此数据的消歧结果，以result.json命名（UTF-8格式），系统会自动计算各队伍的评测结果。

测试结果数据规范为：

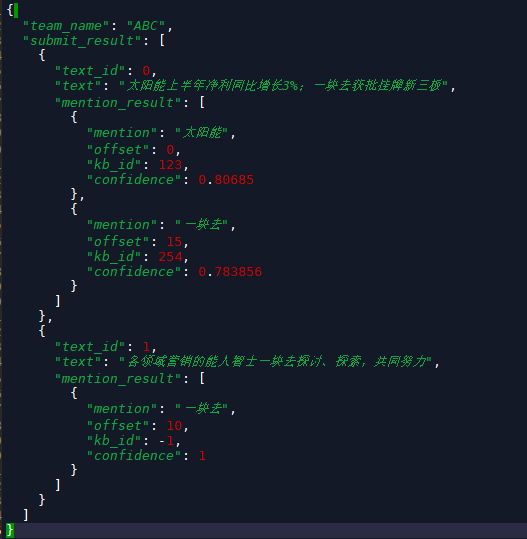
1. 数据返回为dict（字典形式）。team\_name声明队伍名；submit\_result为整个测试集对应结果的list串，每个元素对应一个测试集文本id的消歧结果
2. 每个测试集文本id对应的测试结果同样用dict（字典）形式进行输出。包括，text\_id表示测试集id；text是与text\_id对应的文本；mention\_result代表该测试文本的所有消歧结果构成list结构，可能有多个待消歧实体的消歧结果。
3. 每个待消歧实体的消歧结果为一个dict（字典）。包括，mention：实体提及（文本中出现的待消歧实体）；offset：mention在句子中的起始位置（句首的位置为0）; kb\_id：消歧结果对应在实体表中的编号（**-1表示mention指向并非我们所关系的实体**）; confidence：置信度得分（不计入评分，只作为模型评测的辅助）
4. 提交结果的形式须参照上述数据结构，否则无法开展评测。

代码及方案文档需要打包成一个文件（tar，zip，gzip，rar等均可），用code.xxx命名，要求提交所有的程序代码及相关的配置说明，程序应当可以运行且所得结果与result.txt相符。如果方法使用了额外资源，要求说明并提供资源文件或地址。

下面以两条测试数据为例，展示测试结果：

0 太阳能上半年净利润同比增长3%；一块去获批挂牌新三板

1 各领域营销的能人智士一块去探讨、探索，共同努力



**注：提交的结果需包含该句中所有可能歧义子串的消歧结果（相同子串，位置不同，需分别给出消歧结果）。例如，**

