**基于百度百科的实体消歧**

# 1.项目背景

此项目旨在做“实体消歧”工作，词义歧义是指一词多义，即一个词在不同的上下文中有不同的含义。

消歧在多个领域中都扮演着非常重要的角色，包括机器翻译、信息抽取、主题发现、文本分类等。

百度百科作为目前中国最大的、使用最广泛的在线百科全书，目前已经成为语言信息处理的重要资源。

对于百科消歧可以建立文本-百科页面的余弦相似度模型来实现，也可以通过比较已知实体类别与百科中类别标签并结合上下文词语特征来实现，还可以构建机器学习模型自动获取标注语料，并对语料增加百科特征实现。

## 1.1人物实体的获取

在“计算机领域”知识图谱中，从ccf专委页面中抽取专委名实体及各个专委对应的人物描述，如：刘江——美团大众点评技术学院院长。

## 1.2人物实体非结构化信息抽取

从百度百科抓取人物实体的描述，从中抽取去除标签后、抽取出正文的人物实体非结构化描述信息。

# 2.任务描述

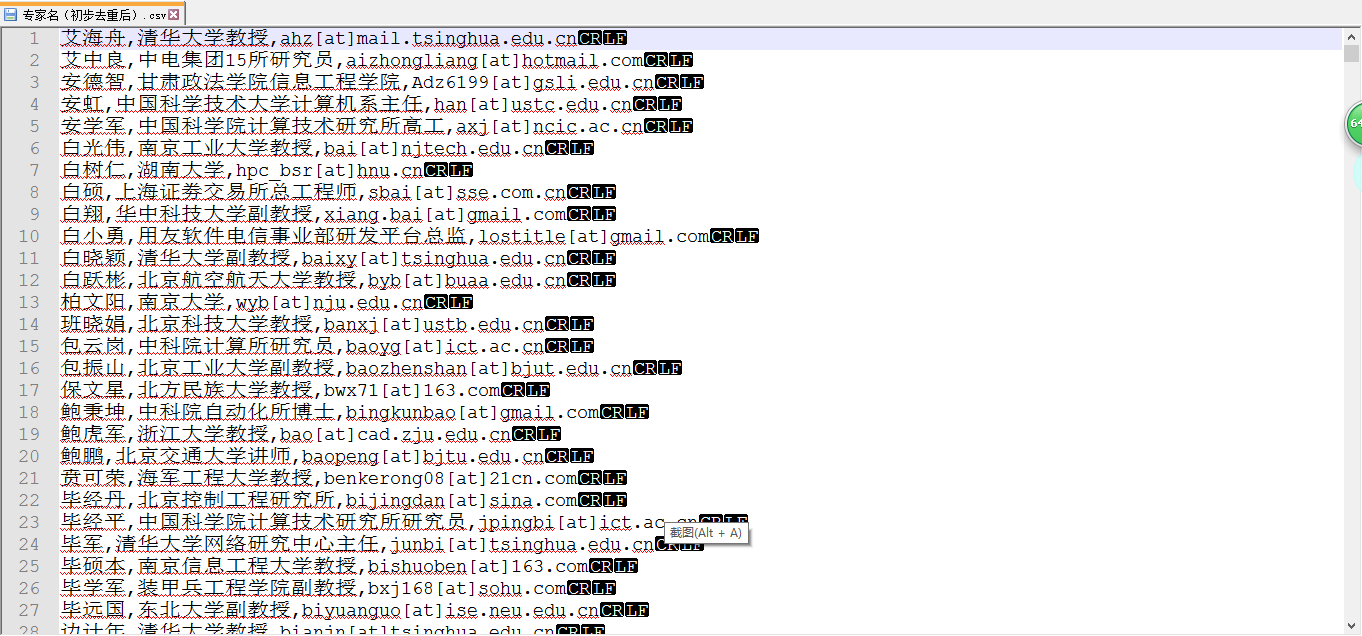
由于人物实体有重名的现象，所以百度百科同一人会有多个义项，如：王勇有203个义项。我们要做的就是将ccf专委页面的带简单描述的人物实体和百度百科中的人物义项一一对应出来。

# 3.数据采集及预处理

## 3.1人物实体的数据采集

对于ccf官网的34个专委的主任、副主任、秘书长、常务委员、资深委员、委员进行抓取，共得到4000+人物名称及对应的人物简短描述。

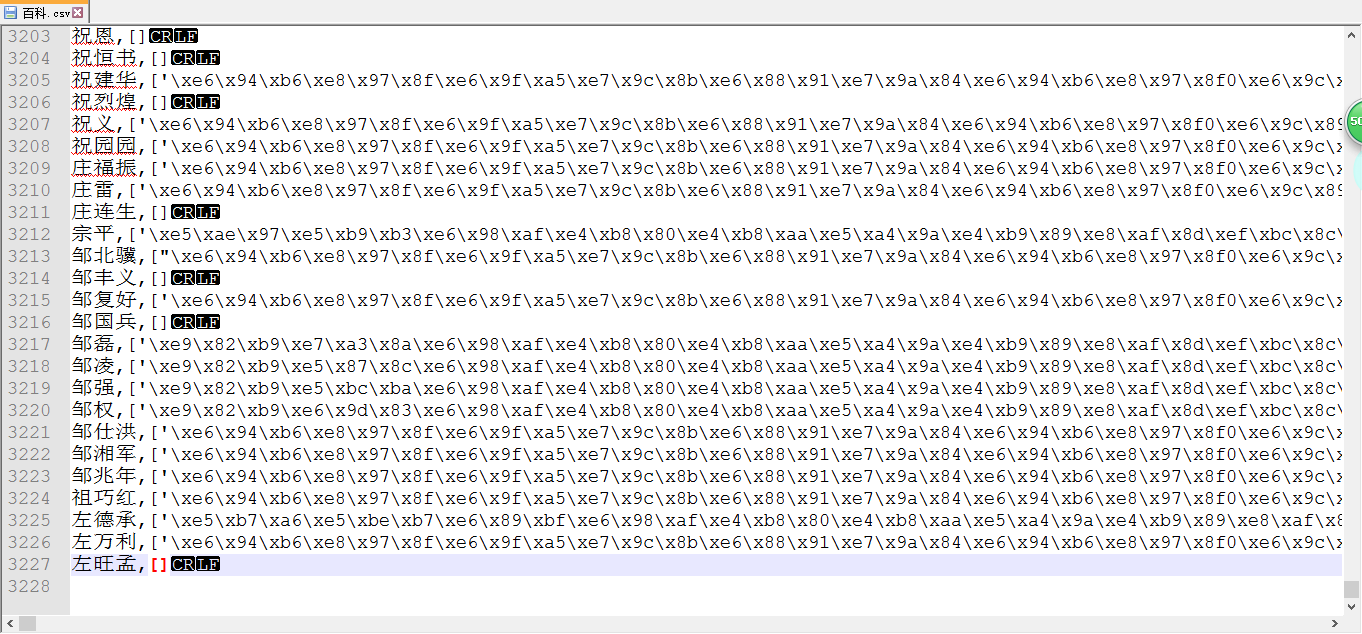
由于同一人物可能参加了不同的专委，所以抓取的名单人物会有重名同一实体的现象，需要做一个简单的去重复预处理，将重复人物的简单描述进行分词，统计词项的异同，如果出现过的词项相同，则将该重复人物进行合并，合并后的人物实体数目为3224个，其中还有200+左右重名情况，效果较好。以下为采集并预处理后的ccf人物实体数据：



## 3.2百科的数据采集

对于人物的百度百科，抓取所有义项的描述，抓取到的页面是带标签的混合非结构化文本。

但是进行后期消歧处理时，需要用义项描述的网页正文作为语料。所以对于目前抓取到的页面数据，用正则表达式将标签去除，利用基于网页密度的算法抽取出网页正文。由于停用词对语料的质量有所影响，所以去除停用词。以下为采集并预处理后的百科数据：



## 3.3训练语料的构建

# 4.词向量表示（Word Embedding/Word Representation）