**基于百度百科的实体消歧**

# 1.项目背景

此项目旨在做“计算机领域”知识图谱，其中包括机构、专家人物等实体及其之间的联系。

本次实践聚焦**ccf专家名单中实体**与**百度百科同名的不同义项**的“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、建立文本-百科页面的余弦相似度模型来实现，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

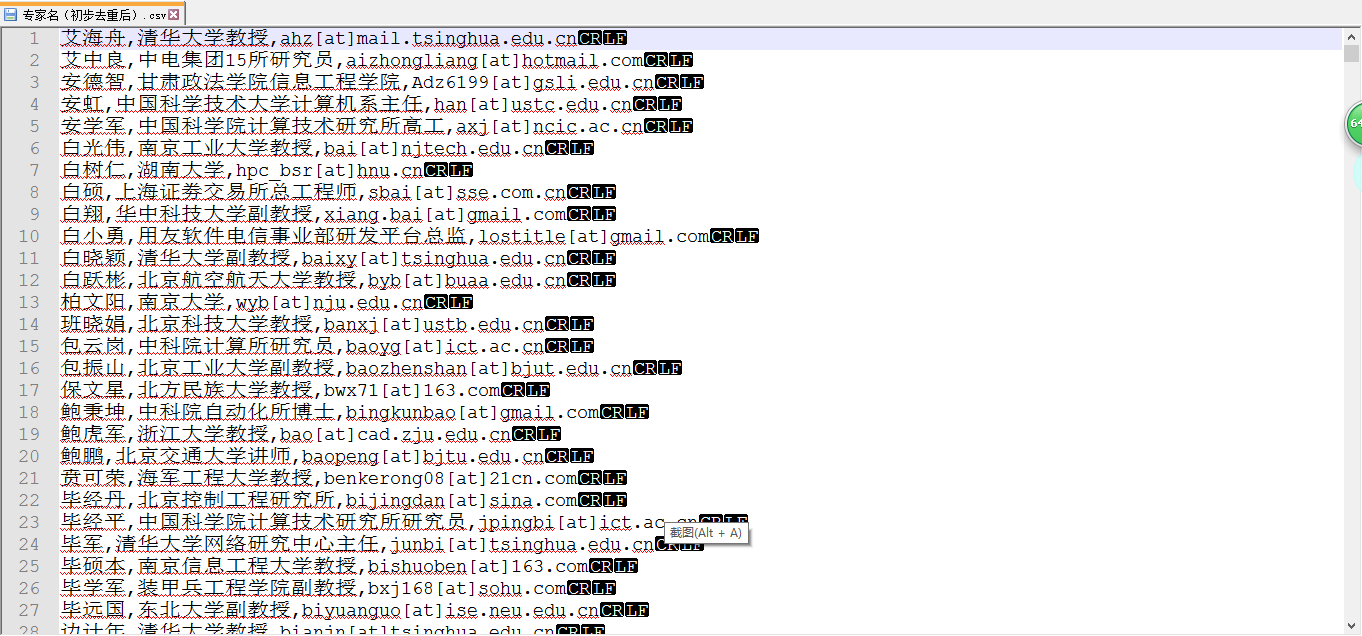
由于人物实体有重名的现象，所以百度百科同一人会有多个义项，如：王勇有203个义项。我们要做的就是将ccf专委页面的带简单描述的人物实体和百度百科中的人物义项一一对应出来。

# 3.数据采集及预处理

## 3.1人物实体数据的构建

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。

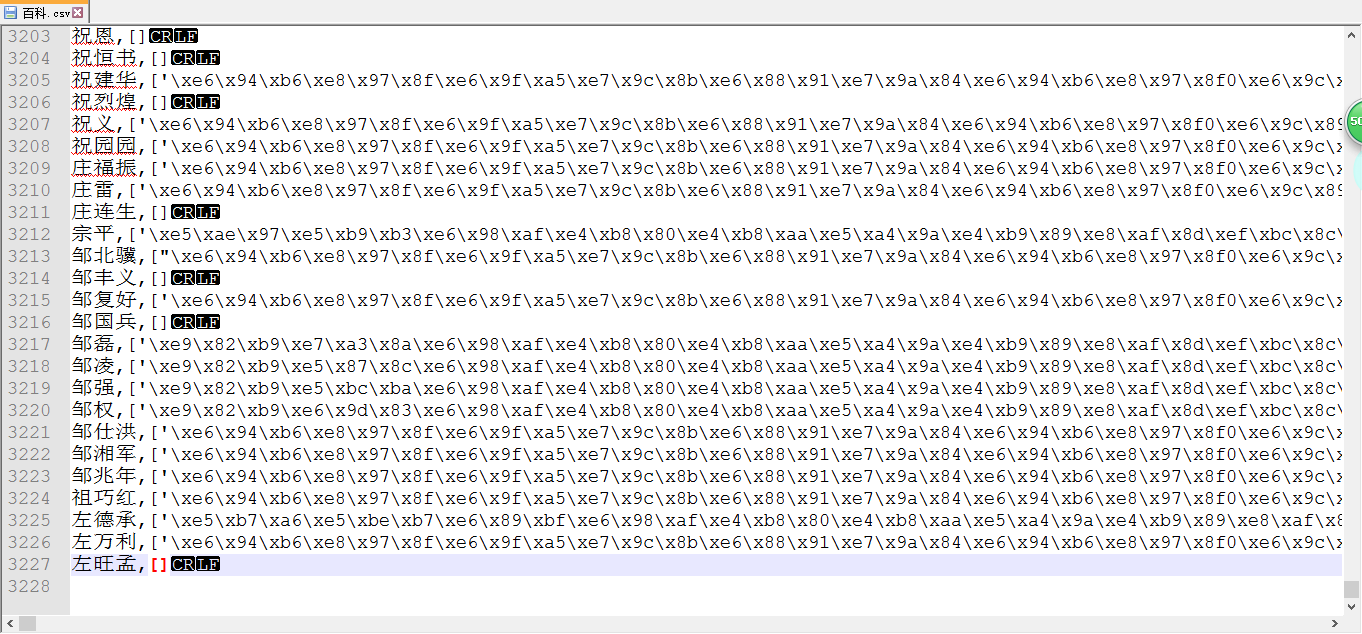
由于同一人物可能参加了不同的专委，所以抓取的名单人物会有重名同一实体的现象，需要做一个简单的去重复预处理，将重复人物的简单描述进行分词，统计词项的异同，如果出现过的词项及其频率相同，则将该重复人物进行合并，合并后的人物实体数目为3224个，其中还有200+左右重名情况，剩余的重复项手工筛选，最后得到的效果较好。以下为采集并预处理后的ccf人物实体数据：



## 3.2百科数据的构建

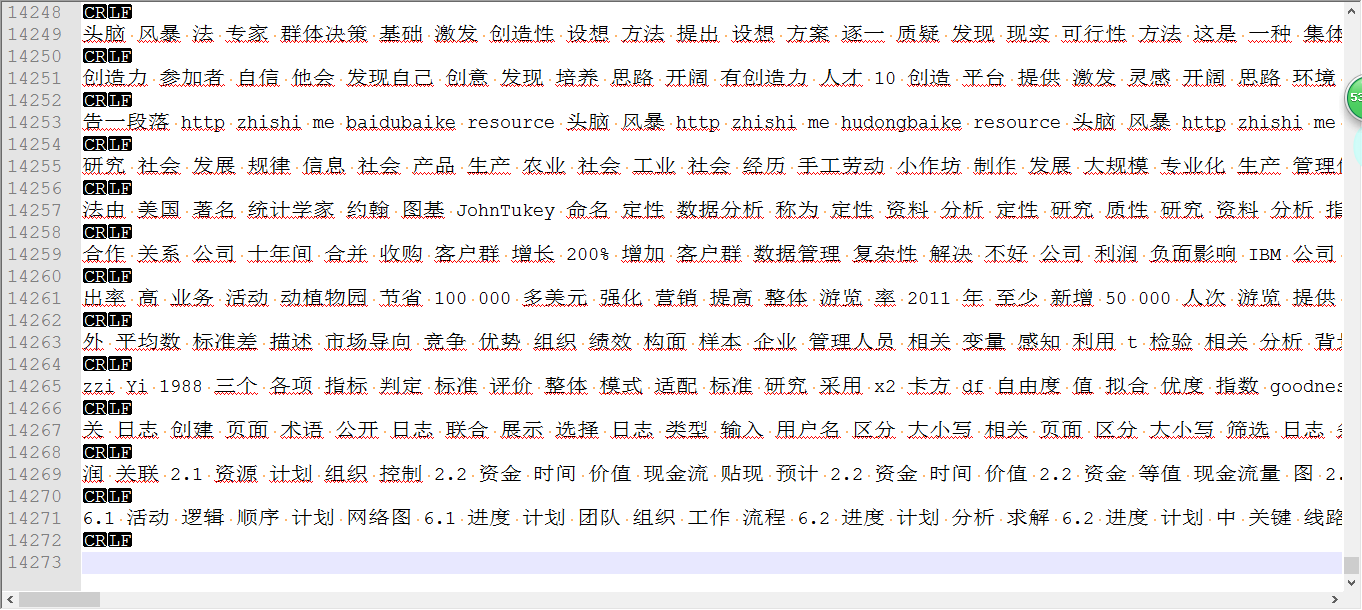
对于人物的百度百科，抓取所有义项的描述，抓取到的页面是带标签的混合非结构化文本。

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，并进行分词操作。以下为采集并预处理后的百科数据：



## 3.3训练语料的构建

抓取ccf官网全站（大概6000+页面），抓取到的页面是带标签的混合非结构化文本。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，并进行分词操作。以下为采集并预处理后的训练语料数据：



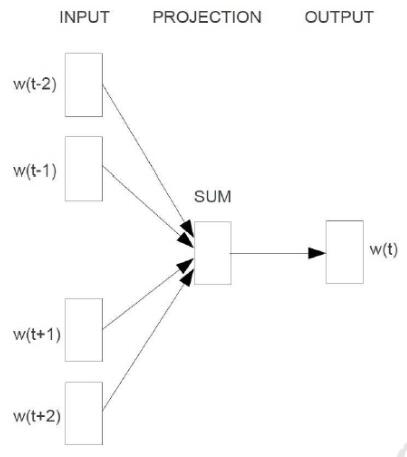
# 4.词向量训练——CBOW模型

训练的过程如下图所示，主要有输入层(input)，映射层(projection)和输出层(output)三个阶段。

输入层：当前词前面c个词，以及后面c个词。

隐藏层：输入层若干向量相加。

输出层：树结构，以语料库中出现的词作为叶子节点。



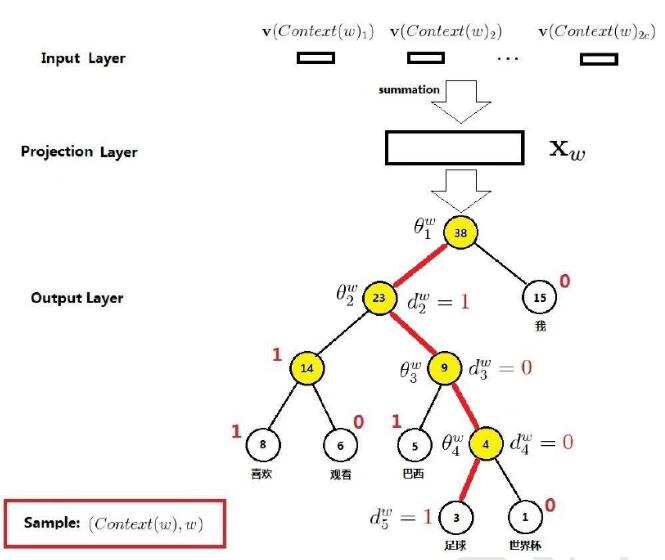
## 4.1从输入层到映射层

输入层即为某个词汇A周围的n-1个单词的词向量。如果n取5，则词A(可记为w(t))前两个和后两个的单词为w(t-2),w(t-1),w(t+1),w(t+2)。相对应的，那4个单词的词向量记为v(w(t-2)),v(w(t-1)),v(w(t+1)),v(w(t+2))。从输入层到映射层比较简单，将那n-1个词向量相加即可。

## 4.2从映射层到输出层

将条件概率用哈夫曼树表示。树中每个节点表示在当前上下文对词进行一次分类的概率，叶子节点表示词在上下文下的条件概率。

从哈夫曼树根节点开始，映射层的值需要沿着Huffman树不断的进行logistic分类，并且不断的修正各中间向量和词向量。



# 5.词向量训练——Skip-gram

skip-gram的输入是当前词的词向量，而输出是周围词的词向量。也就是说，通过当前词来预测周围的词。由于输出有n-1个词，所以要对于一个词来讲，上述沿着huffman树从顶到底的过程要循环n-1遍。

# 6.相似度计算（实体链接）

计算ccf每个人物对应的若干词组成的列表与其百科的各个义项的文档（多个词）的余弦相似度，找到最大的相似度，则其对应的义项就是我们要找的。

这里有个基本假设：ccf每个人物对应的若干词组成的列表与其对应的百科的义项的文档相似度最大。

# 7.结果评估

在一共3000+ccf实体和110000+百度百科义项中进行实体链接，采样200个ccf实体，人工查看正确与否，发现正确率：0.87，召回率0.9，F1值为：0.44。

# 8.程序运行

到项目目录下:

1. **更改settings.py文件中的目录路径等设置。**
2. **抓取ccf人物实体数据并清洗：**

python get\_and\_preprocess\_data\get\_data\_from\_ccf.py

1. **抓取百科义项并清洗：**

python get\_and\_preprocess\_data\get\_data\_from\_baike.py

1. **抓取训练语料并清洗：**

python get\_and\_preprocess\_data\ get\_corpus.py

python model\ merge\_corpus\_and\_baike.py

1. **训练词向量并进行实体链接**

python model\train.py