1. 构建知识图谱

1.1 数据采集

1.1.1 非结构化数据采集

1.1.1.1书籍等

1.1.2 半结构化数据采集

1.1.2.1基于beautifulsoup+requests+Selenium的爬虫系统,再结合正则表达式和Xpath的配合解析得到.本文选择xpath表达式定位所需数据元素。首先我们依据Selenium模仿用户操作行为,进行网页获取,之后通过xpath解析网页数据,对文旅各项信息进行匹配捕捉,并进行爬取。 本文选择xpath表达式定位所需数据元素。首先我们依据Selenium模仿用户操作行为,进行网页获取,之后通过xpath解析网页数据,对汽车各项信息进行匹配捕捉,并进行爬取。 爬取下来的数据存储为CSV格式,方便后续的数据处理

1.1.2.2 后以较专业的百度百科数据作为Scrapy爬取的来源。Scrapy是一种以Python为基础的爬虫引擎框架

1.2 模式层设计

1.2.1 在知识图谱的抽象层上，由不同的概念模型组成的知识图谱范例形成知识本体。本文在充分了解某地区文旅领域知识体系后，基于病文旅模型定义了所需要的知识结构及关系、属性类型，设定对应的约束和清晰的知识提取界限。 文旅领域一般包括一个景区/地区自身，周边以及评价这几个点。根据这几点还能向下进行分类。具体概念层次图如下图

…………………………

1.3数据层构建，建立三元组数据

采用规则的方式提取半结构性的文本数据，并运用深度学习的方式对无结构的文本数据进行抽取。根据获得的数据，将数据分类成多个实体，并且构建实体-关系-实体这样的三元组数据 据语料库的特点,基于BIOES序列标注方法对多文本进行标注,并采用BiLSTM-CRF模型进行文本特征提取,能够有效实现长文本提取多个三元组数据。

1.4数据层构建

将经过处理的汽车领域数据以RDF文件形式存放到Neo4j数据库中。Neo4j是属于NoSQL类型的图形化数据库 ，Neo4j的数据导入方式有多种,包括直接载入CSV数据格式、编写Cypher语句进行导入、使用官方的Neo4j-import工具进行导入。编写Cypher导入速度较慢,查漏补缺可以使用。所以本文使用直接结合Python导入CSV数据。再查漏补缺时自行编写Cypher进行导入

1. 问答系统算法框架

2.1问答系统算法框架

2.1.1

BBA-CRF实体识别算法 （BBA-CRF模型的实现+基于字典匹配的实体识别 +实体链接 ）+B-CNN关系识别算法（B-CNN模型的实现 +相似度计算 ）

2.1.2

问句意图分类算法（基于Word2Vec的词向量模型+意图分类数据集的构建+基于TextCNN的意图分类）+问句命名实体识别（基于AC自动机的苹果问句实体识别+基于BiLSTM-CRF模型的苹果问句实体识别）+基于苹果病虫害知识图谱的答案检索和评估（基于Cypher语句的答案检索）