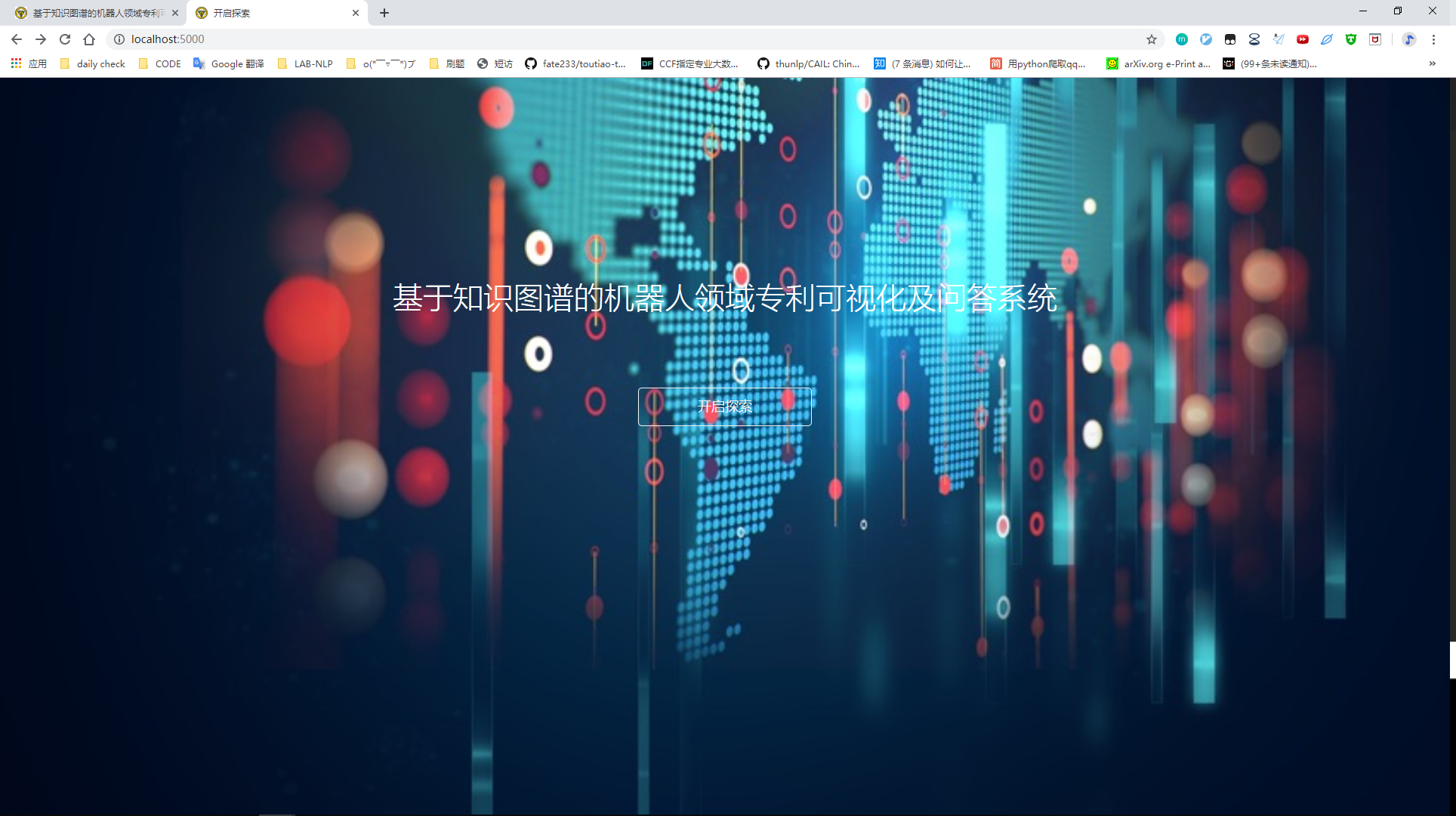
# 机器人专利知识图谱



## 机器人专利知识图谱功能模块

对机器人领域专利进行类别划分。根据类别和专利之间的相似度和CPC分类号分别赋予一定权值，根据权值大小对每个专利文本进行分类。

对机器人手册中介绍的机器人设计构造流程进行结构化提取。

关键技术查询。结合网页结构化语句和分类文本和引用关系等进行关键技术信息返回。

专利文本关键技术提取。从每个专利文本中提取得到关键技术和包含关系等信息。

技术术语推荐。根据键入短语的相似度在知识图谱中检索最相关的单词和短语。

问题解决回答。根据问题提取实体，结合返回设计流程。

整体思路：

构建两个网络：

 分类网络（主）

 引用网络

分类网络

 机器人手册

 专利文本

 专利号

初步构想：

利用专利文本进行信息挖掘，找出不同的（实体-关系-实体/属性）三元组。

根据机器人手册的组织架构，进行专利文本的归类，再用专利号进行一定程度上的验证。

机器人手册的组织架构：分实体层级、方法/属性层级，每层级有一定的术语词描绘。对专利进行分类。

专利号验证：使用专利号所在类别，对分类的结果进行一定程度上的验证。

引用网络：

引用网络来自于专利之间的引用关系，可以据此进行引用网络分析，相关性分析等。

### 1.1 关键技术查询

描述：给定一个术语词，检索得出相似度高的词，并给出这些词所在的专利号，尝试给出这个术语词词频（或专利数目）随着时间的变化趋势。

查询结果：

 相似度高的词（按照相似度大小排序）

 相似度高的词所在的专利号

 查询术语词的词频（或专利数目）随着时间变化趋势图

所用技术：

相似度计算：使用word2vec（或其他技术），获得词语的词向量，使用余弦距离（或其他距离计算公式）计算相似度。

输入关键技术术语，返回得到相对丰富的查询结果。=

结果内容包括：

* 关键技术本身介绍

从网页如维基百科摘取半结构化文档信息进行实体对齐等方式作为介绍。

* 关键技术主题分类

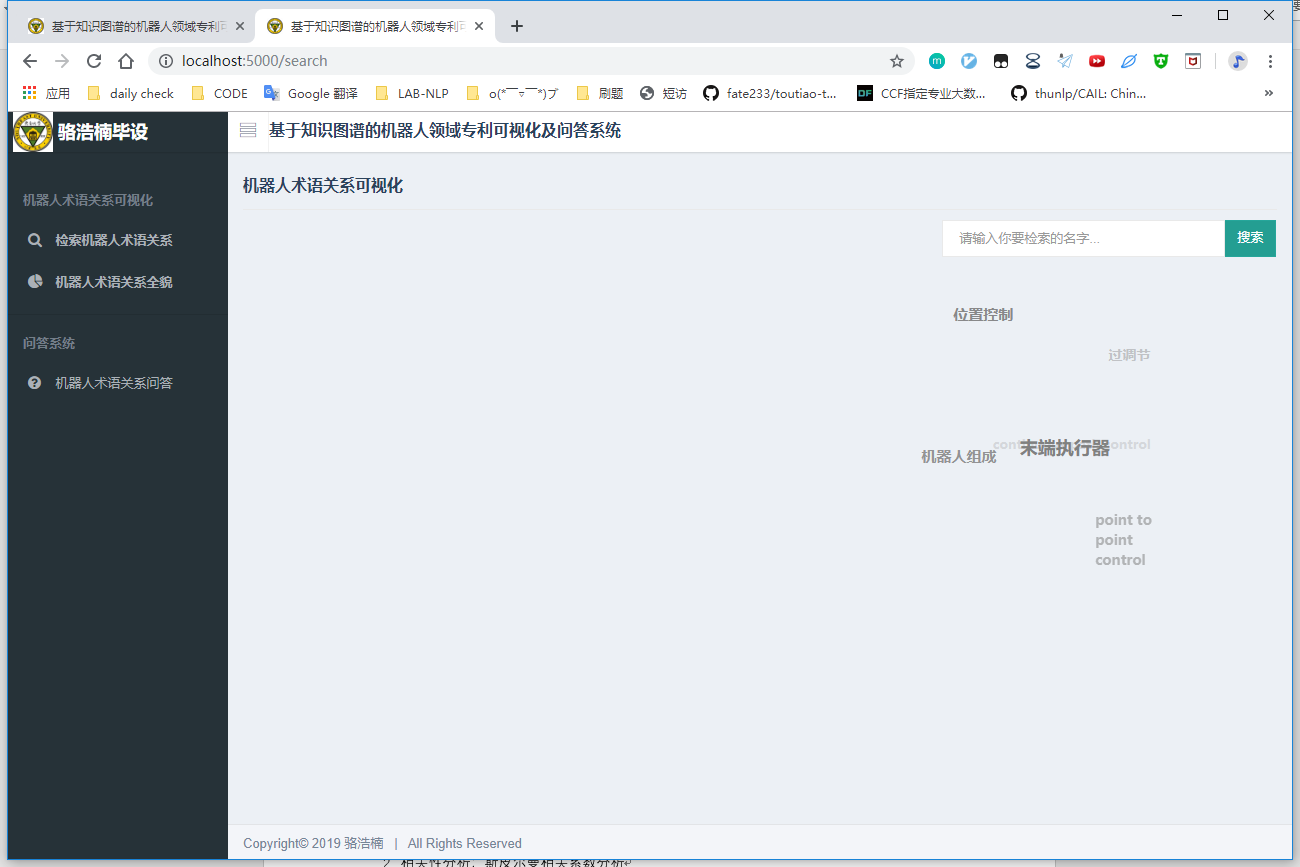
对于语料中目前得到的技术信息进行主题提取和分类，分类算法可采用深度学习方式不断进行优化。

* 相关专利发展演变趋势

通过时间维度对相关专利进行梳理，形成按照时间顺序的相关专利发展脉络，揭示该技术到今天为止的发展变化趋势。

* 相关专利引用关系

通过引用维度对相关专利进行梳理，形成专利引用树，揭示相关技术专利的发展变化情况。



### 1.2 关键技术之间的相似度查询和推荐

#### 1.2.1多个术语词相似度查询

描述：给定多个术语词，给出多个术语词之间的相似度，给出多个术语词的词频（或专利数目）随着时间变化趋势，并判断是否有相关性。

查询结果：

 多个术语词之间的相似度表

 多个术语词的词频（或专利数目）随着时间变化趋势图

 多个术语词随时间变换趋势是否相关性表

所用技术：

1. 相似度计算：使用word2vec（或其他技术），获得词语的词向量，使用余弦距离（或其他距离计算公式）计算相似度。

2. 相关性分析：斯皮尔曼相关系数分析

输入两个关键技术，返回得到相似度。

实现方式：基于在整个语料库中查询词频共现情况，在线计算，得到相似度结果输出。

* 单词的相似度，可直接通过余弦相似度计算得到；
* 短语的相似度计算，可通过识别得到的短语实体，通过返回文档重新生成词向量的方式，进行相似度计算。

#### 1.2.2相关技术术语推荐

根据相关专利和技术相似度的计算结果，对相关技术进行推荐。

实现方式：在后台预先计算当前提取得到的术语，每个术语最相关的其他技术术语词，并选择阈值进行存储。通过查询索引，返回每个词以及相似度。

### 1.3 专利文本查询

描述：给定一个专利文本摘要，进行主题词提取、关系抽取。给出相似专利的专利号和相似度。

查询结果：

 主题词、主题词之间关系

 专利所属主题（分类）

 相似专利的专利号和相似度

 专利主题的引用网络随着时间变化趋势

所用技术：

1. 主题词提取、关系抽取：使用NLTK工具。

2. 专利所属主题：根据预先使用（摘要，主题）数据对训练的模型，进行判断查询专利所属的专利主题。所用工具tensorflow，使用Lstm进行训练。

3. 主题引用网络分析：对各专利，进行引文数目和质量的分析。

### 1.4 问题解决回答

描述：给定一个关于机器人设计的自然语言问题，对问题进行结构化解析，从知识图谱中抽取出和问题相关的文本，进行语义解析，并最终给出答案。

查询结果：

 主题词提取、关系抽取

 相似专利的专利号和相似度

所用技术：

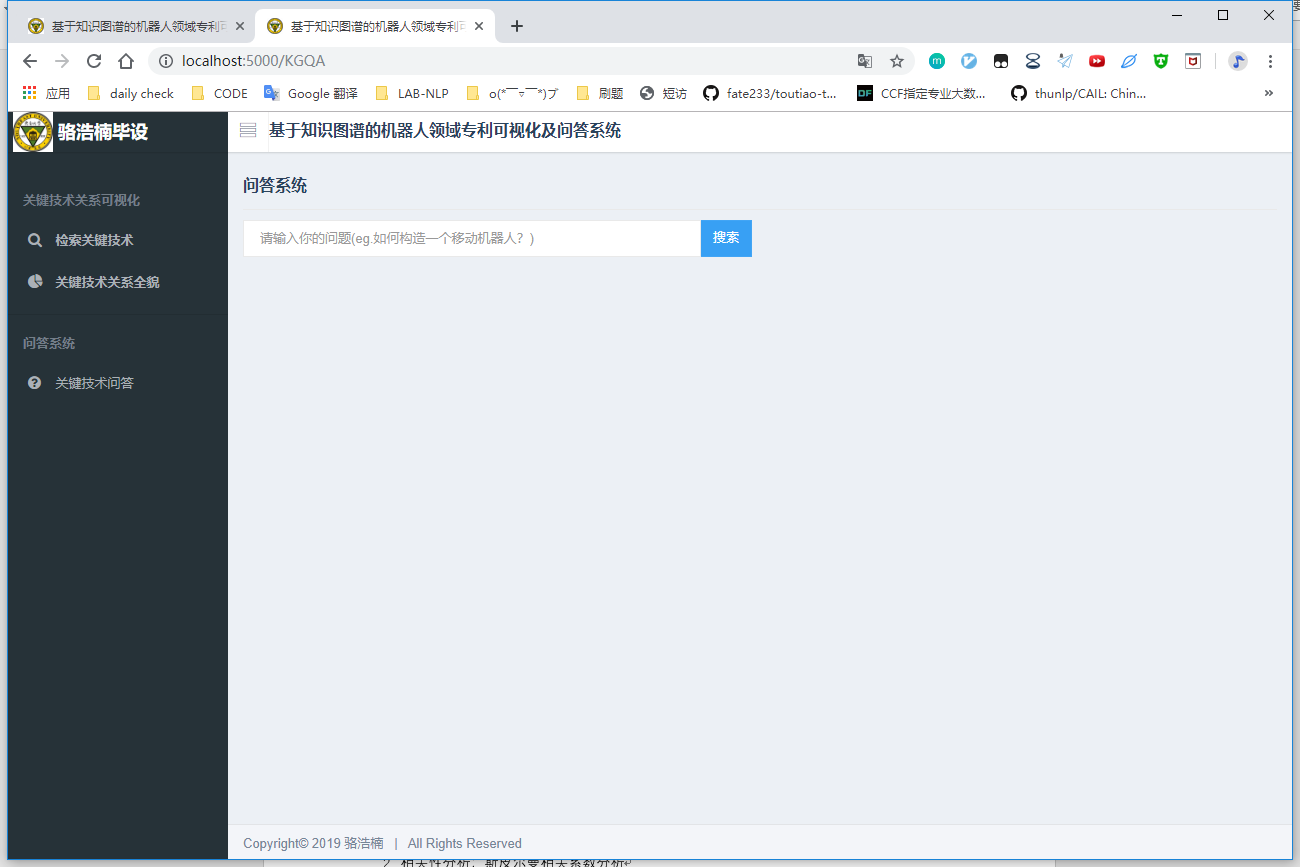
1. 自然语言问题解析：使用NLTK进行关系提取，并最终构成问题三元组。

2. 文本语义解析：从知识图谱中抽取出和问题相关的文本，使用NLTK进行关系提取。

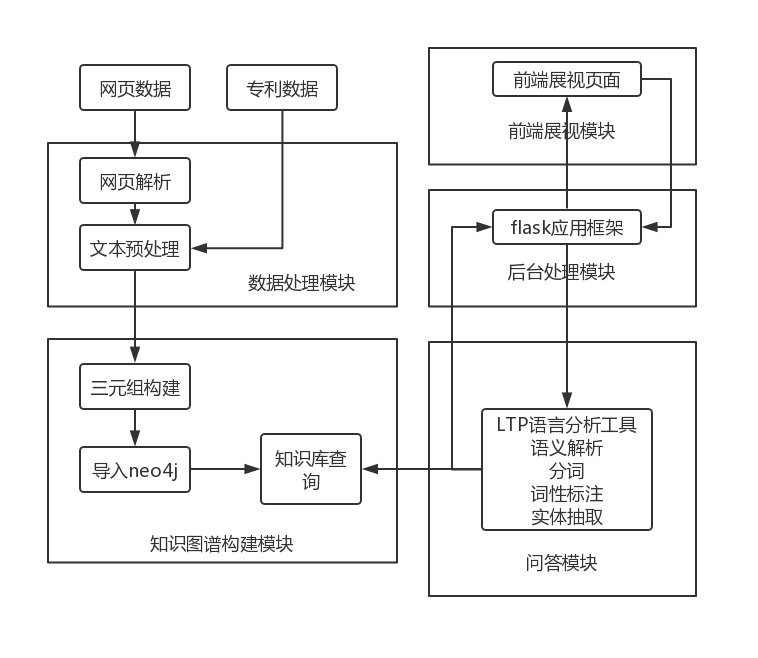
3. 对问题中的研究实体、方法或属性进行匹配，给出按照机器人手册层级结构中的实体、方法答案推荐。

输入问题文本，通过NER对其中的关键问题词进行识别，通过对识别得到的相关术语进行检索，返回得到关键技术查询结果。

通过文本生成返回得到该问题的参考解决方案。方案主要包括相关技术的专利展示。



## 2 机器人专利知识图谱构建过程



* 数据处理模块：

将网页数据和专利数据进行分词词性标注和去除停用词等预处理工作。

* 知识图谱构建模块：

将预处理得到的数据生成三元组，导入neo4j以供知识库查询。

* 后台处理模块：

使用python的flask应用框架，将知识图谱和前端连接结合。

* 前端展示模块：

展示机器人关系术语和知识图谱查询等页面。

* 问答模块：

将前端用户输入的问题，经过后台处理模块之后，进行分词和实体抽取等工作，进入知识库进行查询。

