第一阶段

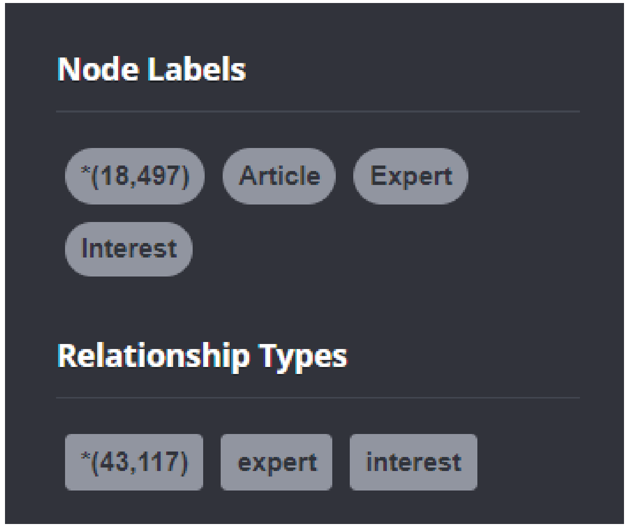
数据集

https://www.aminer.cn/scikg

https://www.aminer.cn/aminernetwork

数据预处理:

去除文本中的特殊字符

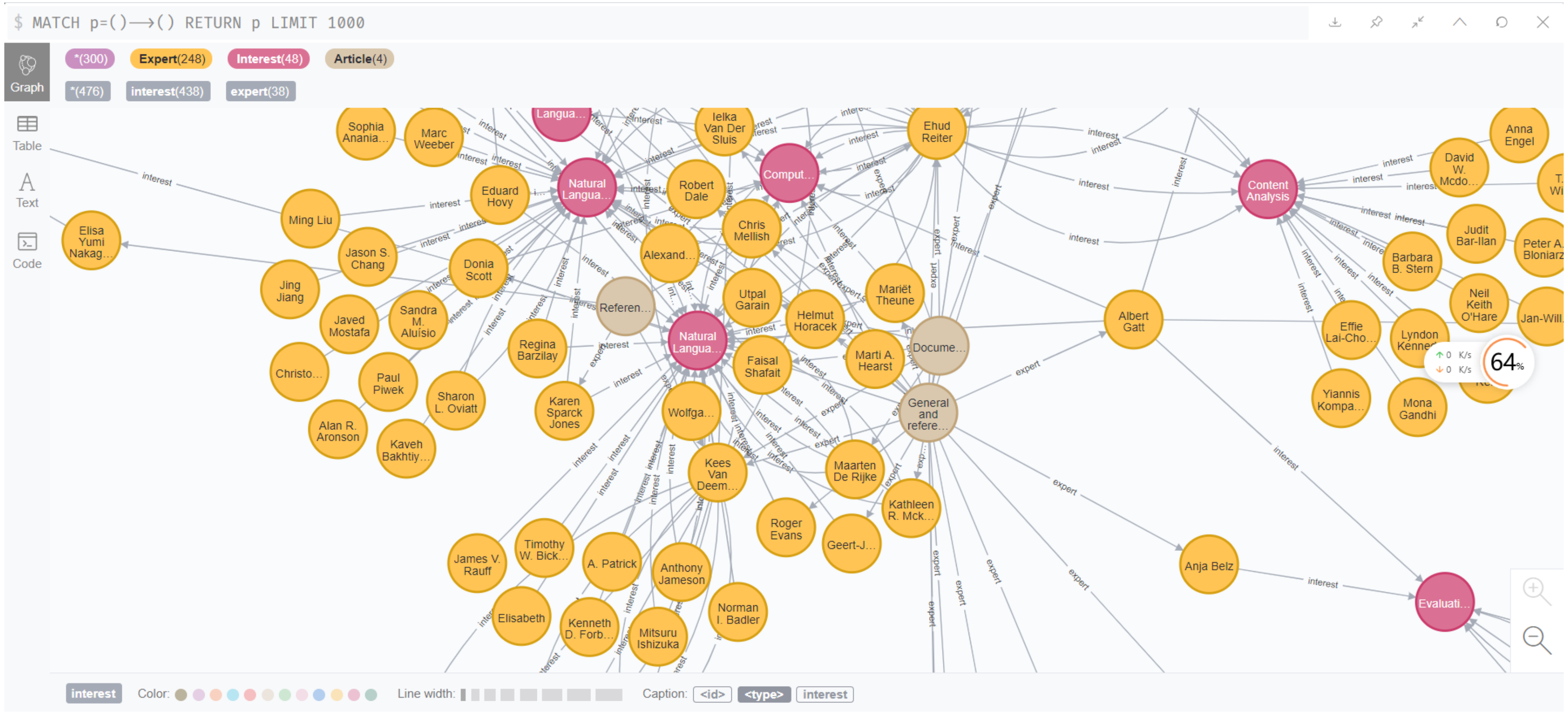


Node Labels

*(18,497) Article Expert Interest

Relationship Types

*(43,117) expert interest

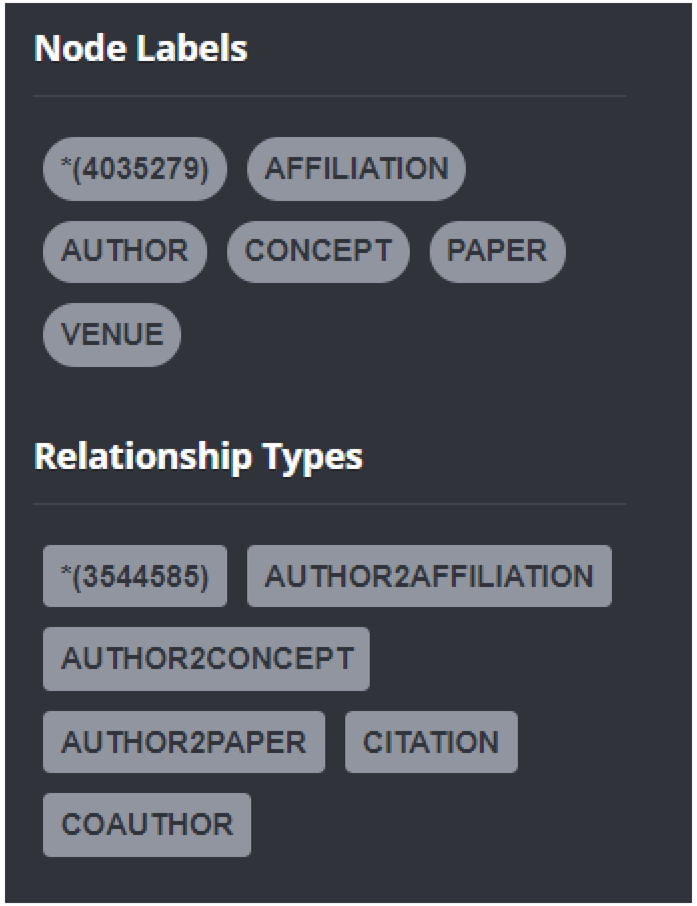


```
$ MATCH p=()⟶() RETURN p LIMIT 1000
```

*(300) Expert(248) Interest(48) Article(4)

*(476) interest(438) expert(38)

interest Color: Line width: Caption: <id> <type> interest



Node Labels

*(4035279) AFFILIATION

AUTHOR CONCEPT PAPER

VENUE

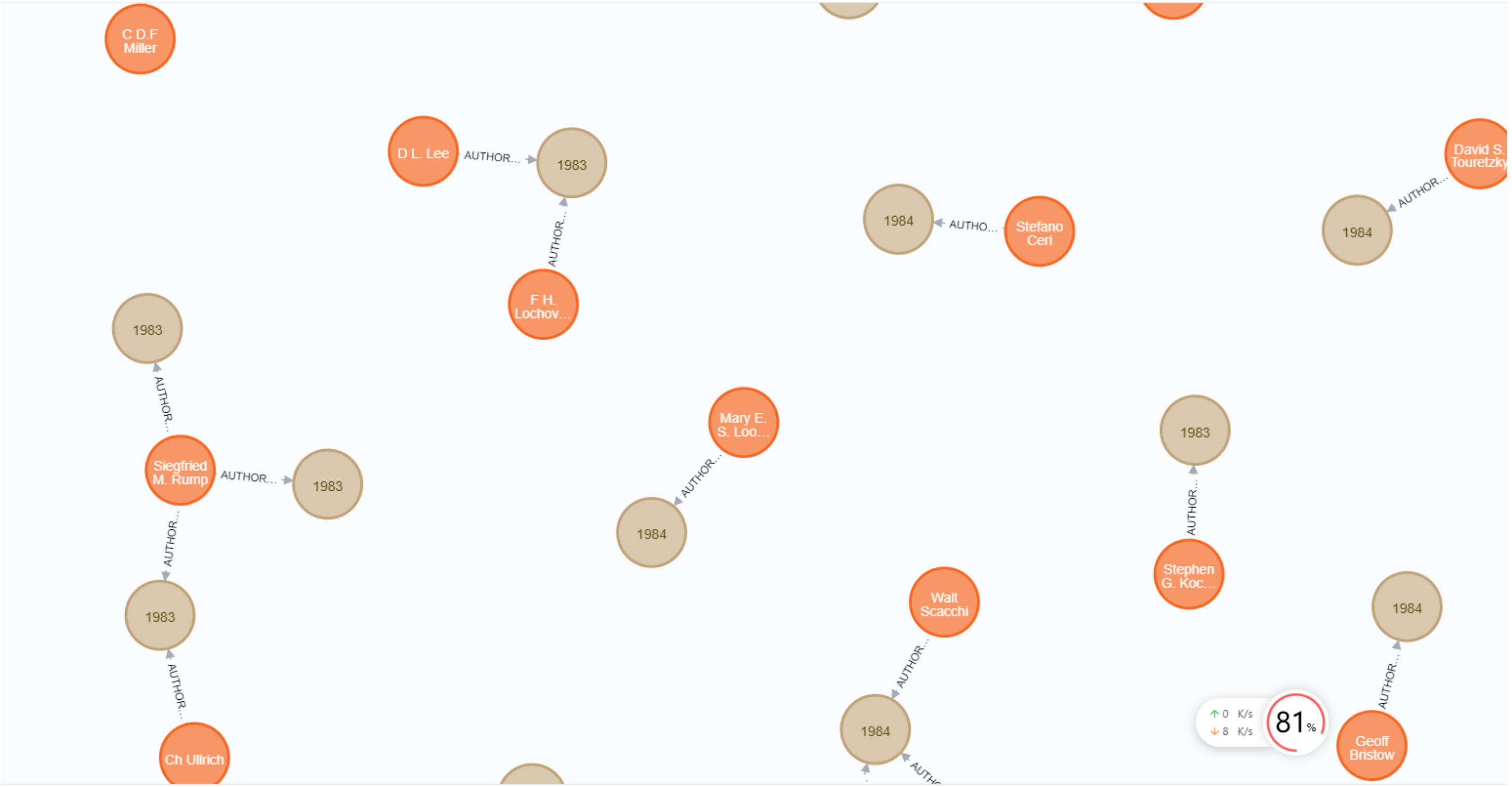
Relationship Types

*(3544585) AUTHOR2AFFILIATION

AUTHOR2CONCEPT

AUTHOR2PAPER CITATION

COAUTHOR



实体:
作者 author
机构 affiliation
知识概念 concept
论文 paper
发生地点 venue

关系
作者-机构
作者-概念
作者-论文
论文引用论文
合作
论文-发生地

数据三元组转化与连接

导入Neo4j

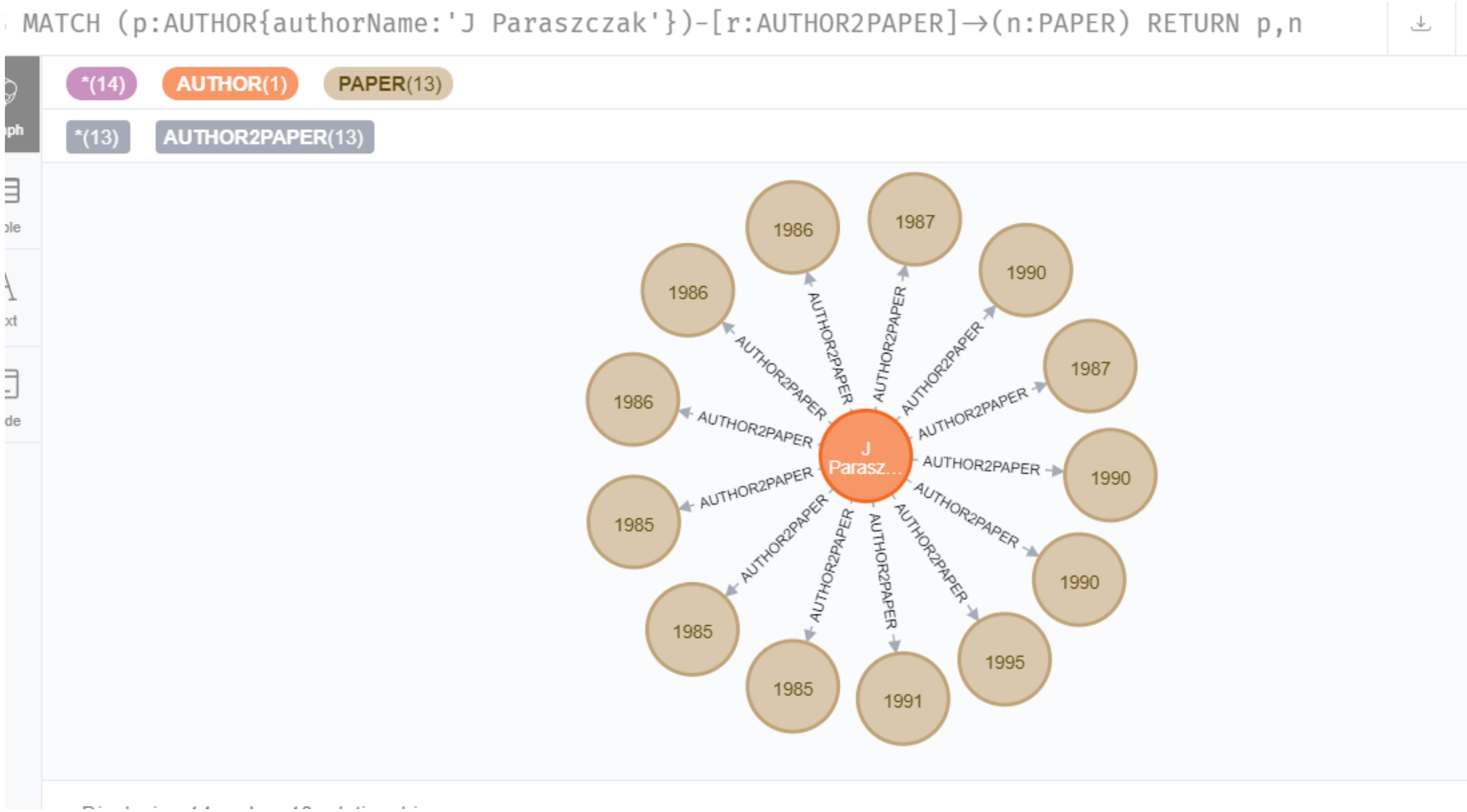
建立索引

To_do_list

- 1 知识图谱可视化
- 2 基于模板和语义的QA

找专家
找论文
找会议
找研究机构/实验室

找关键词 （专家的兴趣和论文的标题摘要）
找合作伙伴（find a person with same interests）



```
MATCH (p:AUTHOR{authorName:'J Paraszczak'})-[r:AUTHOR2PAPER]→(n:PAPER) RETURN p,n
```

- MATCH (p:AUTHOR{authorName:'J Paraszczak'})-[r:AUTHOR2PAPER]->(n:PAPER) RETURN p,n

↓

- ‘J Paraszczak发布了哪些论文

第二阶段

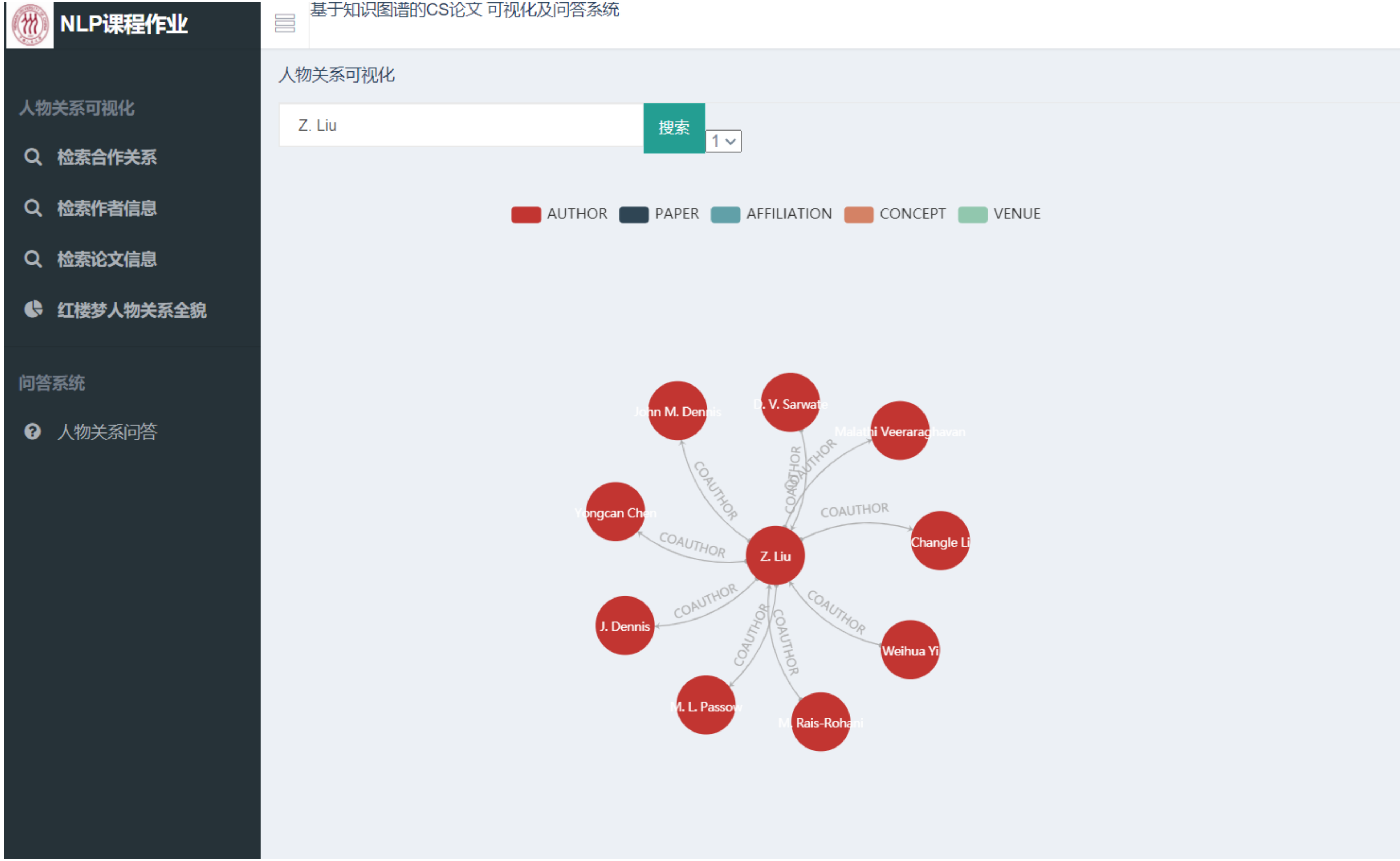
知识图谱可视化以及简单的问答系统

主要技术栈;

Flask
Neo4j
Echart
Elasticsearch

基于知识图谱的CS论文可视化及问答系统

合作关系

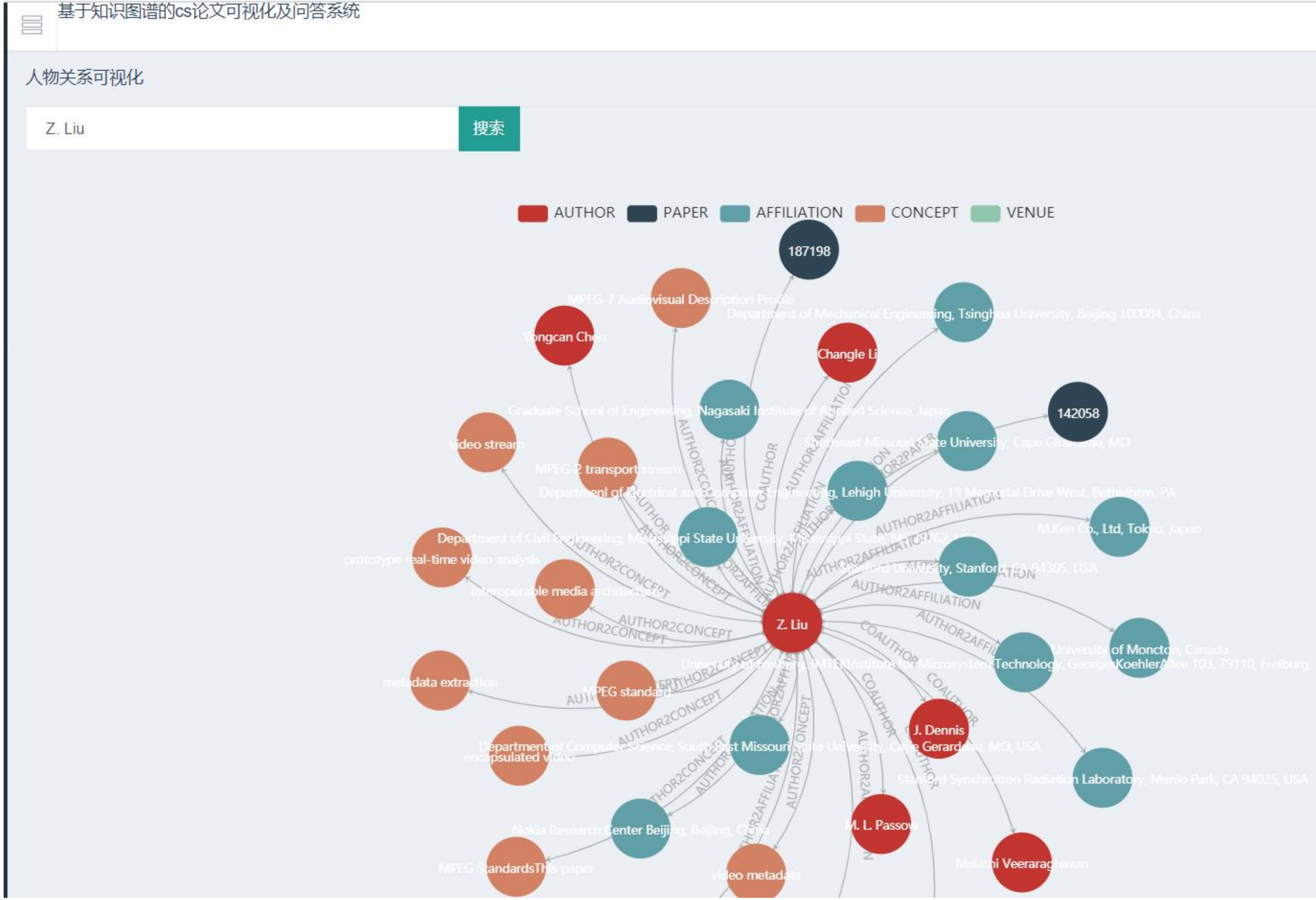


基于知识图谱的CS论文 可视化及问答系统

人物关系可视化

Z. Liu

信息查询

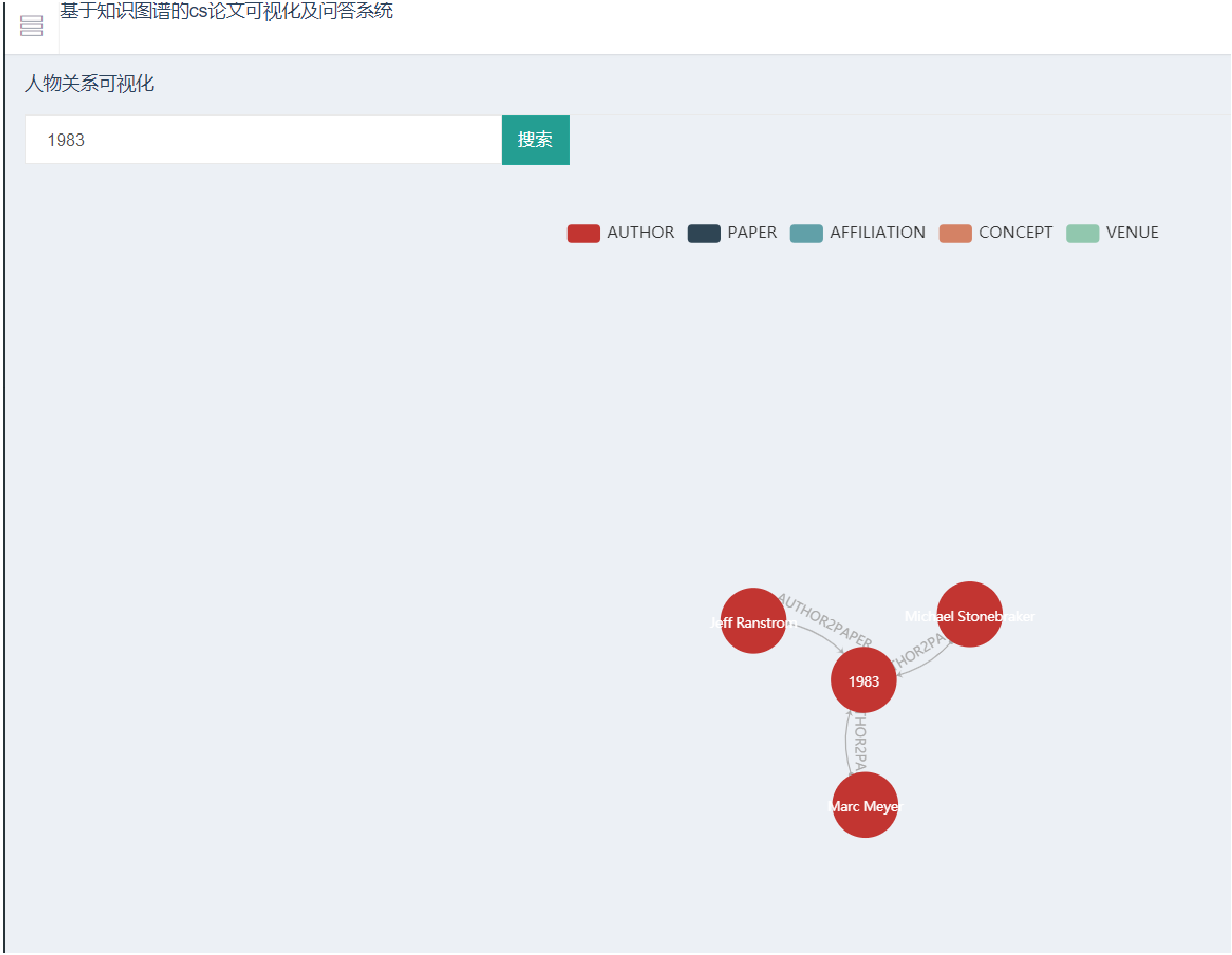


基于知识图谱的cs论文可视化及问答系统

人物关系可视化

Z. Liu

搜索
🟥 AUTHOR
⬛ PAPER
🟦 AFFILIATION
🟧 CONCEPT
🟩 VENUE
187198
MPEG-7 Audiovisual Description Profile
Department of Mechanical Engineering, Tsinghua University, Beijing 100084, China
Yongcan Chen
Changle Li
Graduate School of Engineering, Nagasaki Institute of Applied Science, Japan
142058
Southeast Missouri State University, Cape Girardeau, MO
video stream
MPEG-2 transport stream
Department of Electrical and Computer Engineering, Lehigh University, 19 Memorial Drive West, Bethlehem, PA
M.Ken Co., Ltd, Tokyo, Japan
Department of Civil Engineering, Mississippi State University, Mississippi State, MS 39762, USA
prototype real-time video analysis
Stanford University, Stanford, CA 94305, USA
interoperable media architecture
Z. Liu
University of Moncton, Canada
University of Freiburg, IMTEKInstitute for Microsystem Technology, GeorgesKoehlerAllee 103, 79110, Freiburg,
metadata extraction
MPEG standard
J. Dennis
Department of Computer Science, SouthEast Missouri State University, Cape Gerardeau, MO, USA
encapsulated video
Stanford Synchrotron Radiation Laboratory, Menlo Park, CA 94025, USA
Nokia Research Center Beijing, Beijing, China
M. L. Passow
MPEG StandardsThis paper
video metadata
Malathi Veeraraghavan
AUTHOR2CONCEPT
AUTHOR2AFFILIATION
COAUTHOR
AUTHOR2PAPER

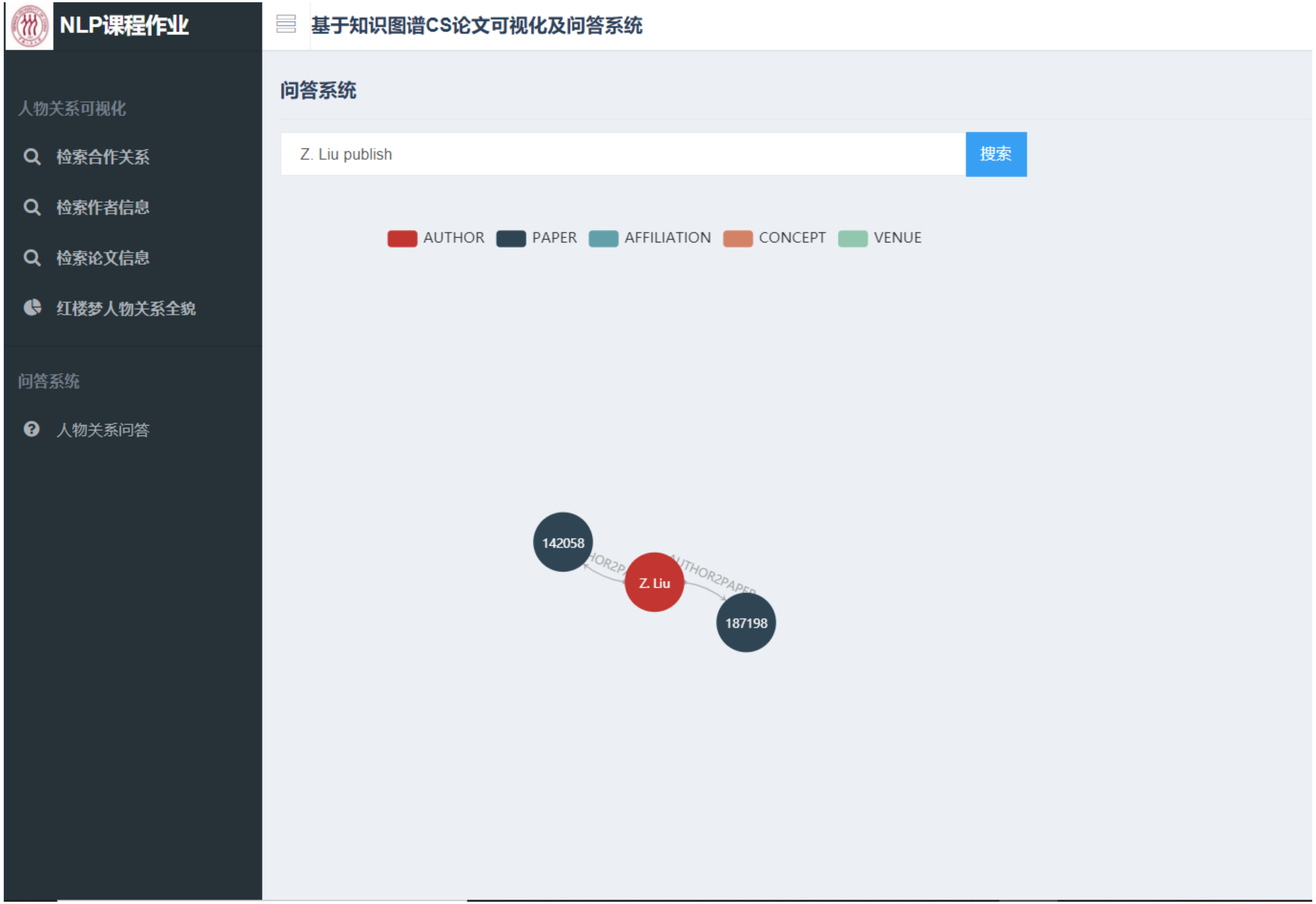


基于知识图谱的cs论文可视化及问答系统

人物关系可视化

1983

问答系统



Z. Liu publish

基于模板和机器学习的问答

1、中文 or 英文
2、同义词语料库 wordnet
3、词性标注和ner nltk or others
4、机器学习 or 深度学习

基于模板的方法

模板定义
模板生成
模板匹配

参考tbsl论文

模板匹配与实例化—实体识别与属性检测

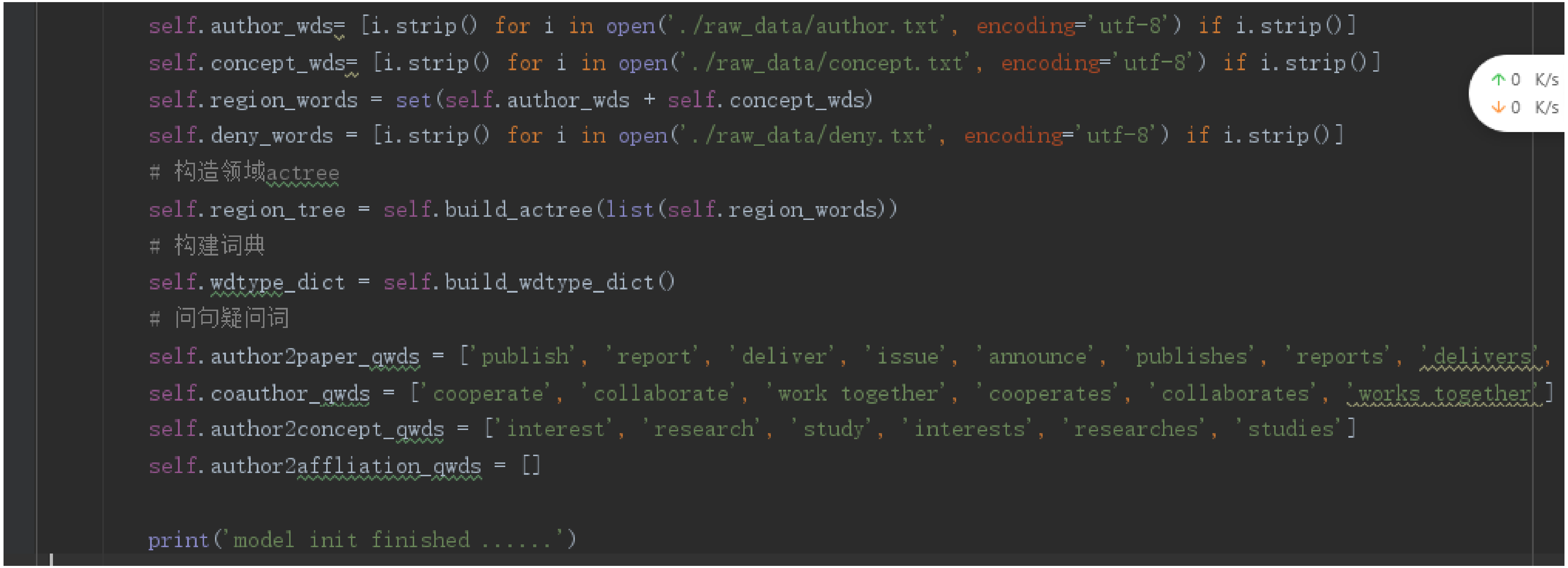
- □ 有了SPARQL模板以后，需要进行实例化与具体的自然语言问句相匹配。即将自然语言问句与知识库中的本体概念相映射的过程。
  - ■ 对于resources和classes,实体识别常用方法:
    - □ 用WordNet定义知识库中标签的同义词
    - □ 计算字符串相似度 (trigram, Levenshtein 和子串相似度)
  - ■ 对于property labels,将还需要与存储在BOA模式库中的自然语言表示进行比较
  - ■ 最高排位的实体将作为填充查询槽位的候选答案

TBSL主要缺点

- □ 创建的模板结构未必和知识图谱中的数据建模相契合

- □ 考虑到数据建模的各种可能性，对应到一个问题的潜在模板数量会非常多，同时手工准备海量模板的代价也非常大

Classify1:基于wordnet列举同义词

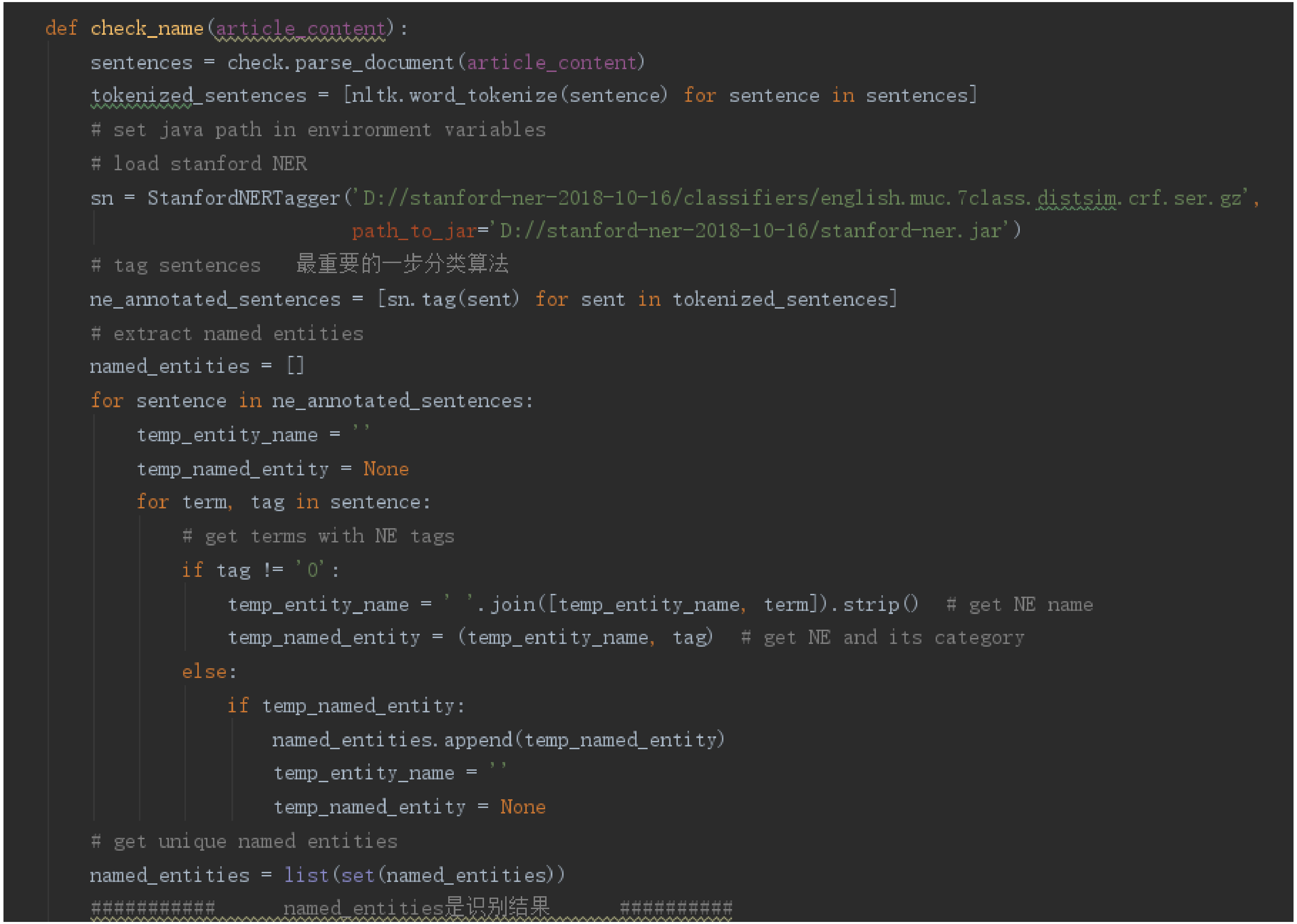


```
        self.author_wds= [i.strip() for i in open('./raw_data/author.txt', encoding='utf-8') if i.strip()]
        self.concept_wds= [i.strip() for i in open('./raw_data/concept.txt', encoding='utf-8') if i.strip()]
        self.region_words = set(self.author_wds + self.concept_wds)
        self.deny_words = [i.strip() for i in open('./raw_data/deny.txt', encoding='utf-8') if i.strip()]
        # 构造领域actree
        self.region_tree = self.build_actree(list(self.region_words))
        # 构建词典
        self.wdtype_dict = self.build_wdtype_dict()
        # 问句疑问词
        self.author2paper_qwds = ['publish', 'report', 'deliver', 'issue', 'announce', 'publishes', 'reports', 'delivers',
        self.coauthor_qwds = ['cooperate', 'collaborate', 'work together', 'cooperates', 'collaborates', 'works together']
        self.author2concept_qwds = ['interest', 'research', 'study', 'interests', 'researches', 'studies']
        self.author2affliation_qwds = []

        print('model init finished ......')
```

遇到的问题：需要先识别实体

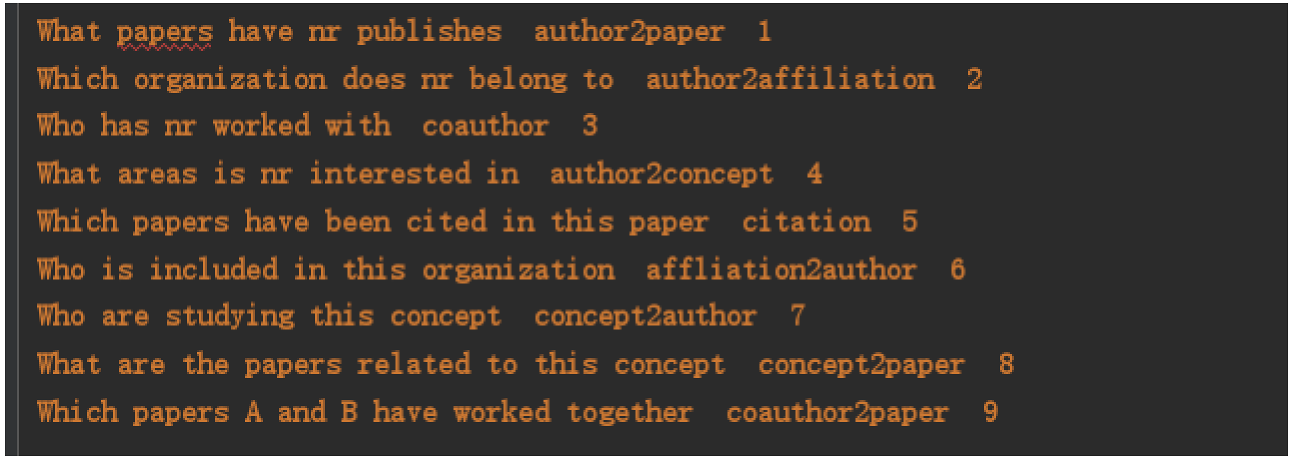
Nltk 和standfordnlp

比较两种方式，我们可以发现，NLTK下的命名实体识别更加倾向于分词和词性标准，虽然它也会将组织名，人名，地名等标注出来，但由于它把文件中的谓语，宾语等成分也标注了出来，造成了输出文本的冗余性，不利于读者很好的识别命名实体，需要我们对文本做进一步处理。NLTK下的命名实体识别的有点时，可以使用NLTK下的treebank包将文本绘制为树形，使结果更加清晰易读。相较而言，我更加倾向于Stanford的命名实体识别，它可以把Time, Location, Organization, Person, Money, Percent, Date七类实体很清晰的标注出来，而没有多余的词性。但由于NER是基于java开发的，所以在用python实现时可能由于jar包或是路径问题出现很多bug。

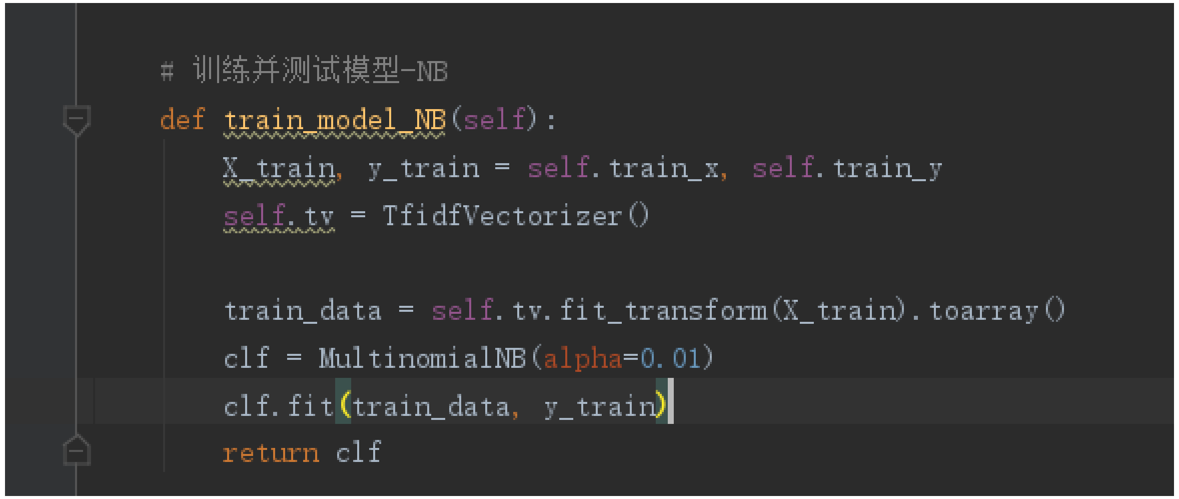


```
def check_name(article_content):
    sentences = check.parse_document(article_content)
    tokenized_sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
    # set java path in environment variables
    # load stanford NER
    sn = StanfordNERTagger('D://stanford-ner-2018-10-16/classifiers/english.muc.7class.distsim.crf.ser.gz',
                           path_to_jar='D://stanford-ner-2018-10-16/stanford-ner.jar')
    # tag sentences   最重要的一步分类算法
    ne_annotated_sentences = [sn.tag(sent) for sent in tokenized_sentences]
    # extract named entities
    named_entities = []
    for sentence in ne_annotated_sentences:
        temp_entity_name = ''
        temp_named_entity = None
        for term, tag in sentence:
            # get terms with NE tags
            if tag != 'O':
                temp_entity_name = ' '.join([temp_entity_name, term]).strip()  # get NE name
                temp_named_entity = (temp_entity_name, tag)  # get NE and its category
            else:
                if temp_named_entity:
                    named_entities.append(temp_named_entity)
                    temp_entity_name = ''
                    temp_named_entity = None
    # get unique named entities
    named_entities = list(set(named_entities))
    ###########     named_entities是识别结果     ##########
```

Classfy2:基于朴素贝叶斯的机器学习分类方法

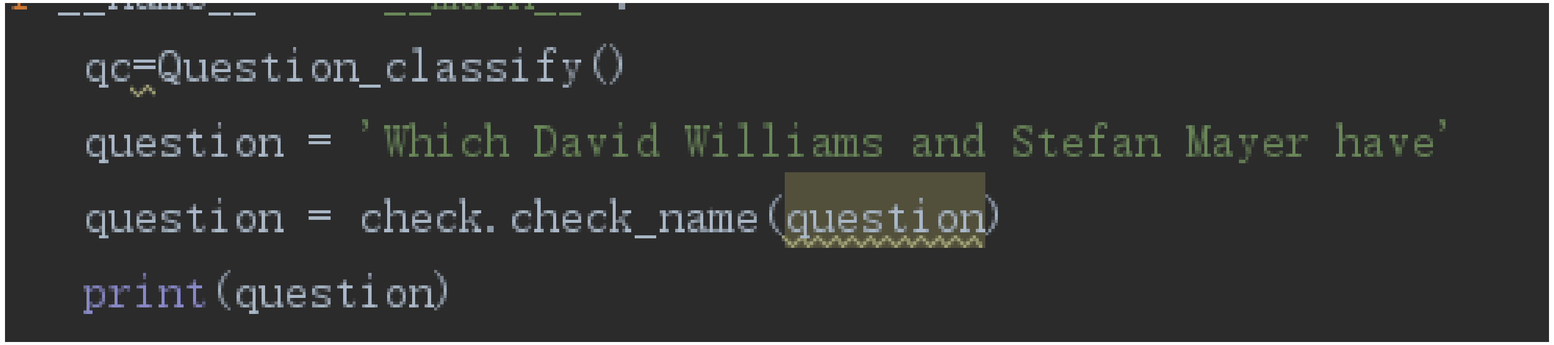


```
What papers have nr publishes  author2paper  1
Which organization does nr belong to  author2affiliation  2
Who has nr worked with  coauthor  3
What areas is nr interested in  author2concept  4
Which papers have been cited in this paper  citation  5
Who is included in this organization  affliation2author  6
Who are studying this concept  concept2author  7
What are the papers related to this concept  concept2paper  8
Which papers A and B have worked together  coauthor2paper  9
```



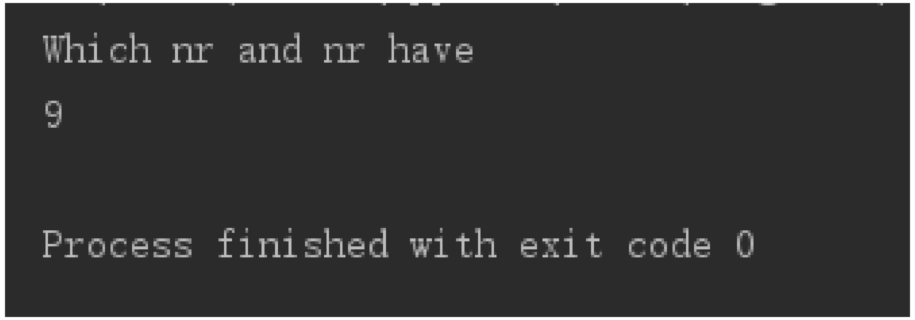
```
# 训练并测试模型-NB
def train_model_NB(self):
    X_train, y_train = self.train_x, self.train_y
    self.tv = TfidfVectorizer()

    train_data = self.tv.fit_transform(X_train).toarray()
    clf = MultinomialNB(alpha=0.01)
    clf.fit(train_data, y_train)
    return clf
```

Input:



```
  qc=Question_classify()
  question = 'Which David Williams and Stefan Mayer have'
  question = check.check_name(question)
  print(question)
```

Output:



```
Which nr and nr have
9

Process finished with exit code 0
```

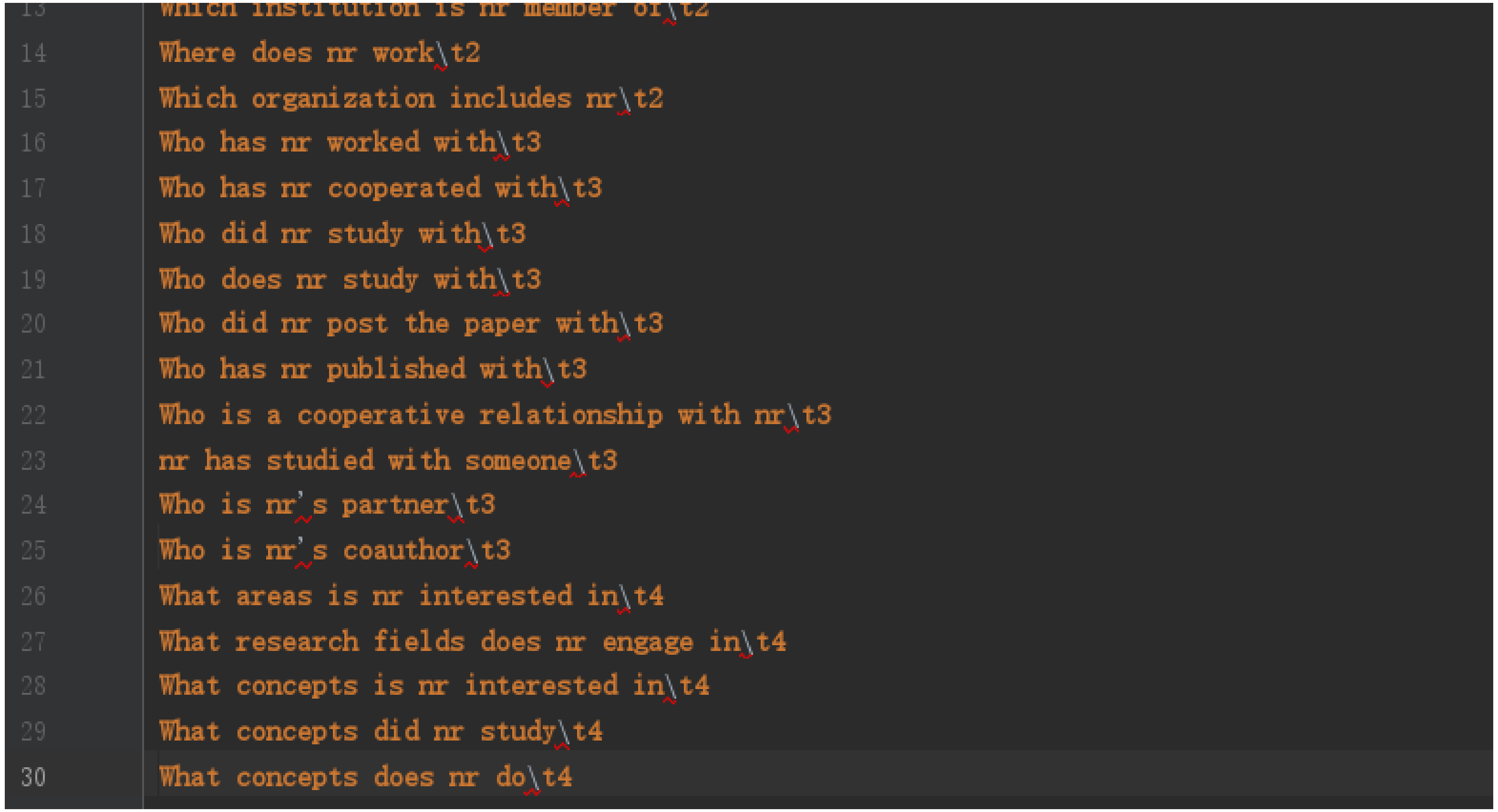
To_do_list

1、增加机器学习的样本量
2、尝试使用其他机器学习算法以及深度学习算法

第三阶段

扩充问答系统

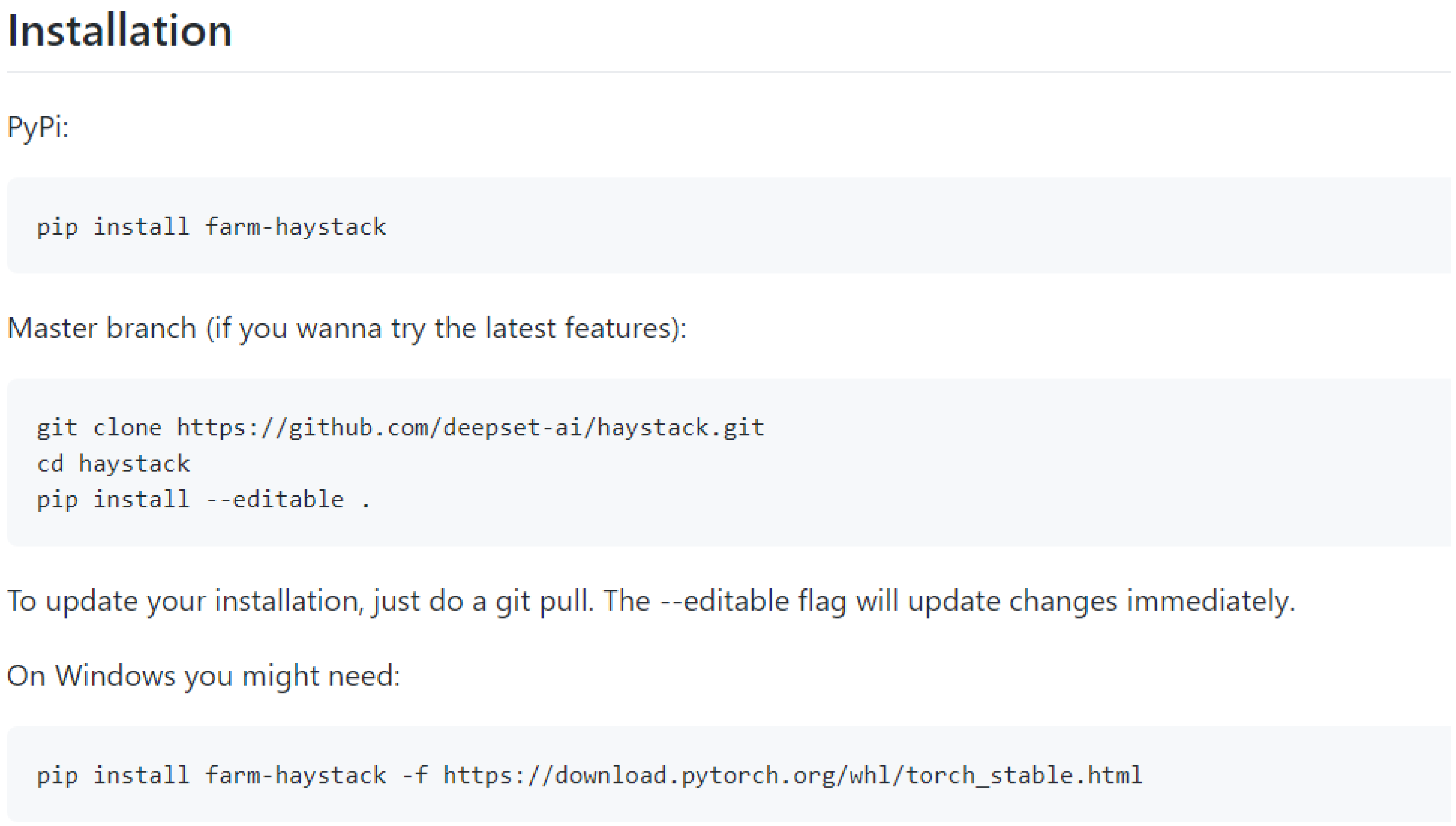
增加机器学习样本量



```
Which institution is nr member of\t2
Where does nr work\t2
Which organization includes nr\t2
Who has nr worked with\t3
Who has nr cooperated with\t3
Who did nr study with\t3
Who does nr study with\t3
Who did nr post the paper with\t3
Who has nr published with\t3
Who is a cooperative relationship with nr\t3
nr has studied with someone\t3
Who is nr’s partner\t3
Who is nr’s coauthor\t3
What areas is nr interested in\t4
What research fields does nr engage in\t4
What concepts is nr interested in\t4
What concepts did nr study\t4
What concepts does nr do\t4
```

https://github.com/facebookresearch/DrQA

https://github.com/deepset-ai/haystack



Installation

PyPi:

```
pip install farm-haystack
```

Master branch (if you wanna try the latest features):

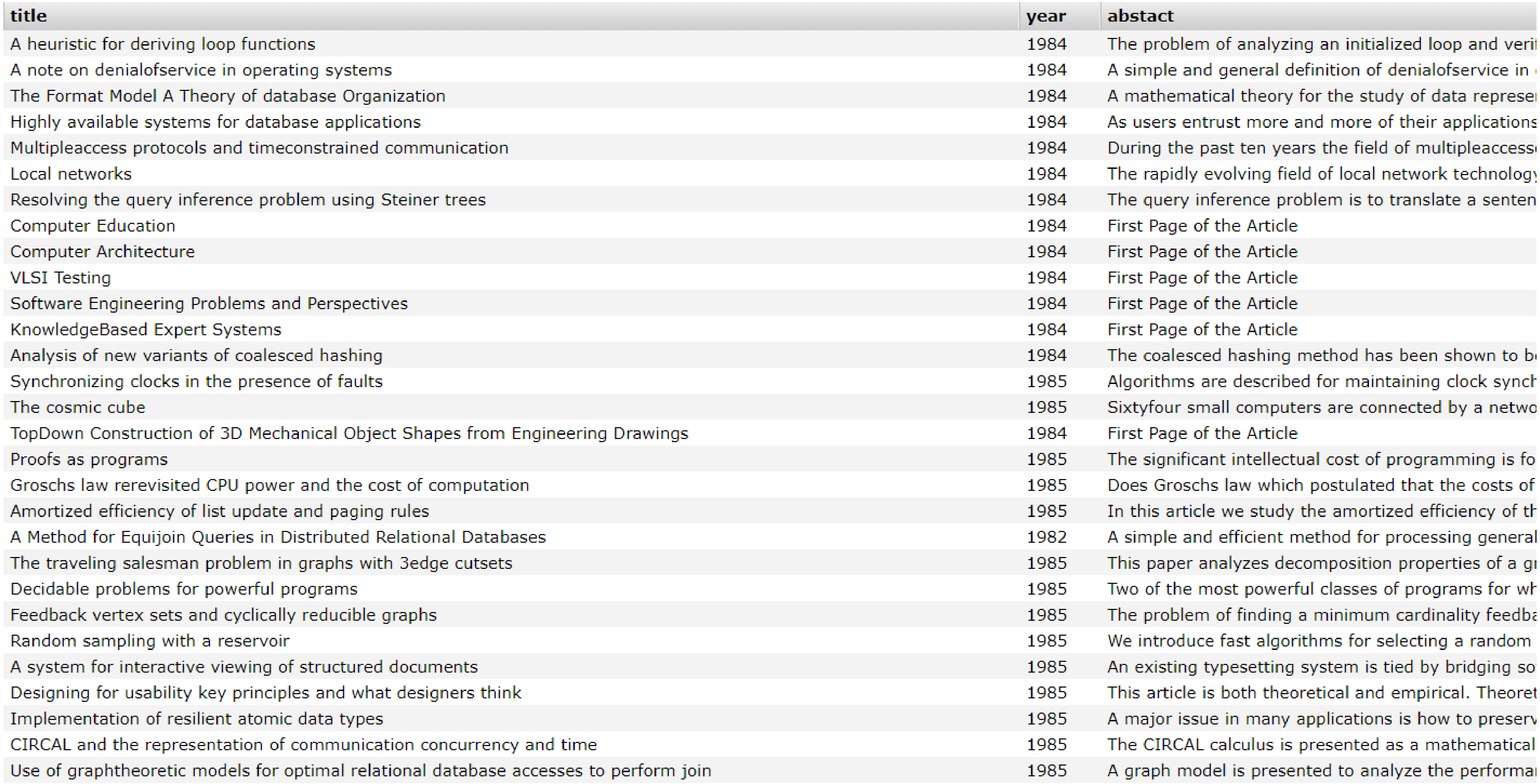
```
git clone https://github.com/deepset-ai/haystack.git
cd haystack
pip install --editable .
```

To update your installation, just do a git pull. The --editable flag will update changes immediately.

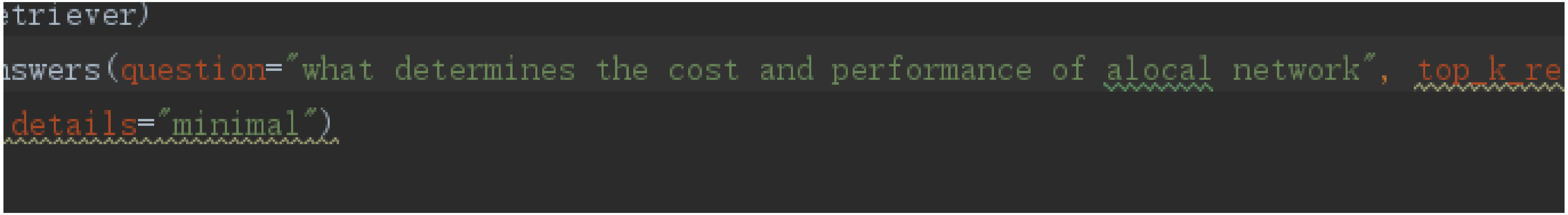
On Windows you might need:

```
pip install farm-haystack -f https://download.pytorch.org/whl/torch_stable.html
```

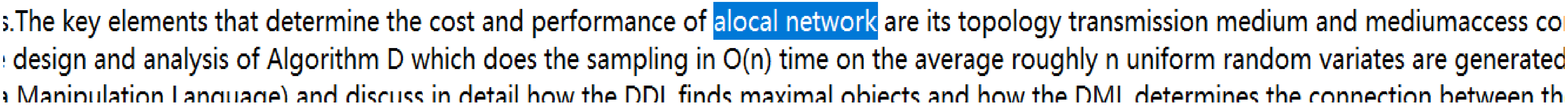
elasticsearch



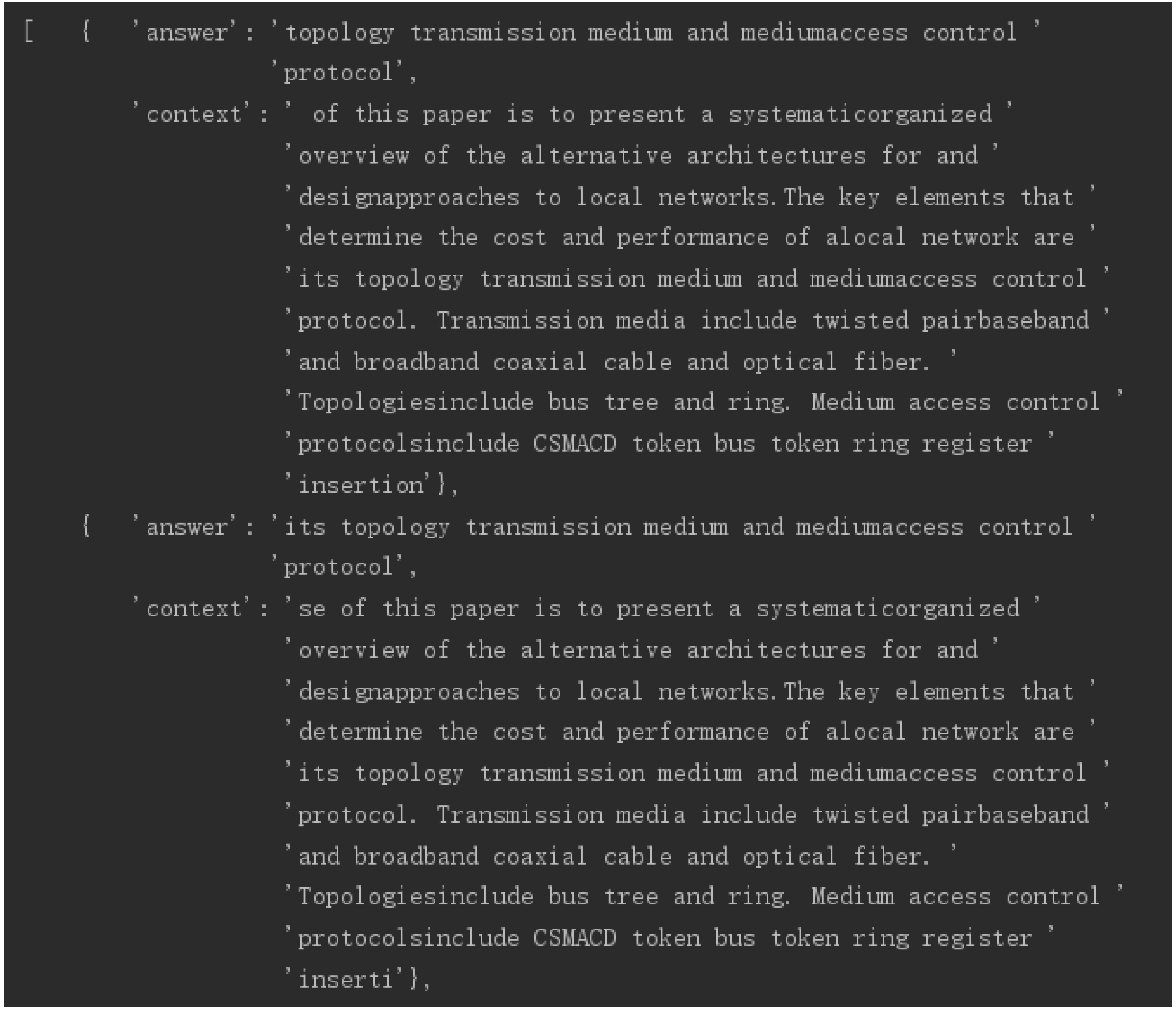
title	year	abstact
A heuristic for deriving loop functions	1984	The problem of analyzing an initialized loop and veri
A note on denialofservice in operating systems	1984	A simple and general definition of denialofservice in
The Format Model A Theory of database Organization	1984	A mathematical theory for the study of data represe
Highly available systems for database applications	1984	As users entrust more and more of their applications
Multipleaccess protocols and timeconstrained communication	1984	During the past ten years the field of multipleaccess
Local networks	1984	The rapidly evolving field of local network technolog
Resolving the query inference problem using Steiner trees	1984	The query inference problem is to translate a senten
Computer Education	1984	First Page of the Article
Computer Architecture	1984	First Page of the Article
VLSI Testing	1984	First Page of the Article
Software Engineering Problems and Perspectives	1984	First Page of the Article
KnowledgeBased Expert Systems	1984	First Page of the Article
Analysis of new variants of coalesced hashing	1984	The coalesced hashing method has been shown to b
Synchronizing clocks in the presence of faults	1985	Algorithms are described for maintaining clock synch
The cosmic cube	1985	Sixtyfour small computers are connected by a netwo
TopDown Construction of 3D Mechanical Object Shapes from Engineering Drawings	1984	First Page of the Article
Proofs as programs	1985	The significant intellectual cost of programming is fo
Groschs law rerevisited CPU power and the cost of computation	1985	Does Groschs law which postulated that the costs of
Amortized efficiency of list update and paging rules	1985	In this article we study the amortized efficiency of th
A Method for Equijoin Queries in Distributed Relational Databases	1982	A simple and efficient method for processing general
The traveling salesman problem in graphs with 3edge cutsets	1985	This paper analyzes decomposition properties of a g
Decidable problems for powerful programs	1985	Two of the most powerful classes of programs for wh
Feedback vertex sets and cyclically reducible graphs	1985	The problem of finding a minimum cardinality feedb
Random sampling with a reservoir	1985	We introduce fast algorithms for selecting a random
A system for interactive viewing of structured documents	1985	An existing typesetting system is tied by bridging so
Designing for usability key principles and what designers think	1985	This article is both theoretical and empirical. Theoret
Implementation of resilient atomic data types	1985	A major issue in many applications is how to preserv
CIRCAL and the representation of communication concurrency and time	1985	The CIRCAL calculus is presented as a mathematical
Use of graphtheoretic models for optimal relational database accesses to perform join	1985	A graph model is presented to analyze the performa



```
triever)
swers(question="what determines the cost and performance of alocal network", top_k_re
 details="minimal")
```



.The key elements that determine the cost and performance of alocal network are its topology transmission medium and mediumaccess co
design and analysis of Algorithm D which does the sampling in O(n) time on the average roughly n uniform random variates are generated
Manipulation Language) and discuss in detail how the DDL finds maximal objects and how the DML determines the connection between th



```
[   {   'answer': 'topology transmission medium and mediumaccess control '
                  'protocol',
        'context': ' of this paper is to present a systematicorganized '
                   'overview of the alternative architectures for and '
                   'designapproaches to local networks.The key elements that '
                   'determine the cost and performance of alocal network are '
                   'its topology transmission medium and mediumaccess control '
                   'protocol. Transmission media include twisted pairbaseband '
                   'and broadband coaxial cable and optical fiber. '
                   'Topologiesinclude bus tree and ring. Medium access control '
                   'protocolsinclude CSMACD token bus token ring register '
                   'insertion'},
    {   'answer': 'its topology transmission medium and mediumaccess control '
                  'protocol',
        'context': 'se of this paper is to present a systematicorganized '
                   'overview of the alternative architectures for and '
                   'designapproaches to local networks.The key elements that '
                   'determine the cost and performance of alocal network are '
                   'its topology transmission medium and mediumaccess control '
                   'protocol. Transmission media include twisted pairbaseband '
                   'and broadband coaxial cable and optical fiber. '
                   'Topologiesinclude bus tree and ring. Medium access control '
                   'protocolsinclude CSMACD token bus token ring register '
                   'inserti'},
```

To_do_list

后续的界面完善