

Group 17 Course Knowledge Graphs Report

Zi Gu zigu@usc.edu

Haili Wang hailiwan@usc.edu

Contents

1. Building Knowledge Graphs

1. Web Crawling - USC, UCLA, UCB
2. Ontology
3. HTMLs to Json-LD
4. Entity Matching - Instructor Matching
5. PDF Extraction - Syllabus => email & office hour

2. Elasticsearch & GUI App

1. Elasticsearch
2. Python GUI - Tkinter, Json Viewer, dot, graphviz

1. Building Knowledge Graphs

1.1 Web Crawling - USC, UCLA, UCB

Base URL

USC: <https://classes.usc.edu/term-20191/>

UCLA: <https://www.registrar.ucla.edu/Academics/Course-Descriptions>

UCB: <http://guide.berkeley.edu/courses/>

Extraction Details - USC

Classes Offered

group by school

show all programs

sort by prefix

Academic Medicine

ACMD

Accounting

ACCT

Advanced Dental Education Conjoint Program

ADNT

Aerospace and Mechanical Engineering

AME

Aerospace Studies

AEST

American Language Institute

ALI

American Studies and Ethnicity

AMST

Anesthesiology

ANST

Academic Medicine (ACMD)


+ expand all

- collapse all

- ACMD 503: Leading Change in Academic Medical Centers (3.0 units)

Exploration and practice of skills for promoting programs within academic medicine and health professions' education; building trust, organizational change, conflict resolution, negotiation, and managing resources.


● **Restriction:** Registration open to the following major(s): Academic Medicine

Section	Session	Type	Time	Days	Registered	Instructor	Location	Syllabus	Info
40254R	098	Lecture	5:00-8:00pm	Wednesday	9 of 30	Jerry Gates, Samuel Yanofsky, Julie Nyquist			

- ACMD 511: Competencies in Academic Medicine and Health I (3.0 units)

Acquisition of cognitive knowledge and problem solving skills in health professions worldwide; instructional methods, assessment techniques, designing curricula for health professions education.

● **Restriction:** Registration open to the following major(s): Academic Medicine

Section	Session	Type	Time	Days	Registered	Instructor	Location	Syllabus	Info
40264R	098	Lecture	5:00-8:00pm	Thursday	9 of 30	Julie Nyquist, Cathy Jalali			

+ ACMD 512: Competencies in Academic Medicine and Health II (3.0 units)

+ ACMD 592: Implementing Research on Innovation in Academic Medicine (2.0 units)

Extraction Details - UCLA

Browse By Subject Area

- A
- B
- C
- D
- E
- F
- G
- H
- I
- J
- K
- L
- M
- N
- O
- P
- R
- S
- T
- U
- V
- W
- Y

A

- Aerospace Studies
- African American Studies
- African Studies
- Afrikaans
- American Indian Studies
- American Sign Language
- Ancient Near East
- Anesthesiology
- Anthropology
- Applied Linguistics
- Arabic
- Archaeology
- Architecture and Urban Design
- Armenian
- Art
- Art History
- Arts and Architecture
- Arts Education
- Asian
- Asian American Studies
- Astronomy
- Atmospheric and Oceanic Sciences

B

- Bioengineering
- Bioinformatics (Graduate)
- Biomathematics
- Biomedical Research

Home / Academics / Course Descriptions / Course Details

AEROSPACE STUDIES (AERO ST)

Lower Division Courses Upper Division Courses

A. Leadership Laboratory


Units: 0
(Formerly numbered Z.) Laboratory, three hours. Mandatory for and limited to Air Force ROTC cadets. Provides cadets with practical command and staff leadership experiences through performance of various tasks within framework of organized cadet corps. As integral part of aerospace studies curriculum, provides experiences designed to develop leadership potential and serves as orientation to active duty. P/NP grading.

1A. Heritage and Values

Units: 2
Lecture, one hour. Introduction to U.S. Air Force. Examination of general aspects of Department of Air Force, leadership, benefits, and opportunities for officers. Foundation for becoming airmen by outlining heritage and values. Provides historical perspective through lessons on war and U.S. military, Air Force operations, principles of war, and airpower. Provides students with understanding for employment of air and space power, from institutional, doctrinal, and historical perspective. Students are introduced to Air Force way of life and gain knowledge on what it means to be airmen. P/NP or letter grading.

Courses

The course catalog below includes all courses currently approved to be taught at UC Berkeley.

Please Note: Only a subset of courses that appear are offered each semester. To see a list of course offerings for a current or future term, please see the [Class Schedule](#). 

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X	Y	Z							

A

- Aerospace Studies (AEROSPC)
- African American Studies (AFRICAM)
- Agricultural and Resource Economics (A,RESEC)
- American Studies (AMERSTD)
- Ancient History and Mediterranean Archaeology (AHMA)

Aerospace Studies (AEROSPC)

[Expand all course descriptions \[+\]](#)

AEROSPC 1A Foundations of the U.S. Air Force 1 Unit [-]

Terms offered: Fall 2019, Spring 2019, Fall 2018

This course introduces students to the United States Air Force (USAF) and Air Force Reserve Officer Training Corps (AFROTC) with an overview of the basic characteristics, missions, and organization of the Air Force; additional topics include officership and professionalism, Air Force career opportunities, military customs and courtesies, and an introduction to USAF basic communication skills. Additionally, AFROTC cadets must attend weekly Leadership

[Read More \[+\]](#)

AEROSPC 1B Foundations of the U.S. Air Force 1 Unit [-]

Terms offered: Spring 2019, Spring 2018, Spring 2017

A survey course designed to introduce cadets to the U.S. Air Force and the Air Force Reserve Officer Training Corps (AFROTC). Featured topics include the history and structure of the U.S. Air Force, the Air Force's capabilities, career opportunities, benefits, Air Force installations, and communications skills. Additionally, AFROTC cadets must attend Leadership Lab. Leadership Lab is a weekly laboratory that touches on the topics of Air Force

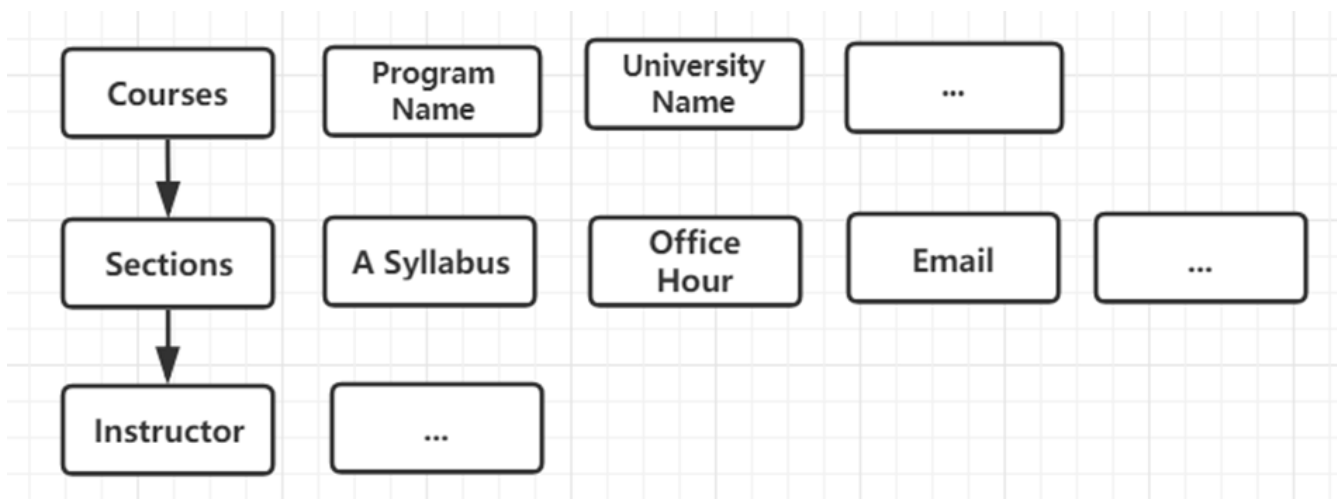
[Read More \[+\]](#)

AEROSPC 2A The Evolution of U.S. Air Force Air and Space Power 1 Unit [+]

1.2 Ontology

Overview

Use three main objects with attributes and links to each other. Implement by Protege.



Object Details

Course: Course Abbreviation, Course Name, Description, Program Name, University Name, Section IDs

Section: Time, Location, Syllabus Link, Email, Session Type, Office Hour, Instructor ID

Instructor: Instructor, Name, Instructor Link

1.3 HTMLs to Json-LD

Build python scripts convert html files to Json-LD format.

```

1 {
2   "@context": {
3     "ckb": "http://www.semanticweb.org/rice/ontologies/2019/3/CourseKB",
4     "xsd": "http://www.w3.org/2001/XMLSchema"
5   },
6   @graph: [
7     // list of Course, Section Objects so far. Instructors are added later.
8   ]
9 }

```

1.4 Entity Matching - Instructor Matching

Each section in previous Json-LD contains a individual Instructor object. This step matches Instructors by similarity of Instructor name. Extract Instructor object from Section object, add unique IDs, and put in @graph list. Use InstructorID in Section object as linkage. Here, three main objects, Course, Section, Instructor are listed properly in @graph.

Similarity function of Instructor name is based on edit distance.

1.5 PDF Extraction - Syllabus => email & office hour

Instructor Email and office hour are extracted from PDF syllabus. PyPDF2 is used to convert PDF files to text. Then apply regular expression to extract email and office hour.

```
1 # office hour regular expression
2 r'(?i)office hour[s]?[:]?[\s\n]*?([\d\w\-\.\,\:\t]+[\d]*)'
3
4 # email
5 r'[e|E][-]?[m|M]ail[:]?[\s]*?([\w\.-]+@[w\.-]+)
```

```
{
  "@id": "_:96",
  "@type": "ckb:Section",
  "ckb:Days": "Friday",
  "ckb:InstructorID": "_:17987",
  "ckb:OfficeHour": "Friday",
  "ckb:Syllabus": "PDF",
  "ckb:SyllabusLink": "https://web-app.usc.edu/soc/syllabus/20191/60558.pdf",
  "ckb:Time": "11:00-12:50pm",
  "ckb:email": "marisama@usc.edu",
  "ckb:sectionID": "60558D",
  "ckb:sessionID": "001",
  "ckb:sessionLink": "https://classes.usc.edu/term-20191/session/001/",
  "ckb:sessionType": "Lecture"
},
```

Elasticsearch & GUI App

2.1 Elasticsearch

Data Deployment

Based on previous Json-LD structure, we build a python script to extract objects in @graph and build index in Elasticsearch. Objects from three universities are list together. So finally, there is an index which contains a list of Course, Section, Instructor objects from three universities.

2.2 Python GUI - Tkinter, Json Viewer, dot, graphviz

Screen Shots - Search Window

tk

Breath Search by

id
 Program
 Course
 Section
 Instructor

Depth Search by

Course
 Section
 Instructor

Screen Shots - Search Result - Json Viewer

tk

Breath Search by

id
 Program
 Course
 Section
 Instructor

Depth Search by

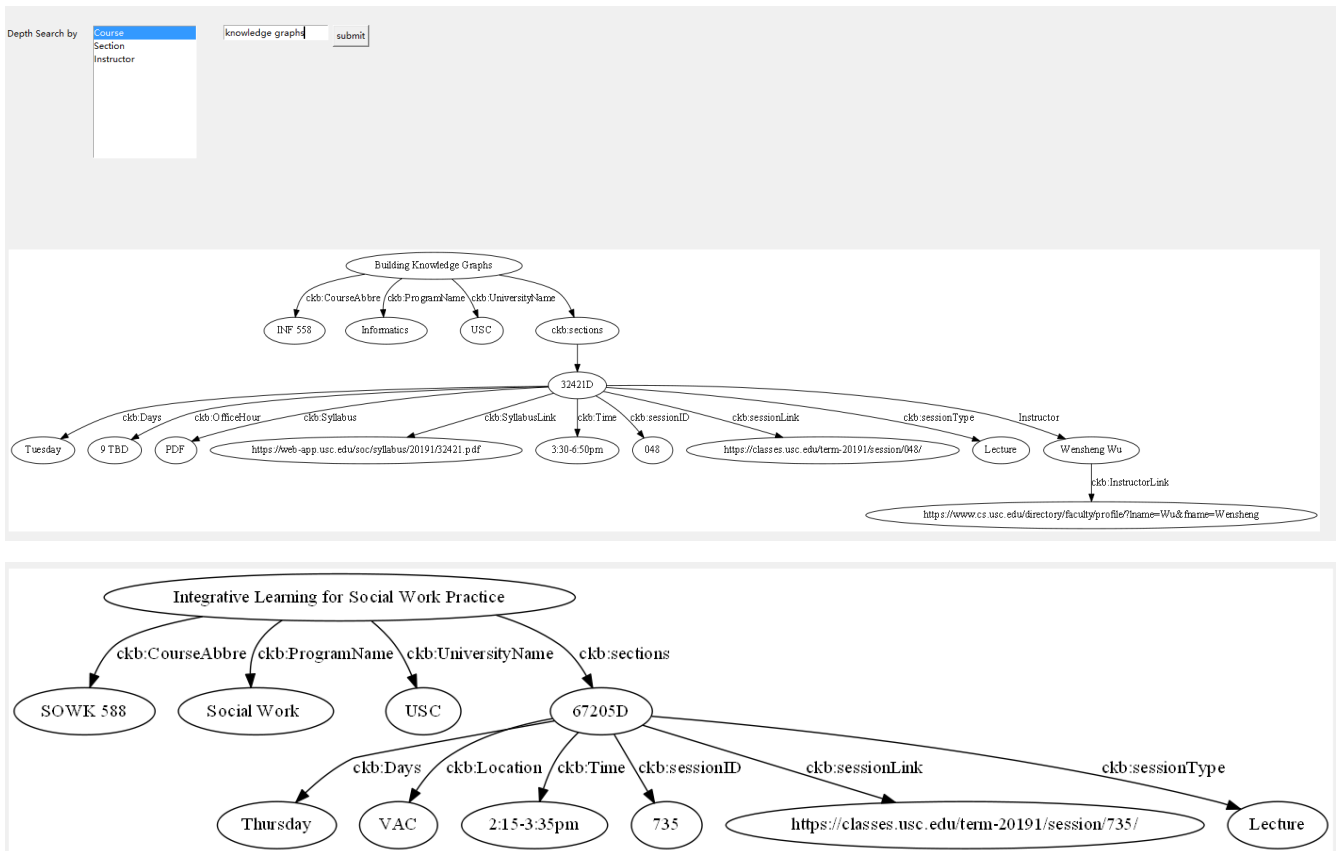
Course
 Section
 Instructor

JSON Viewer

test.json

Key	Value
Root	
0	
@id	._:371
@type	ckb:Course
ckb:CourseAbbre	INF 558
ckb:CourseName	Building Knowledge Graphs
ckb:Description	Foundations, techniques, and algorithms for building knowled...
ckb:ProgramName	Informatics
ckb:UniversityName	USC
ckb:sections	._:373
1	
@id	._:6295
@type	ckb:Course
ckb:CourseAbbre	CSCI 563
ckb:CourseName	Building Knowledge Graphs
ckb:Description	Foundations, techniques, and algorithms for building knowled...
ckb:ProgramName	Computer Science
ckb:UniversityName	USC
ckb:sections	._:6297
2	
3	

Screen Shots - Search Result - Tree Graph



Implementation Details

Basic search window is built by tkinter.

Json Viewer is an open source github repo from <https://github.com/ashwin/json-viewer>.

Tree graphs is built based on graphviz and dot language. With query result, rebuilt that into a dot file and generate image by graphviz.

Query Details - Breadth Search

Breadth search locates all results that are relevant to user query. Rank results by similarity score. Result shows a list of one object.

Query in code is formed as follow. Two matches restrict @type and search keyword.

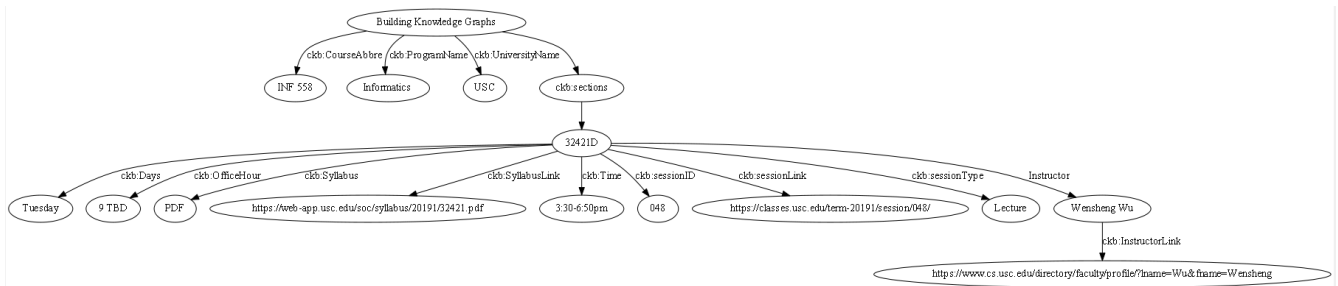
```

1  basicBody = {
2      'query': {
3          'bool': {
4              'must': [
5                  {'match': {}},
6                  {'match': {}}
7              ]
8          }
9      }
10 }
```

Query Details - Depth Search

Depth search locate one result which is most relevant to user query. Then it search related parent nodes and child nodes and combine them together as Json object result.

For example, this is a result from depth search, which search course name 'knowledge graphs'.



Query in code is a nested for loop structure, which iteratively detect parent nodes and child nodes.

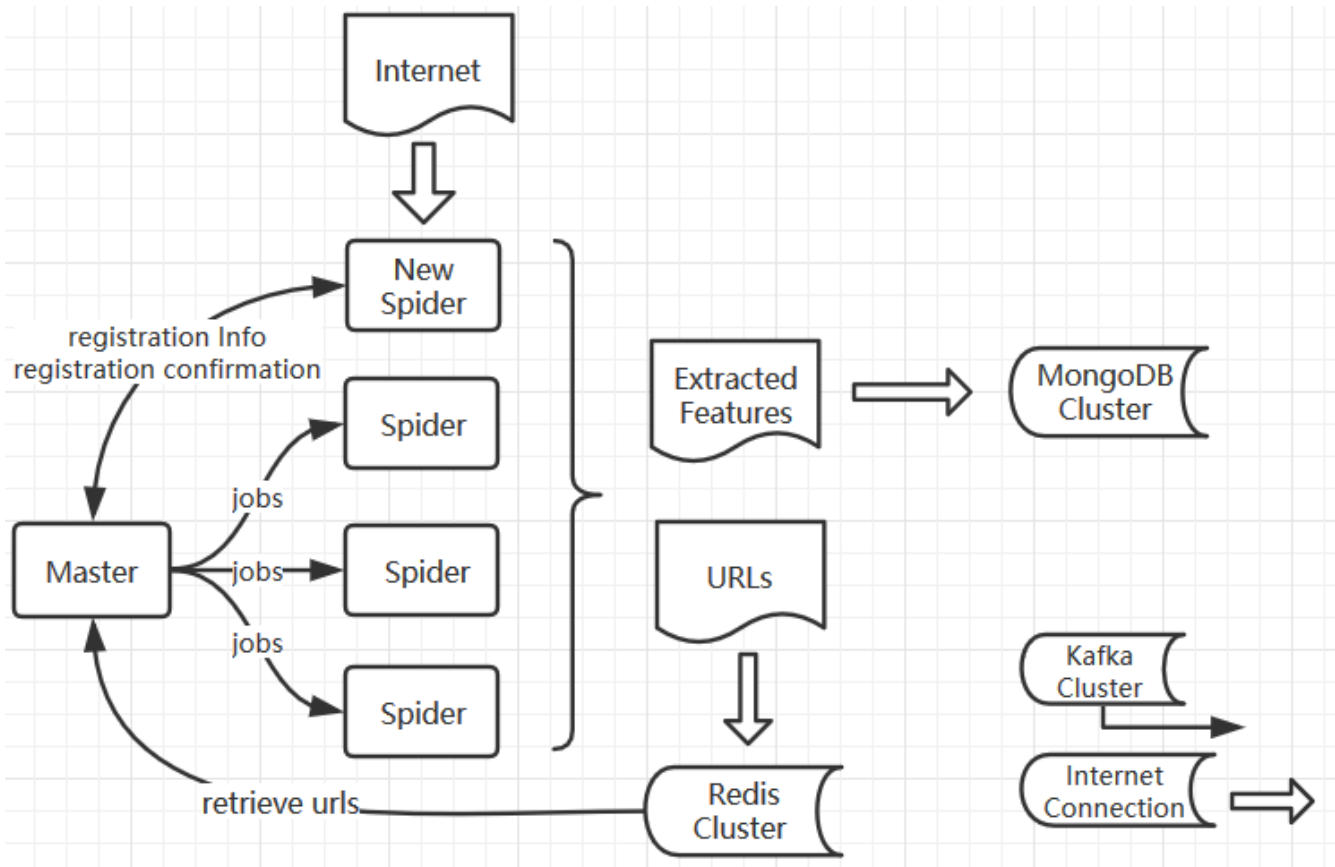
3. Distributed Web Crawler

1. Overall Structure
2. Spider Logic
3. Master Logic

Distributed Web Crawler

This part is built before midterm progress report, so content here will be similar to previous report

3.1 Overall Structure

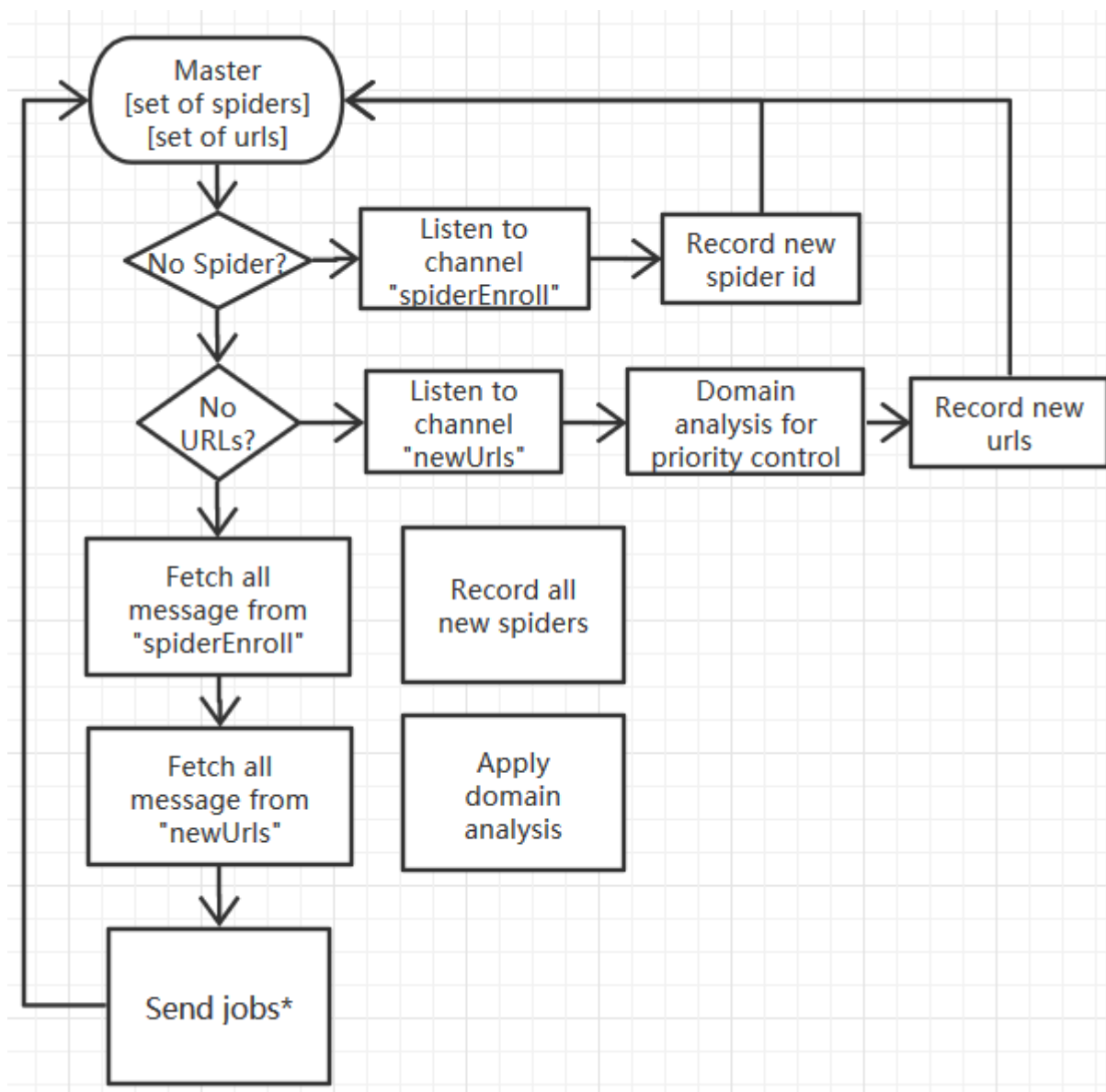


- Master program
 - Distribute URLs(jobs) to Spiders with care of priority and politeness
 - Only one Master program in total.
- Spiders
 - Listen to Master program, receive jobs, fetch data from Internet, parse data and save data in Redis and Mongo DB.
 - Number of spider almost has no limitation.
- Redis Cluster
 - A in-memory distributed database used to maintain URLs and document fingerprints.
- Mongo DB Cluster
 - A normal distributed No-SQL database used to maintain crawled data from web.
- Kafka Cluster
 - A distributed message queue application used to maintain message connection among individual components.
 - All solid arrows are connection based on Kafka Cluster. Hollow arrows are direct Internet connection.

Steps to Run

1. Run component Master program with configuration of crawling speed, priority information, start URLs.
2. Run component Spider with a unique ID.
3. Spider will send message through Kafka to Master as a registration process.
4. Master distributes jobs to spiders with care of priority and politeness.
5. Spiders receive jobs, analysis content, save new URLs into Redis, save other useful data into Mongo DB.
6. Master will fetch new URLs from Redis and distribute them to spiders again.

3.2 Master Logic



3.3 Spider Logic

