



엘라스틱서치를 활용하여 고급 검색엔진 만들기

Henry Lee



Overview

- 검색 엔진 기술의 개요
- 엘라스틱서치를 사용하여 간단한 쇼핑몰 검색 기능을 만들기 [실습 위주]
- 지식그래프의 개요 [실습 포함]
- 이미지 처리를 활용하여 검색 결과를 향상 시키기 [실습 위주]
- 검색 랭킹 [실습 포함]



Prerequisites

기본지식

- 파이썬 (Python), 자바스크립트 (JavaScript), PHP, HTML 사용가능
- 호스트, 포트, 서비스 등 네트워킹 관련 기본지식 및 서버 개발에 대한 기본 이해지식
- RESTful 아키텍처에 대한 이해
- Docker 에 대한 기본 지식

사용되는 도구

- 윈도우나 맥, 또는 리눅스 처럼 Docker 가 지원되는 운영체제 (맥과 윈도우 예제 설치 설명은 첨부됨)
- Apache, WordPress, Elasticsearch, Kibana, 인터넷 브라우저, TensorFlow, MySQL, Docker
- Optional: Postman, MySQL Workbench, Microsoft Visual Studio Code



What you will learn

- 검색 기술에 대한 기본 지식
- 엘라스틱서치를 활용한 검색 시스템 구축
- 지식 그래프, 부속 처리 기술들을 사용한 고급 검색 기술 활용법
- 검색 랭킹의 기본 지식, 검색 결과의 개인화에 대한 개요
- 예제를 통해 위 기술들에 대한 기본적인 실습



Why You Should Take This Course

- 빅데이터의 경쟁력은 데이터를 최대한 빠르고 효율적으로 정리해 주는 기술들이다
- 현존하는 데이터의 규모는 현재도 폭발적인 성장을 하고 있다
- 빅데이터를 다룰 줄 아는 자가 빅데이터의 패러다임을 이끄는 기업의 핵심 전문가가 된다



About Myself

- 1980년대 때부터 개발 경험 시작
- 대략 20년 전부터 시작된 엔지니어 실무 경력
- 대략 15년 전부터 검색 엔진 기술 개발
- 실리콘밸리 FAANG 기업들에서 검색 엔진 엔지니어로 10년 정도 근무
- 대략 3년정도 실리콘밸리의 창업 경험

1. 검색 엔진 기술의 개요



Overview

- 검색 엔진 기술이란?
- SQL, NoSQL 과의 비교
- 검색 엔진 기술의 기본 아키텍처
- 실리콘밸리에 검색 기술 활용 사례



검색 엔진 시스템이란?

컴퓨터 시스템에 저장된 정보를 찾아주는 것을 도와주도록 설계된 정보 검색 시스템

[출처: [위키피디아](#)]



검색 시스템들의 사례

- 오프라인
 - 파일 검색, 데스크탑 검색 ([마이크로소프트 코타나](#))
- 검색 서비스
 - 웹 검색, 이미지 검색, 비디오, 오디오 검색 등등
- 인터페이스형
 - 인공지능 개인비서: 아마존 알렉사, 삼성 빅스비 등
 - 지도형: 카카오맵, 배달앱 등
- 추론형
 - 부동산: Zillow, Redfin
 - Inference-as-a-service: 팔란티어, 케임브리지 애널리티카

SQL로는 안되나요?

—

Scale, Speed, and Usefulness

—



SQL vs 검색엔진

SQL

- ACIDic, even in read replica
- Slows down as data size increases $O(\log n)$
- Realtime up-to-date data
- Advanced search features / ranking very difficult
- Linear increase in traffic can cause exponential decay in service speed

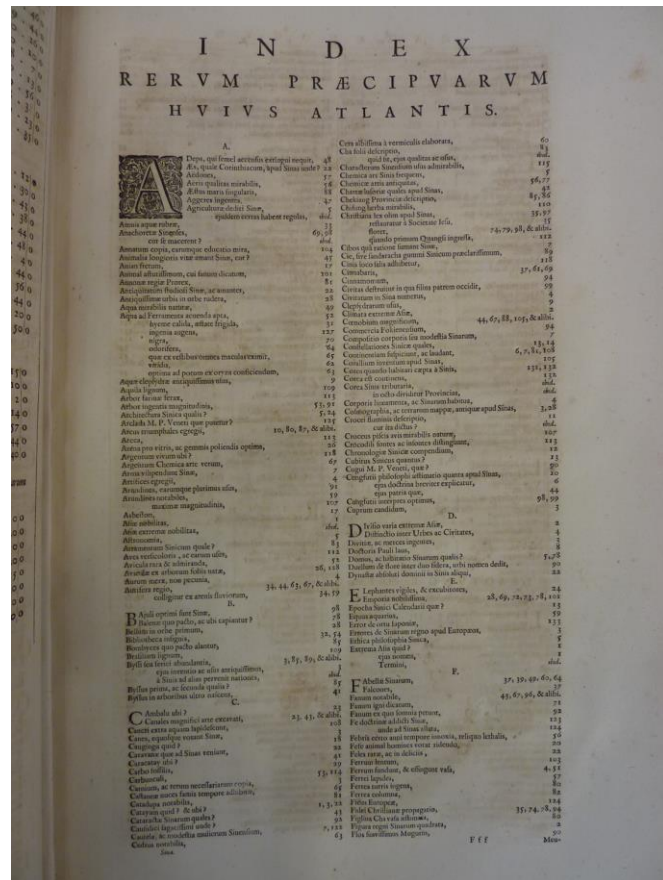
검색엔진

- Queries returned in near constant time $\sim O(1)$
- Data can expand almost infinitely
- Data can be stale
- Traffic increase is unrelated to the system itself, serving systems can grow linearly with the traffic

검색 엔진 시스템이 대체 무엇인가요?

—

1655 년 출판된 Atlas Maior의 색인
[출처: [위키피디아](#)]





Index (색인) 란

A		50
Aaron	51	
About	52	
B		32, 50
Baby		12
Bit		23, 50, 77
Bitten	30, 23	
Bitter	55	
...		

색인과정

[Page 40]

... The quick brown fox jumps over the lazy dog ...

...

Brown

51, 40

...

Dog

32, 50, 40

...

Fox

23, 50, 77, 40

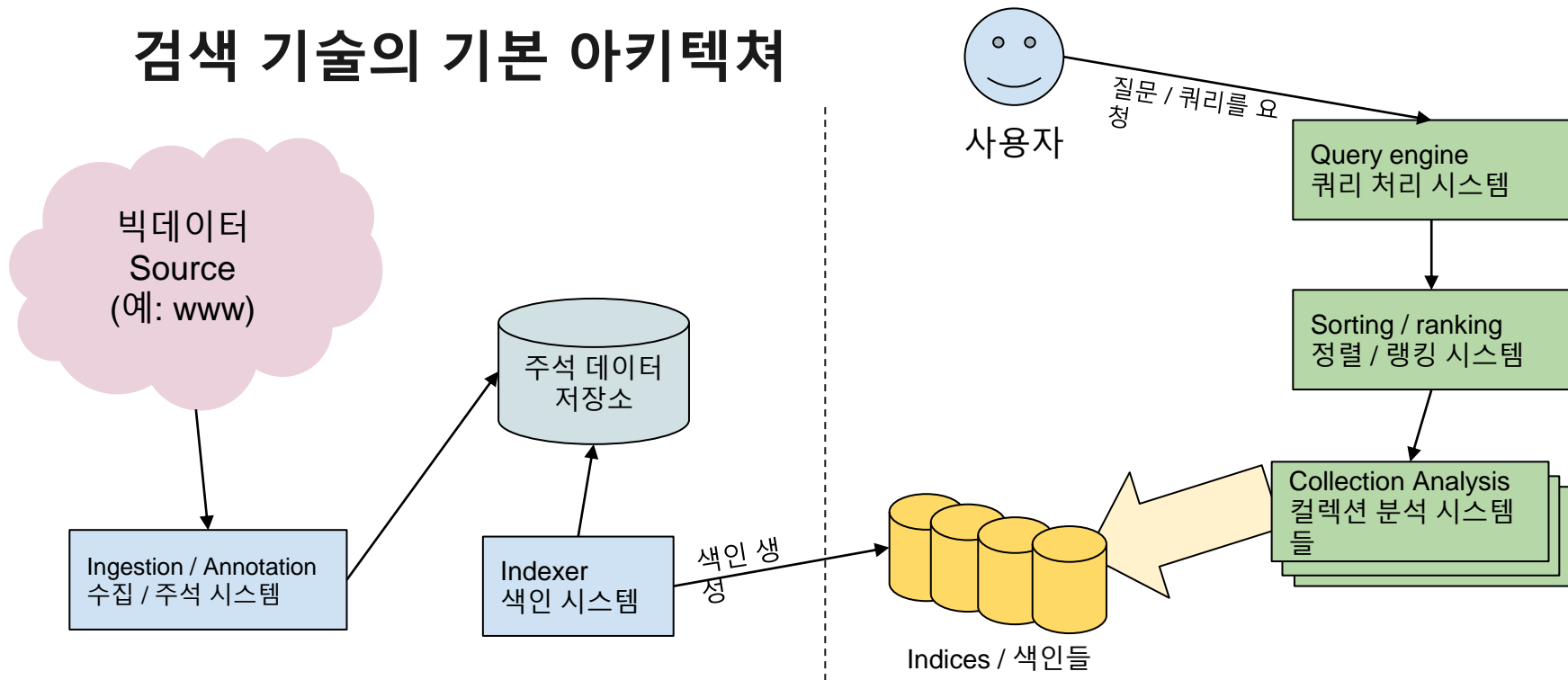
...

Over

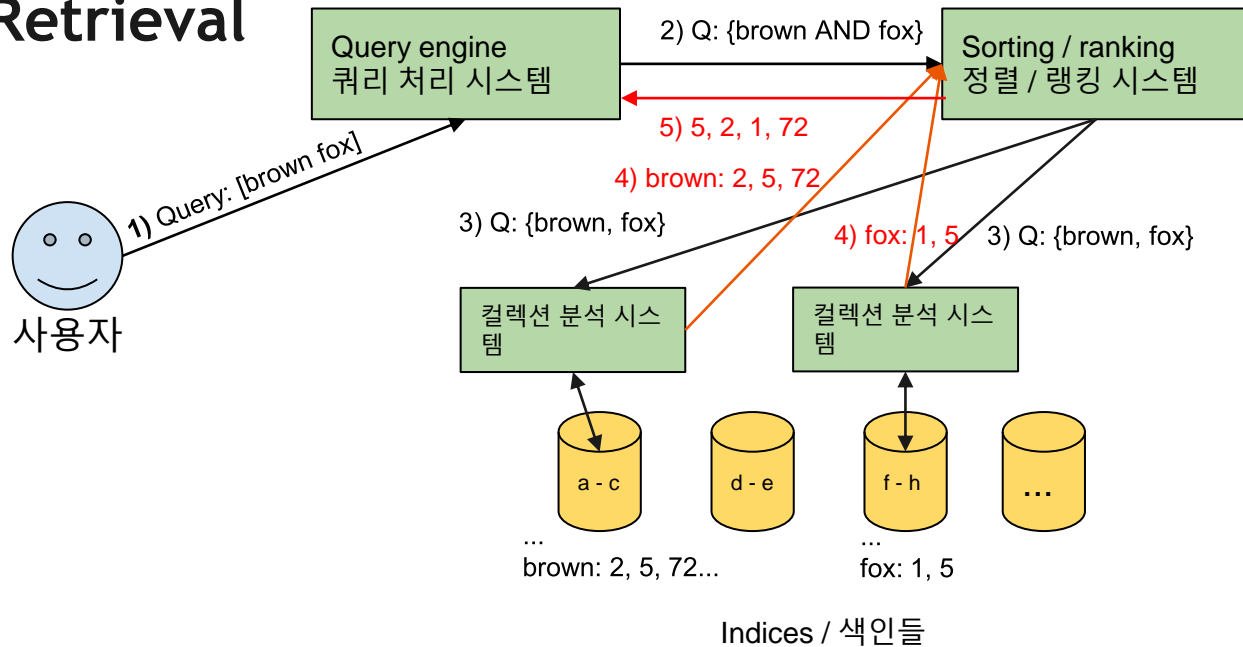
55, 40

...

검색 기술의 기본 아키텍처

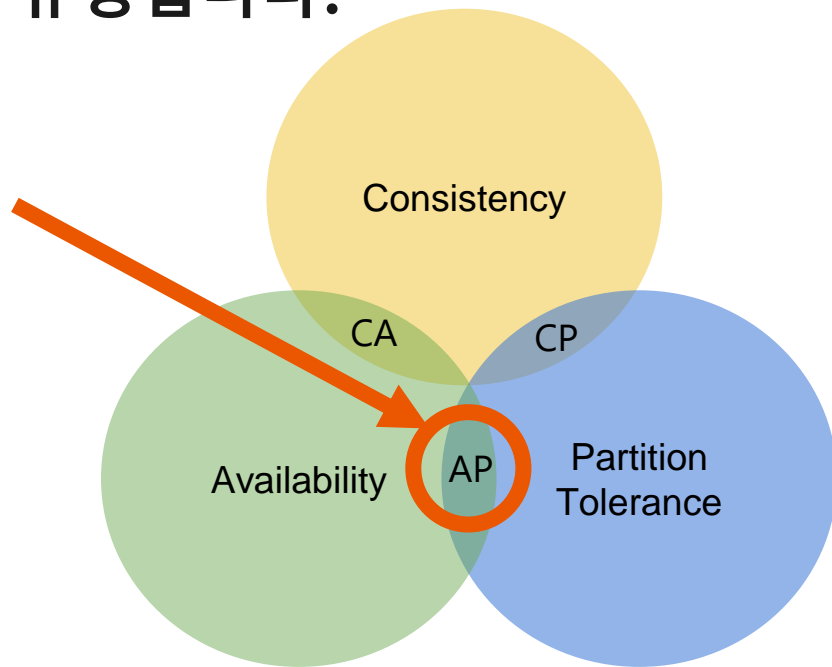


Information Retrieval



NoSQL DB들과 어떻게 다른가요?

검색엔진은 전문화된 NoSQL의 유형입니다.



검색 엔진 기술의 활용사례

Quora

자연어 / 관련 검색

quora.com/Whats-the-best-way-to-cook-salmon?q=how%20do%20%20cook%20salmon

Quora

Salmon Fish (as food) Learning to Cook Fish Recipes Cooking

What's the best way to cook salmon?

Answer Follow 89 Request

Ad by Grammarly

Writing something important?
Join 30 million Grammarly users and see what better, clearer writing can do for you. Free!

Download

88 Answers

Chris O'Leary, Gifted Amateur Cook.
Answered May 28


Sous Vide.

First things first though, pay attention to the salmon you buy, the fresher the better, and leave the skin on.

ZERO Question on this one, I have done Salmon 600 different ways, and until I got my sous vide machine I would have told you, Cedar Plank, on the Grill.

But Salmon was one of the first things I did in the Sous Vide bath that showed how great an investment the machine was.

(Sorry, there's not going to be a lot of pictures on this one like normal, it will be a while before I do Salmon again, my wife is mildly allergic to fish)



Search: how do i cook steak

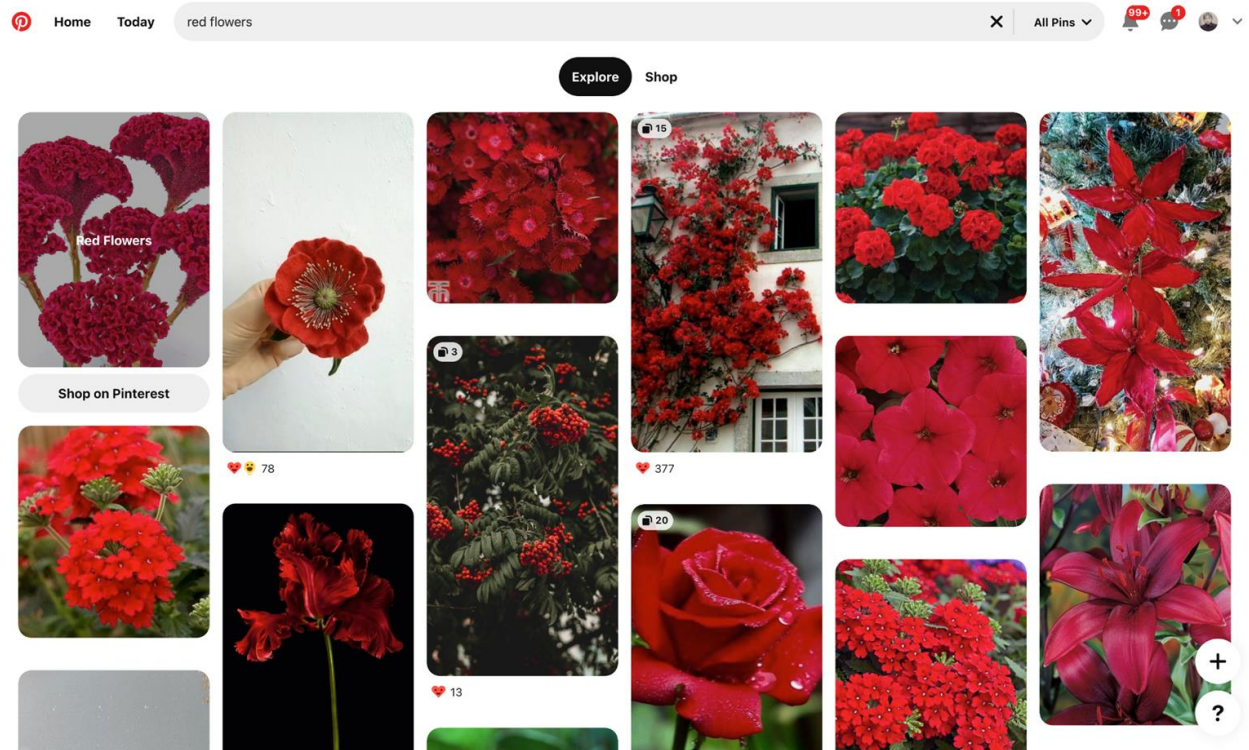
- What are the best ways to **cook steak** in an oven?
- How do I cook steak** in a pan?
- How do you cook steak** Umms?
- What is the best way to **cook a steak**?
- What is the best way to prepare **steak**? What is the best cut to use? What seasonings are best and in what proportions? What is the ideal degree of **cooking**, from blue to well done? I'm looking for optimum deliciousness.

+ Add New Question

자료출처: [www.pinterest.com]

Pinterest

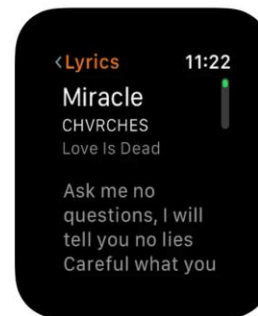
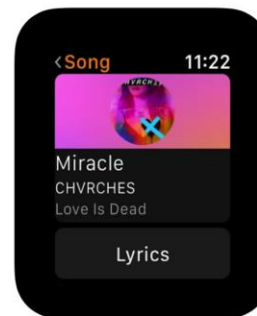
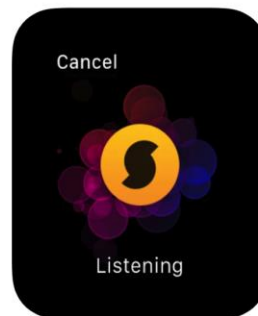
이미지 검색



자료출처: [www.pinterest.com]

Soundhound

음원 검색



자료출처: [play.google.com, apps.apple.com]



Palantir

Data Insights

n/a



Frequently Asked Questions

[Edit profile](#)

DNA 검색

Get started with your predicted relationships, then connect and message to learn more.

Filters

Showing 382 of 382 relatives

Sort by

Strength of Relationship

Search keywords

Name, location, notes

Notifications

Profile features and activity

Mother's side/Father's side ⓘ

Ancestor birthplaces

Connections ⓘ

2nd Cousin

2.77% DNA shared, 9 segments

4th Cousin

0.40% DNA shared, 2 segments

4th Cousin

0.40% DNA shared, 1 segments

4th Cousin

0.36% DNA shared, 1 segments

4th Cousin

0.32% DNA shared, 1 segments

4th Cousin

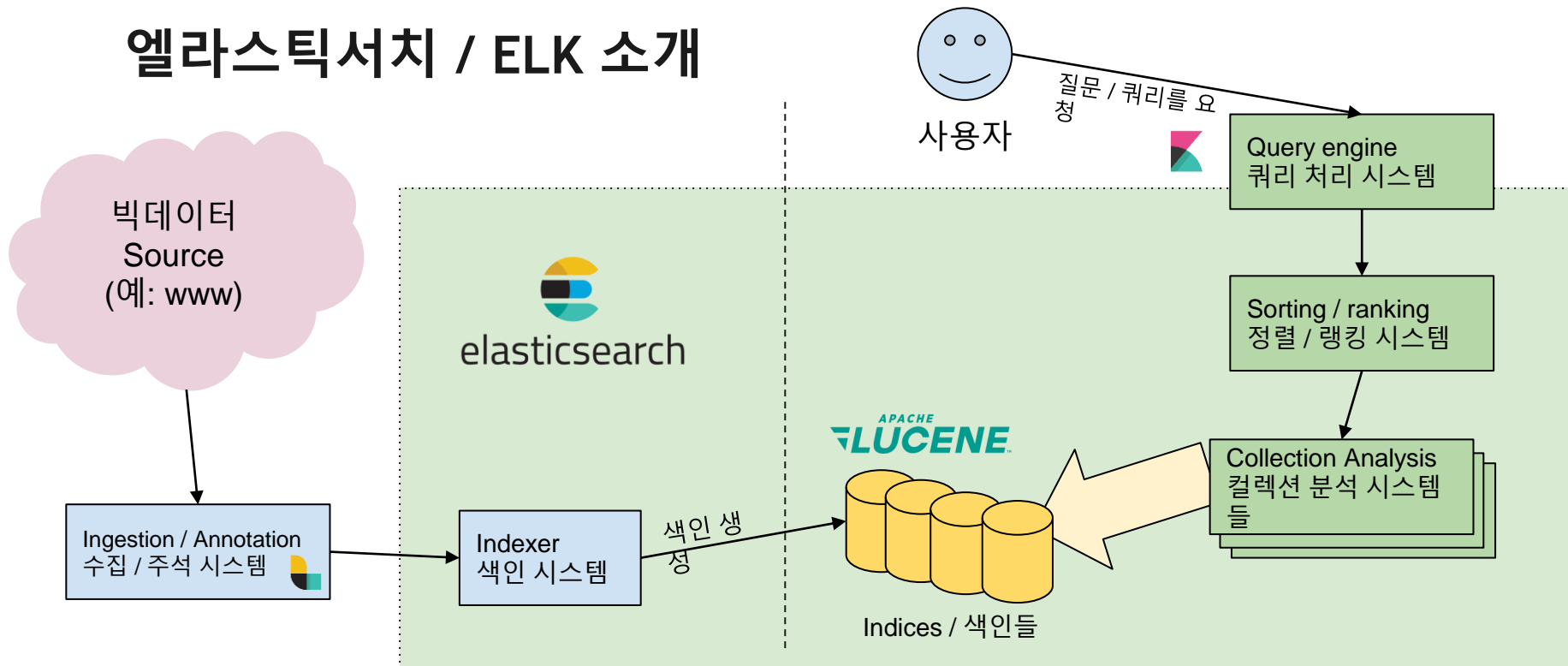
0.31% DNA shared, 2 segments

4th Cousin

자료출처: [23andme.com]

2. 엘라스틱서치를 사용하여 검색 기능을 추가하기

엘라스틱서치 / ELK 소개



플라워몰

감색기술 관련 예제 사이트

결제 내 계정 상점 장바구니 홈페이지

Login/Register

🛒 (0)



플라워몰 예제

http://localhost:
8000/

검색 ...

최신 글

안녕하세요!

최신 댓글

안녕하세요의 워드프레스 댓글 작성자

보관함

2021년 7월

카테고리

▶ n1보르

7개 결과 출력

기본순



꽃병에 담긴 해바라기 2송이

38,000₩



대형 테라코타 화분

~~60,000₩~~ 30,000₩



데이지 꽃 세트

12,000₩



민들레 및 여러가지 꽃과 화병

~~65,000₩~~ 40,000₩





플라워몰 실행하기

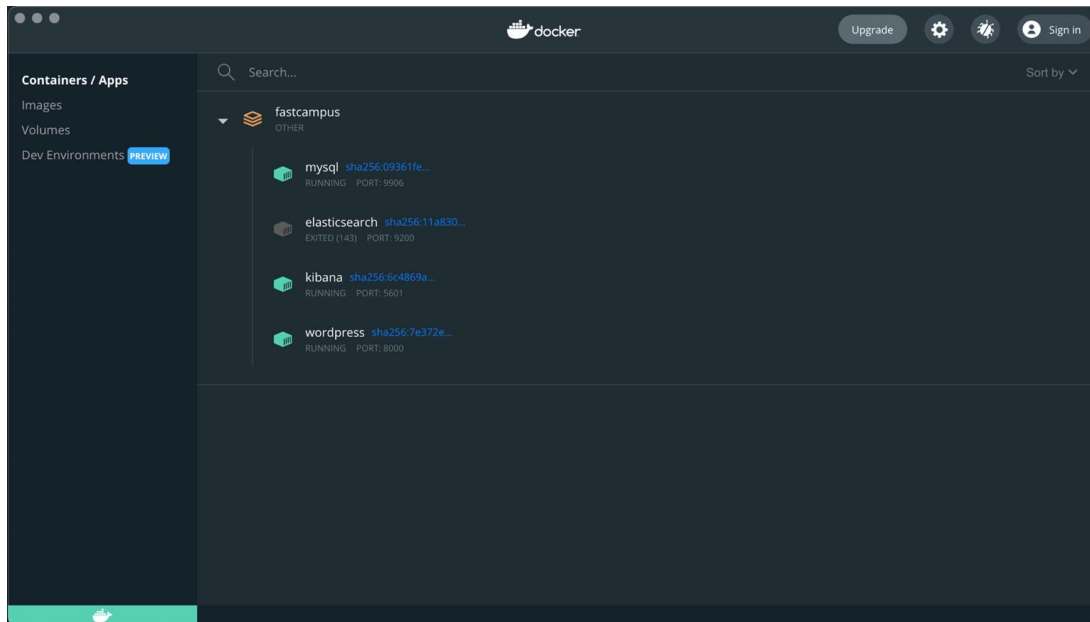
Download and install Docker at:

<https://www.docker.com/products/docker-desktop>

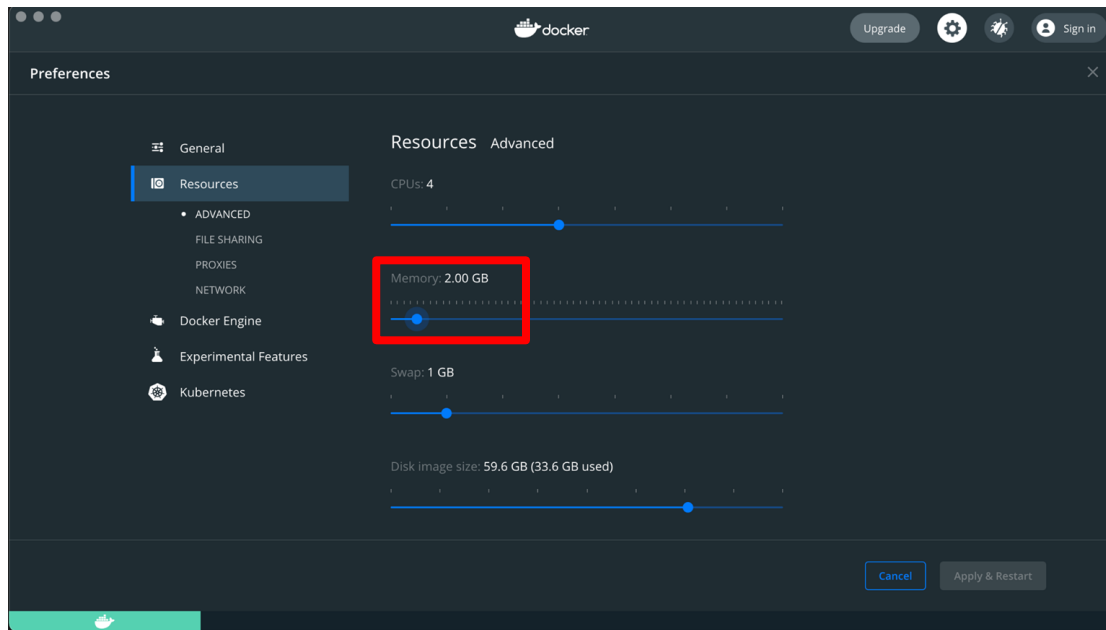
Then run

```
> docker compose up -d
```

엘라스틱서치가 실행이 안될때



엘라스틱서치가 실행이 안될때





플라워몰 웹사이트

`http://localhost:8000/`



플라워몰에 검색 기능 추가하기

색인 프로그램 `/indexer/ingestor1.py`

1. `> cd ingestor`
2. Install Python3 [Mac, Windows]
<https://www.python.org/downloads/>
3. `pip3 install mysql-connector`
`pip3 install requests`
4. `> python3 tools/create_index1.py`
(https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-index_.html)
5. `> Python3 ingestor1.py`
(https://www.elastic.co/guide/en/elasticsearch/reference/current/docs-index_.html)



엘라스틱서치의 인덱스 확인

```
> python3 tools/check_es_index.py
```

```
{ "took":1, "timed_out":false, "_shards":{"total":1, "successful":1, "skipped":0, "failed":0}, "
hits":{"total":{"value":7, "relation":"eq"}, "max_score":1.0, "hits":[{"_index":"products", "
_type":"_doc", "_id":"90063200dc4ce5dc4271f7bb1b9750e481a5c8f6", "_score":1.0, "_source":{"
  "content": "\ub300\ud615 \ud654\ubd84\uc785\ub2c8\ub2e4. \uc0ac\uc774\uc988\ub294
13\" \uc785\ub2c8\ub2e4.",
    "id": 10,
    "image_file": "2021/07/terra-cotta-planter.jpg",
    "modified_date": "2021-07-02T15:28:15",
    "post_date": "2021-07-01T18:55:54",
    "shipped_from": "\ud574\uc678",
    ...
```



엘라스틱서치에서 검색하기

```
http://localhost:9200/products/_search?q=장미
```

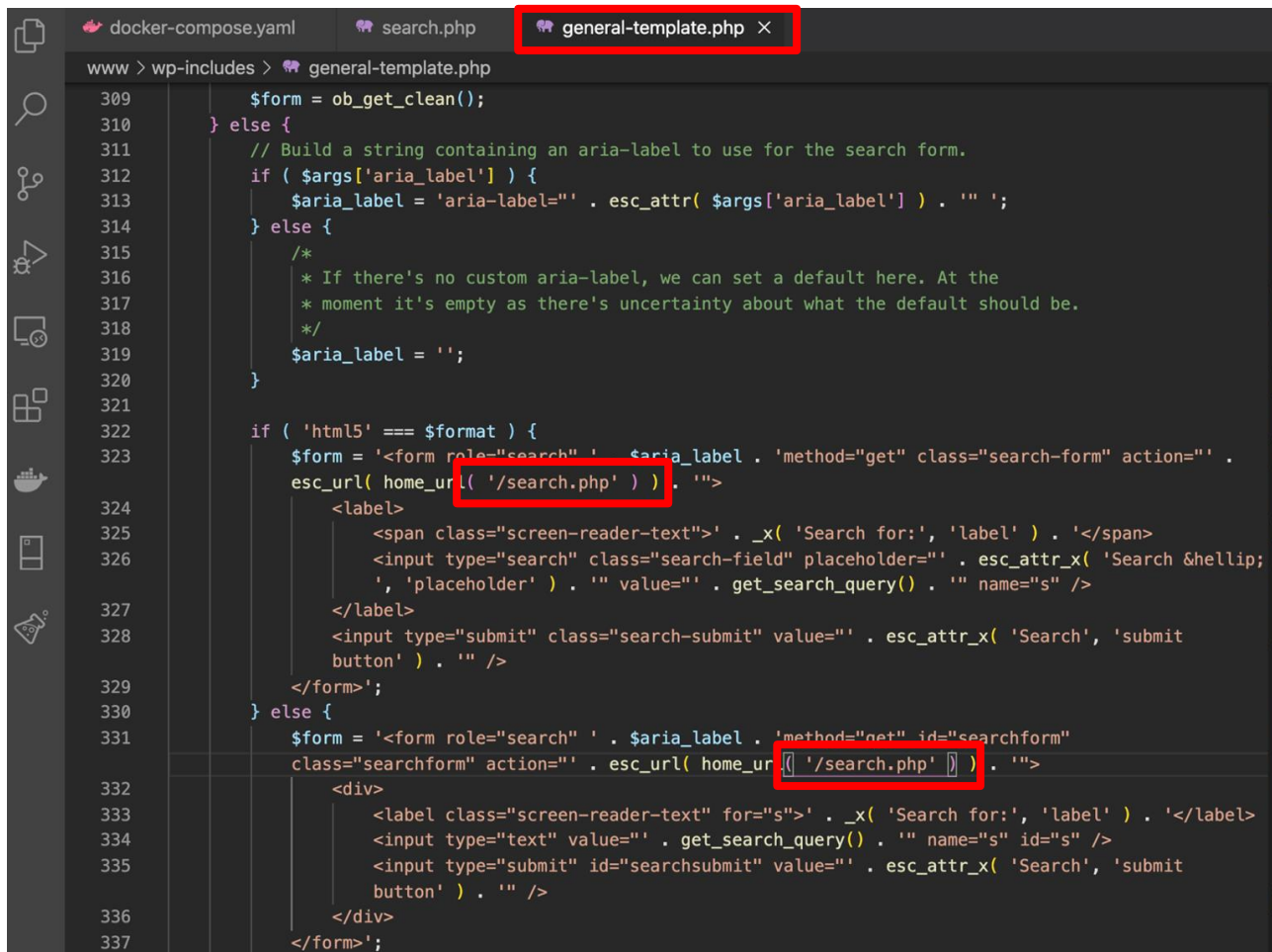


플라워몰 웹사이트에 엘라스틱서치 검색을 연결하기

- 1) 쇼핑몰: `http://localhost:8000/`
- 2) 검색 페이지 추가: `www/search.php`

검색 연결하기

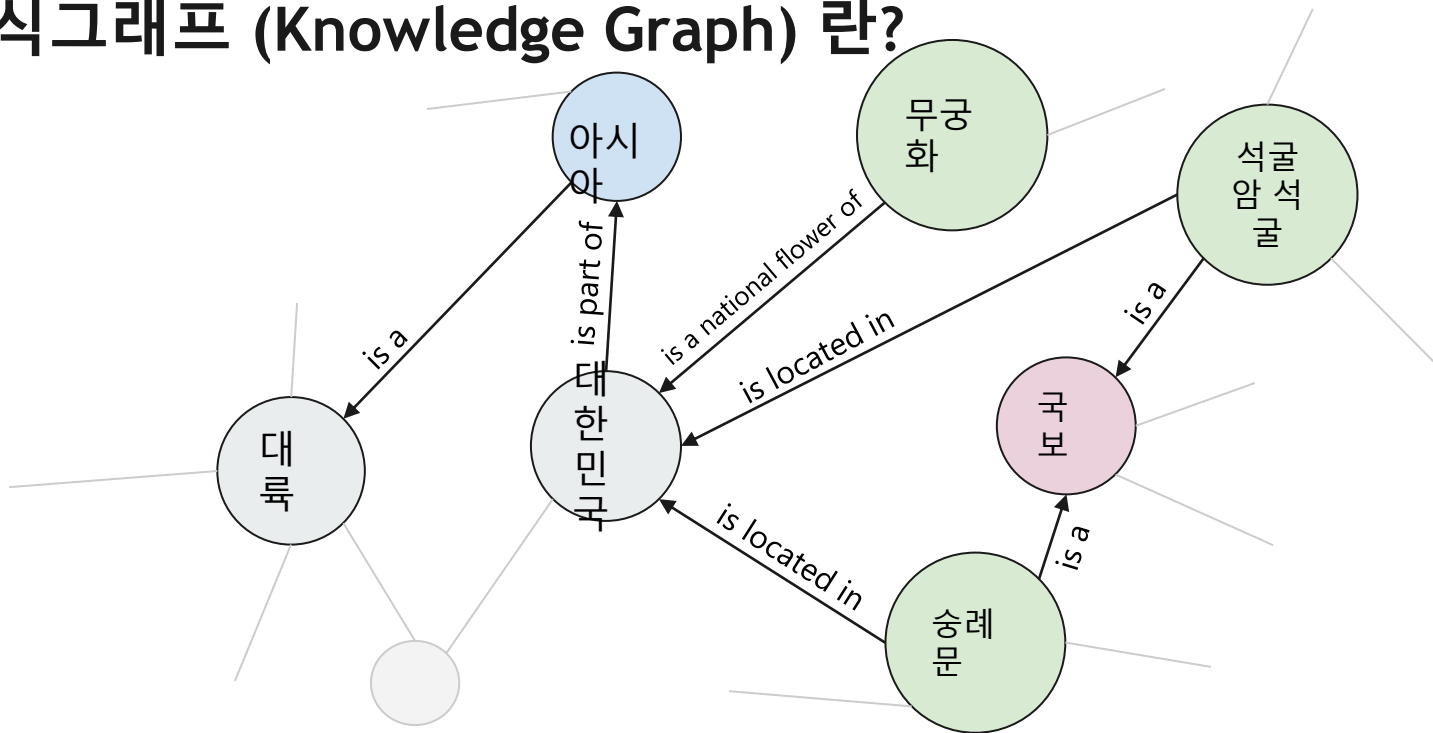
www/wp-
includes/general
-
template.php:322
-335



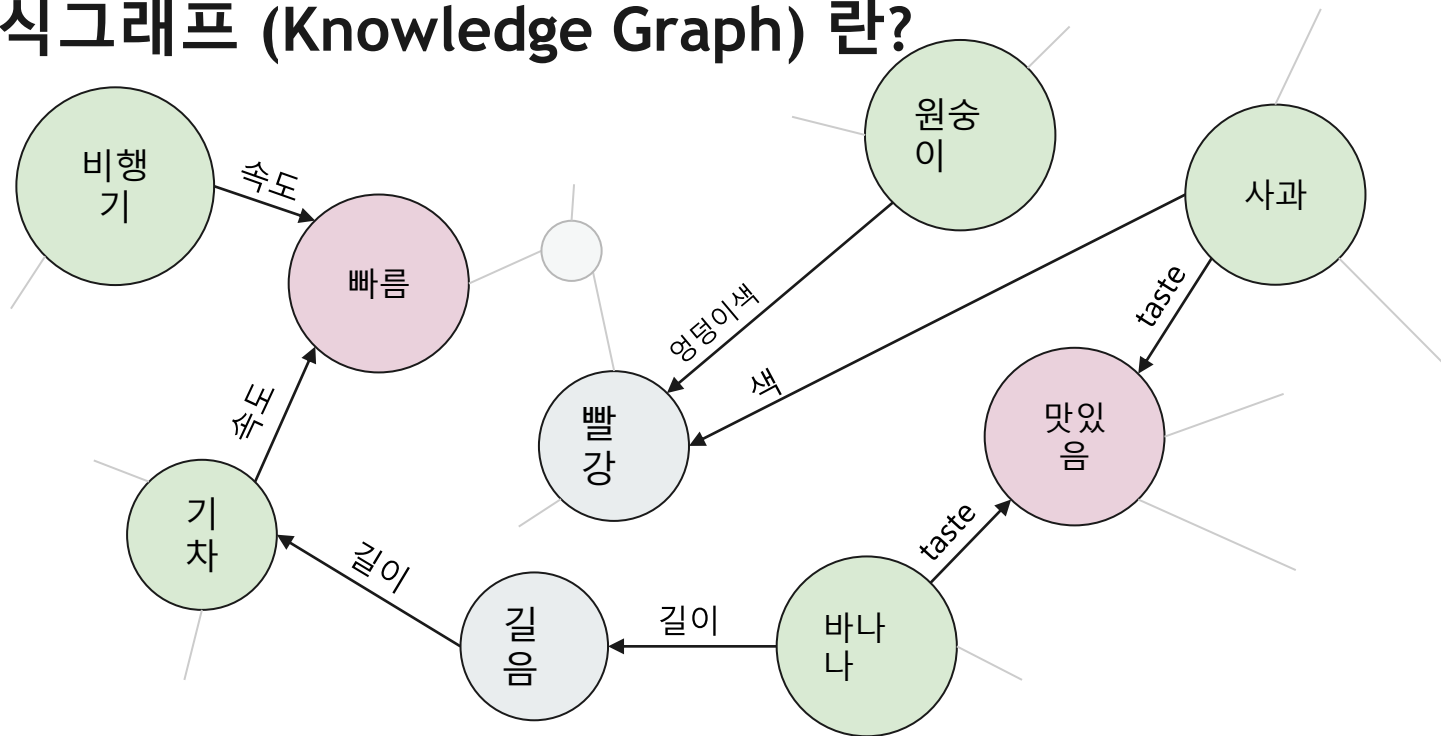
```
309     $form = ob_get_clean();
310 } else {
311     // Build a string containing an aria-label to use for the search form.
312     if ( $args['aria_label'] ) {
313         $aria_label = 'aria-label=' . esc_attr( $args['aria_label'] ) . ' ' ;
314     } else {
315         /*
316          * If there's no custom aria-label, we can set a default here. At the
317          * moment it's empty as there's uncertainty about what the default should be.
318          */
319         $aria_label = ' ';
320     }
321
322     if ( 'html5' === $format ) {
323         $form = '<form role="search" ' . $aria_label . 'method="get" class="search-form" action="' .
324             esc_url( home_url( '/search.php' ) ) . '">
325             <label>
326                 <span class="screen-reader-text">'. _x( 'Search for:', 'label' ) . '</span>
327                 <input type="search" class="search-field" placeholder="' . esc_attr_x( 'Search &hellip;',
328                     'placeholder' ) . '" value="' . get_search_query() . '" name="s" />
329             </label>
330             <input type="submit" class="search-submit" value="' . esc_attr_x( 'Search', 'submit
331                 button' ) . '" />
332             </form>';
333     } else {
334         $form = '<form role="search" ' . $aria_label . 'method="get" id="searchform"
335             class="searchform" action="' . esc_url( home_url( '/search.php' ) ) . '">
336             <div>
337                 <label class="screen-reader-text" for="s">'. _x( 'Search for:', 'label' ) . '</label>
338                 <input type="text" value="' . get_search_query() . '" name="s" id="s" />
339                 <input type="submit" id="searchsubmit" value="' . esc_attr_x( 'Search', 'submit
340                     button' ) . '" />
341             </div>
342             </form>';
```

3. 지식그래프 (Knowledge Graph) 를 활용하여 검색품질을 향상하기

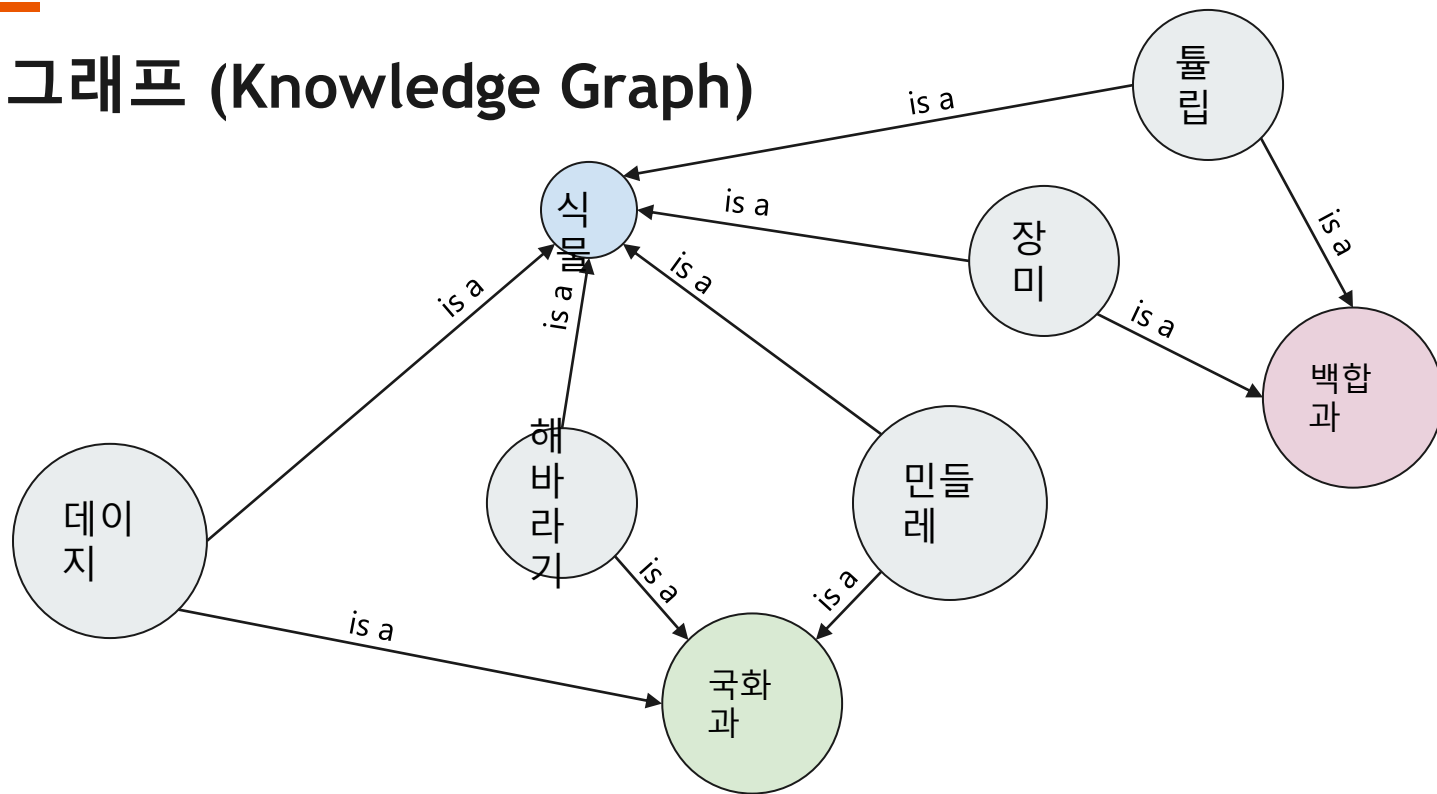
지식그래프 (Knowledge Graph) 란?



지식그래프 (Knowledge Graph) 란?



지식그래프 (Knowledge Graph)



지식그래프의 활용사례

스마트 스피커



출처: 위키피디아

지식그래프의 활용사례

- SNS
- 개인화된 교육
- 날씨 예측
- 선거 공약
- 교통 정책
- 정부 오픈 데이터
- 반테러/보안
- 엔터테인먼트
- 금융

The screenshot shows the Amazon Prime Video interface for the show 'Upload'. The page is in dark mode. At the top, the 'prime video' logo is on the left, and navigation links (Home, Store, Channels, Categories, My Stuff, Deals) are on the right. The show title 'UPLOAD' is prominently displayed in large, stylized letters. Below the title, the season information 'Season 1' is shown with a star rating of 4.6 out of 5 stars (11,276 reviews), an IMDb rating of 8.0, and the year 2020. A description of the show follows: 'From the Emmy-winning Greg Daniels (The Office, Parks & Rec) comes a hilarious new sci-fi comedy. In the future people can upload their consciousness to a luxurious digital afterlife. When party boy Nathan gets uploaded to a virtual resort he meets the down-to-earth Nora who starts as his customer service "angel", but becomes so much more as she helps him find friendship, love and purpose.' Below the description, there are sections for 'Starring' (Robbie Amell, Andy Allo, Allegra Edwards), 'Genres' (Comedy), 'Subtitles' (English [CC], العربية, more...), and 'Audio languages' (English, English [Audio Description], more...). A 'Watch now' button is visible. Below this, it says 'Season 1 included with Prime'. Further down, there are navigation tabs for 'Episodes', 'Related', and 'Details'. The 'Details' tab is selected, showing a table of information about the show, including directors, supporting actors, producers, season year, network, content advisory, purchase rights, format, and devices. The 'Reviews' section at the bottom shows a 4.6 out of 5 stars rating and a link to 'Top reviews from the United States'.

prime video Home Store Channels Categories ▾ My Stuff Deals

UPLOAD

Season 1 ★★★★★ (11,276) IMDb 8.0 2020 [X-Ray] [HDR] [UHD] [AD] [16+]

From the Emmy-winning Greg Daniels (The Office, Parks & Rec) comes a hilarious new sci-fi comedy. In the future people can upload their consciousness to a luxurious digital afterlife. When party boy Nathan gets uploaded to a virtual resort he meets the down-to-earth Nora who starts as his customer service "angel", but becomes so much more as she helps him find friendship, love and purpose.

Starring Robbie Amell, Andy Allo, Allegra Edwards
Genres Comedy
Subtitles English [CC], العربية, more...
Audio languages English, English [Audio Description], more...

prime Season 1 included with Prime

▶ Episode 1 Watch now

By ordering or viewing, you agree to our Terms. Sold by Amazon.com Services LLC.

Episodes Related Details

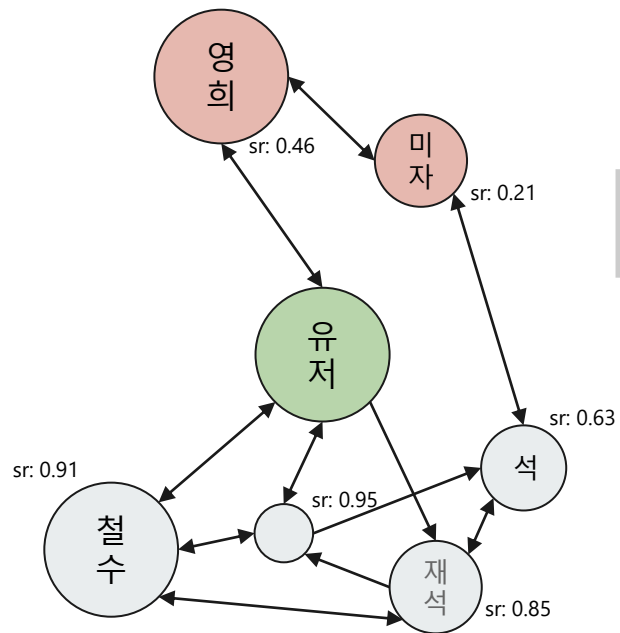
More details

Directors	Greg Daniels, Kacie Anning, Jonathan van Tulleken, David Rogers, Jeffrey Blitz, more...
Supporting actors	Zainab Johnson, Kevin Bigley, Chloe Coleman
Producers	Greg Daniels, Howard Klein, Shepard Boucher, Arielle Boisvert, Shawn Williamson, more...
Season year	2020
Network	Amazon Studios
Content advisory	Alcohol use, foul language, sexual content, nudity, substance use, more...
Purchase rights	Stream instantly Details
Format	Prime Video (streaming online video)
Devices	Available to watch on supported devices

Reviews Sorted by: Top reviews ▾

★★★★★ 4.6 out of 5 stars Top reviews from the United States

지식그래프 사례: SNS



사용자

질문 [동남아 여행]



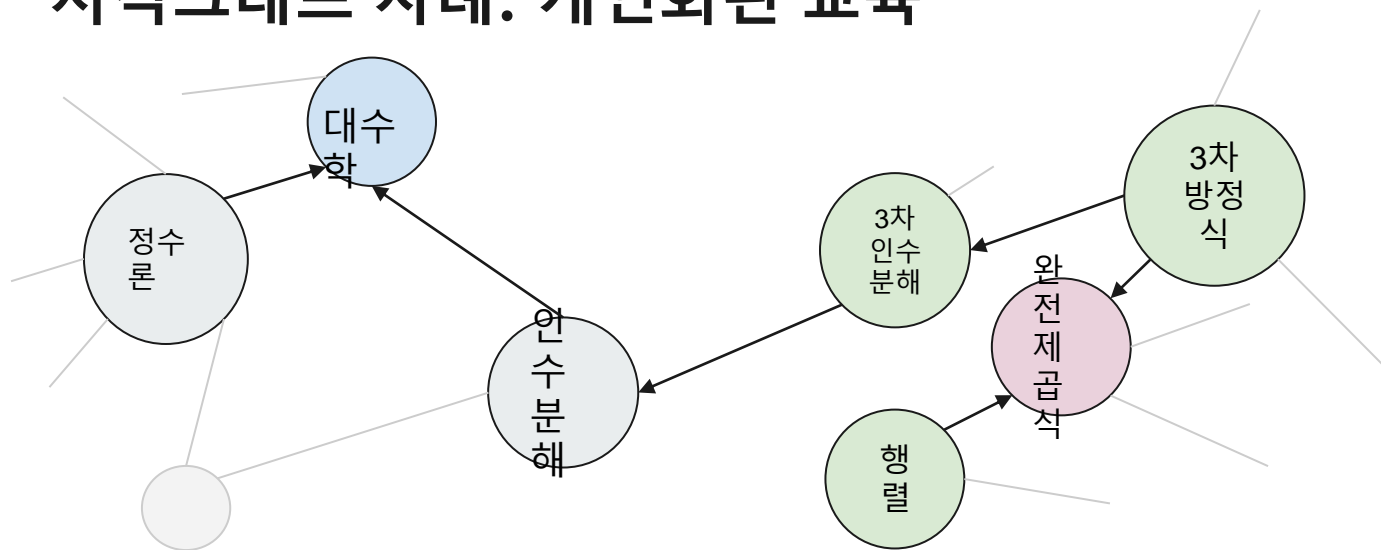
Social rank DB

Query engine
쿼리 처리 시스템

Sorting / ranking
정렬 / 랭킹 시스템

컬렉션 분석 시스템

지식그래프 사례: 개인화된 교육





한글 위키미디어 데이터덤프

`https://dumps.wikimedia.org/kowiki/`



실습: 위키미디어 데이터를 사용하여 관련 키워드 확장하기

```
> create_index2.py
```

```
> ingestor2.py
```

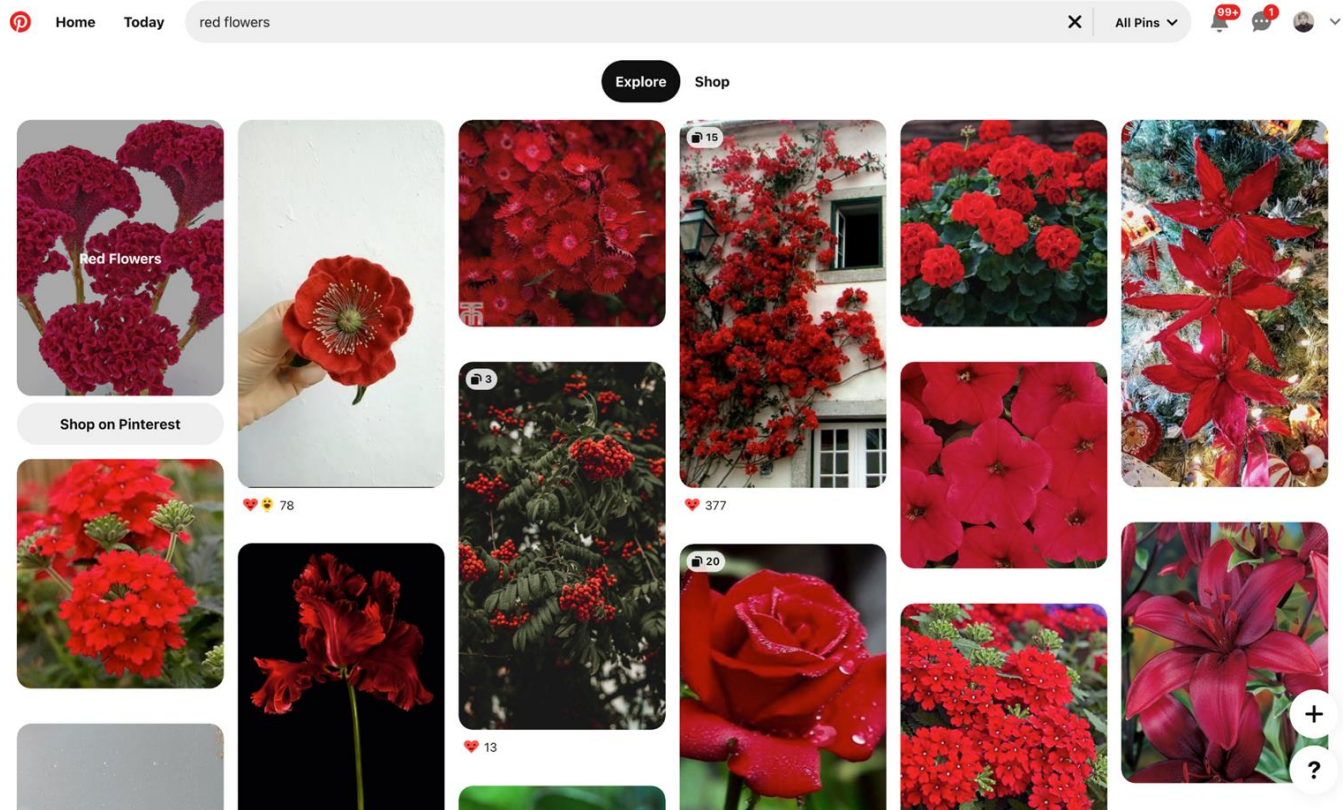


실습: 위키미디어 데이터로 문서의 속성을 키우기

```
> ingestor2.py
```

4. 검색 부속 처리 사례: 이미지 검색

이미지 검색

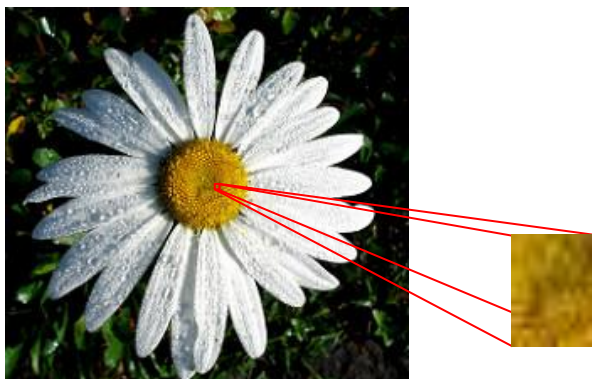




이미지 검색 활용도

- 유사 이미지 검색
- 이미지 / 동영상 저작권 검색
- 유사한 제품 검색
- 시장 조사 (티비 시청률 / SNS 인지도)
- 메타데이터 추론 및 추출

이미지 검색 사례: 유사 이미지 검색



이미지출처 [www.flickr.com]



이미지 검색 사례: 유사 이미지 검색



이미지출처 [www.flickr.com]





실습: 이미지 색상 검색

```
> ingestor3.py
```



실습: 간단한 이미지 Classification 실습

```
> pip3 install -q tf-nightly
```

5. 검색 랭킹

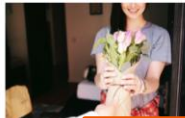


Ranking and Relevance

- Showing results that are more relevant first
- Ranking and relevance is usually marked in numeric values
- The numeric values usually comprises of many factors, combined into a single number

Ranking and Relevance

17ms 안에 3개의 검색결과를 찾았습니다.



[플립 부케](#) (score:0.98423386)

주문시 2인 정도의 배송 시간이 필요합니다. 국내배송.



[핑크빛 장미 부케](#) (score:0.8571708)

크고작은 핑크색 장미들로 부케를 만들어 보았습니다.



[흰색 장미 부케](#) (score:0.8571708)

국내에서 배송되는 흰색 장미의 부케입니다.

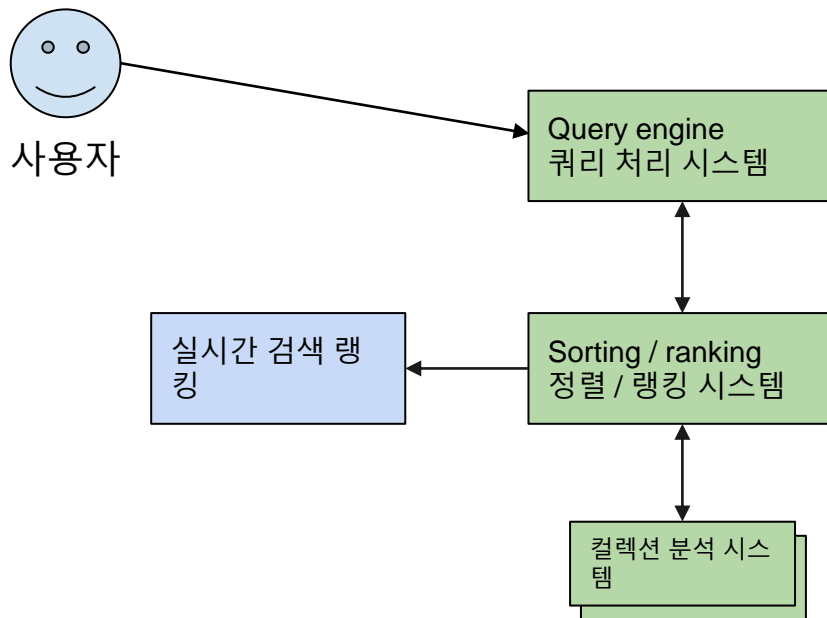


Finding user's intent

- For a query [데이지]
- Do they want to know about the flower?
- Do they want a text containing the word?
- Do they want recent news on the flower?
- Should a document with frequency of the query word be ranked higher?

Ask the user for intent

- 실시간 검색
- 통합 검색





Ranking: TF/IDF

TF: Term-frequency 단어 빈도

특정한 단어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값

$$tf = \sqrt{\text{termFreq}}$$

IDF: Inverse document frequency 역문서 빈도

한 단어가 문서 집합 전체에서 얼마나 공통적으로 나타나는지를 나타내는 값

$$idf = 1 + \ln(\text{maxDocs} / (\text{docFreq} + 1))$$



실습: Local-boost

```
> ingestor4.py
```



실습: Demotion / Promotion (by color)

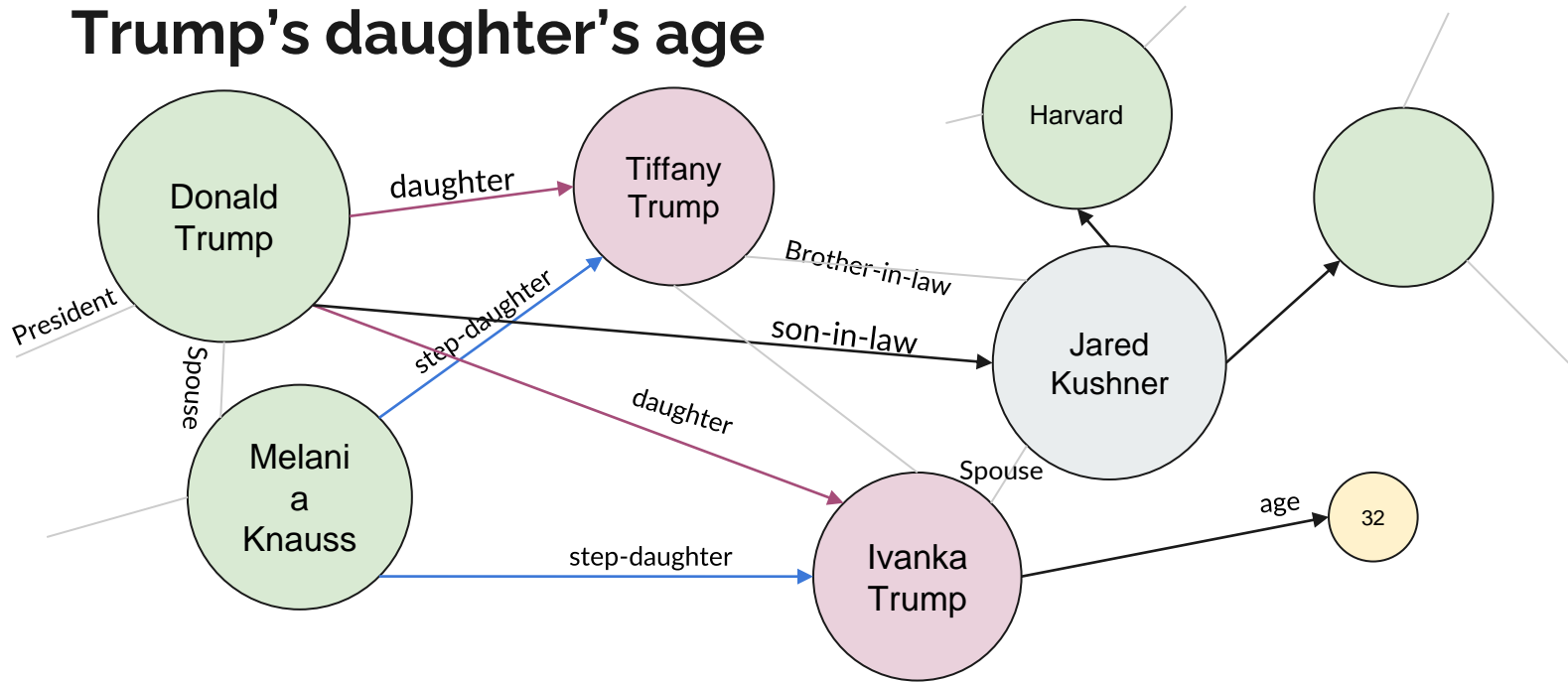
```
> create_index3.py
```

```
> ingestor4.py
```

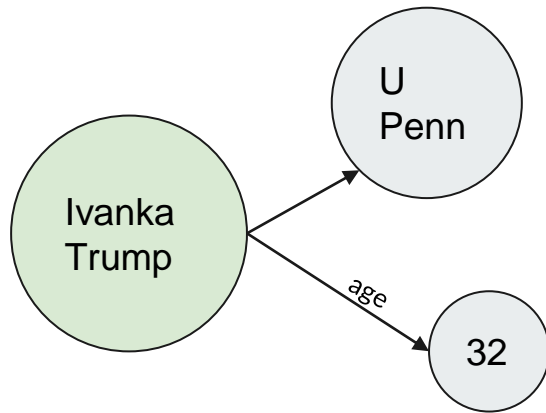
```
> search2.php
```

Knowledge graph: Handling complex queries

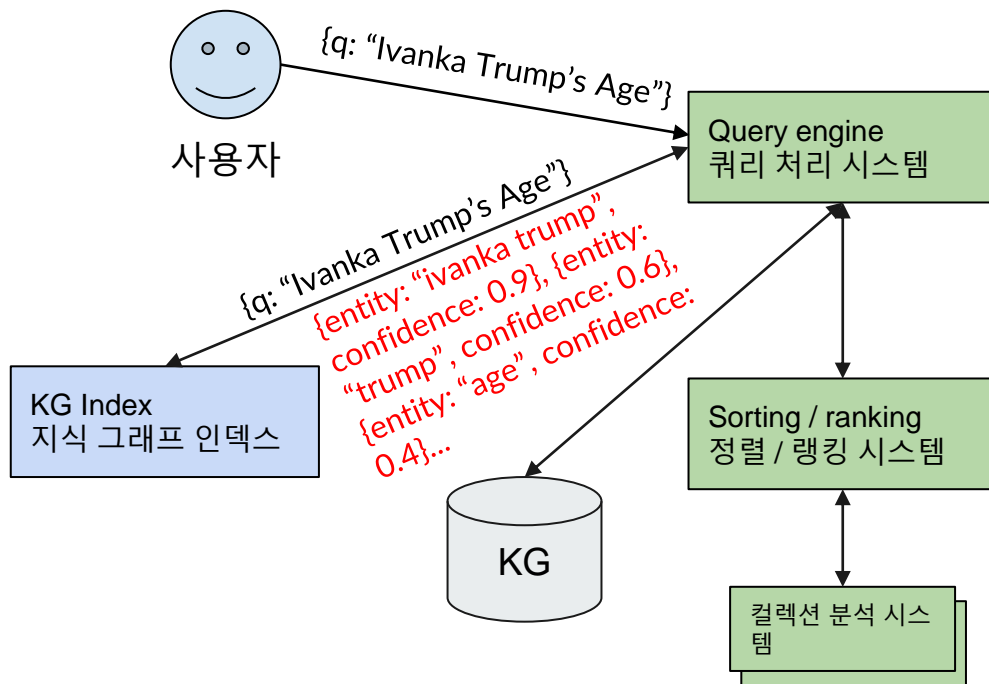
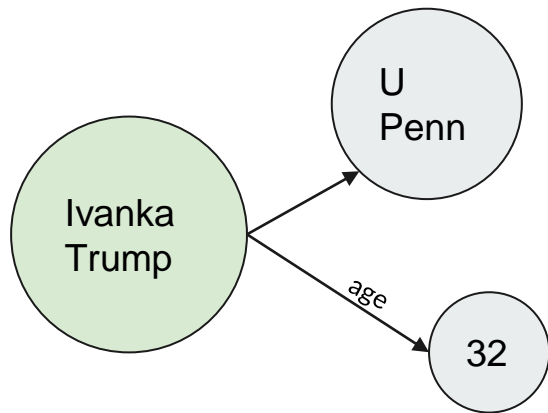
Trump's daughter's age



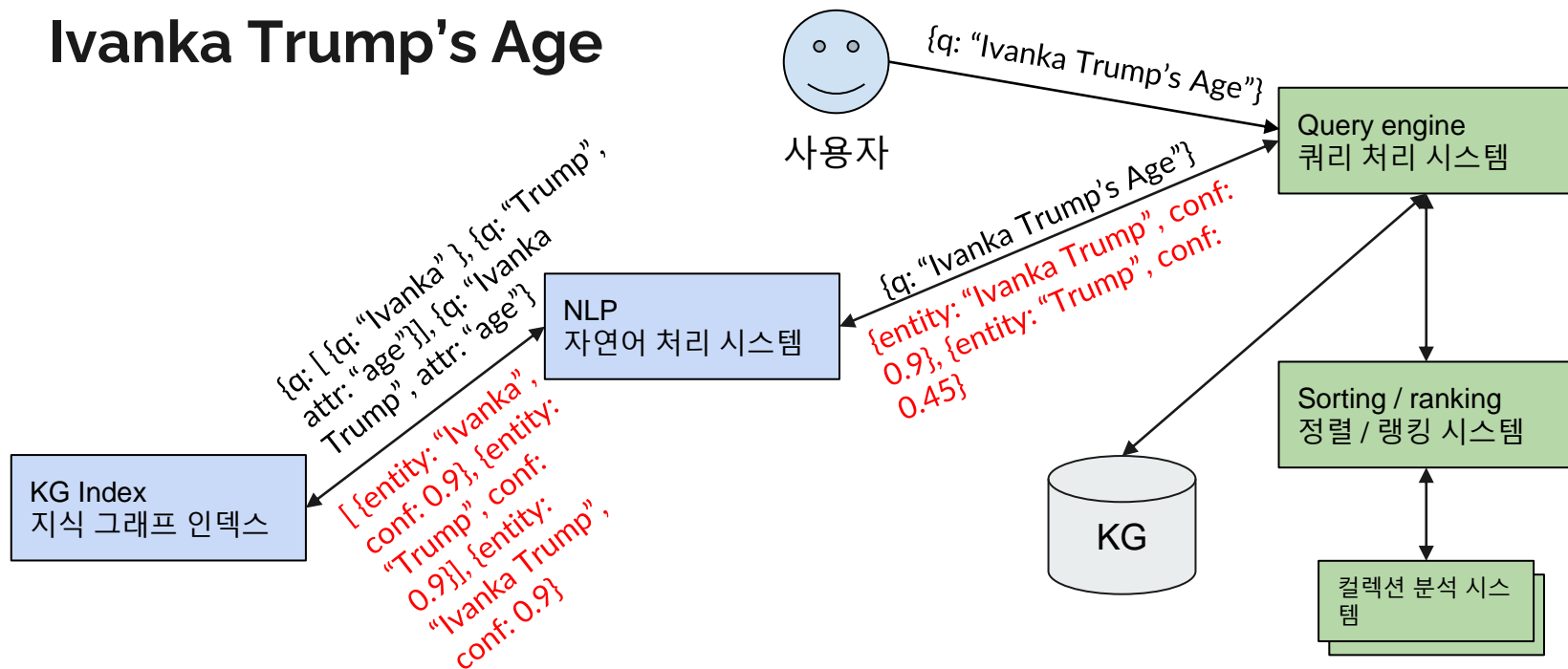
Ivanka Trump's Age



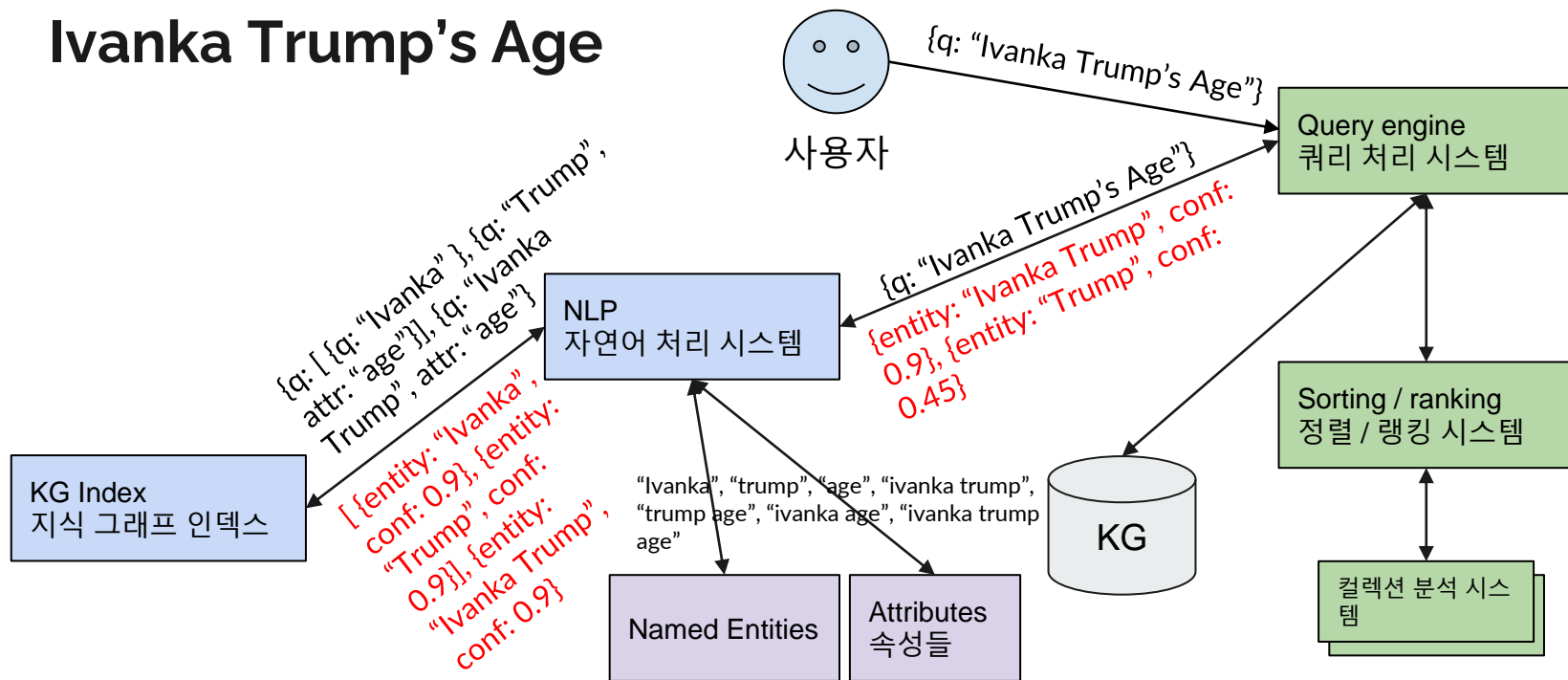
Ivanka Trump's Age



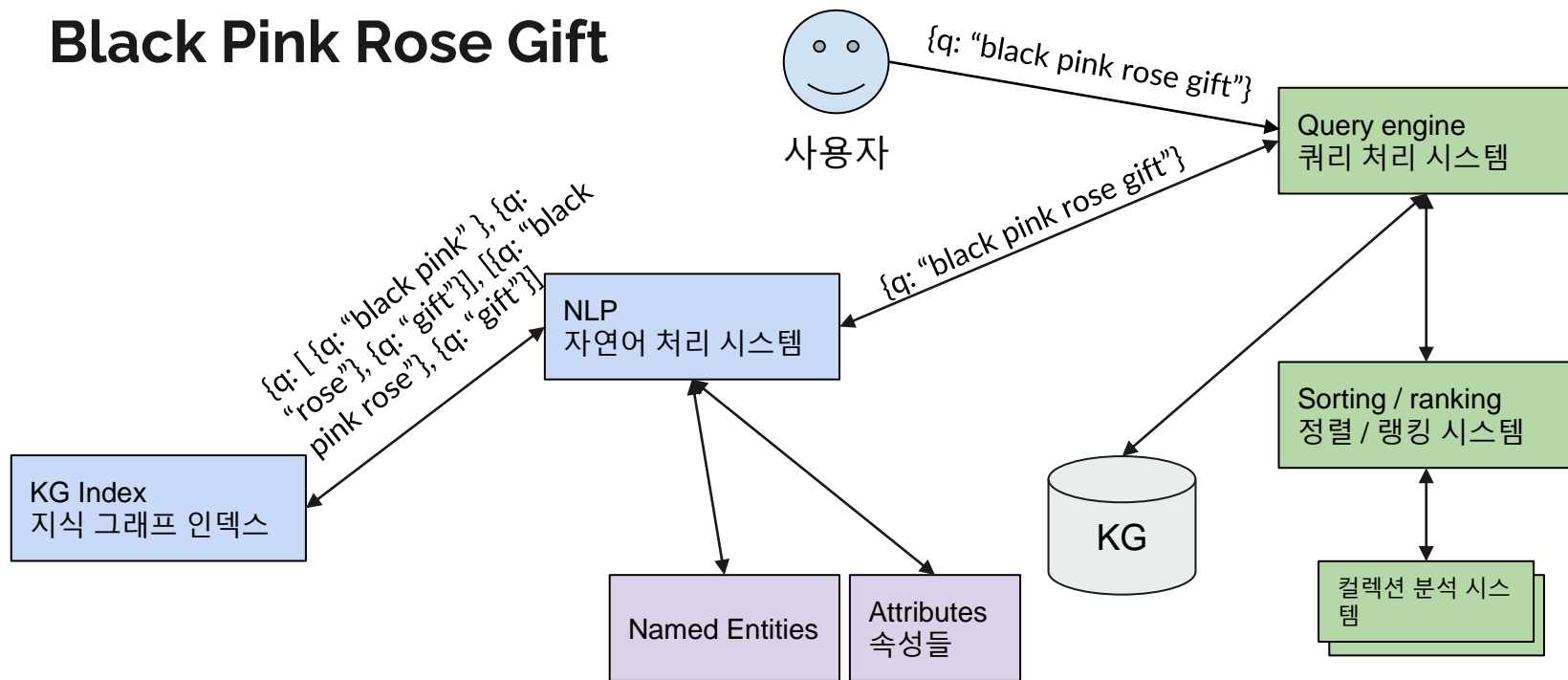
Ivanka Trump's Age



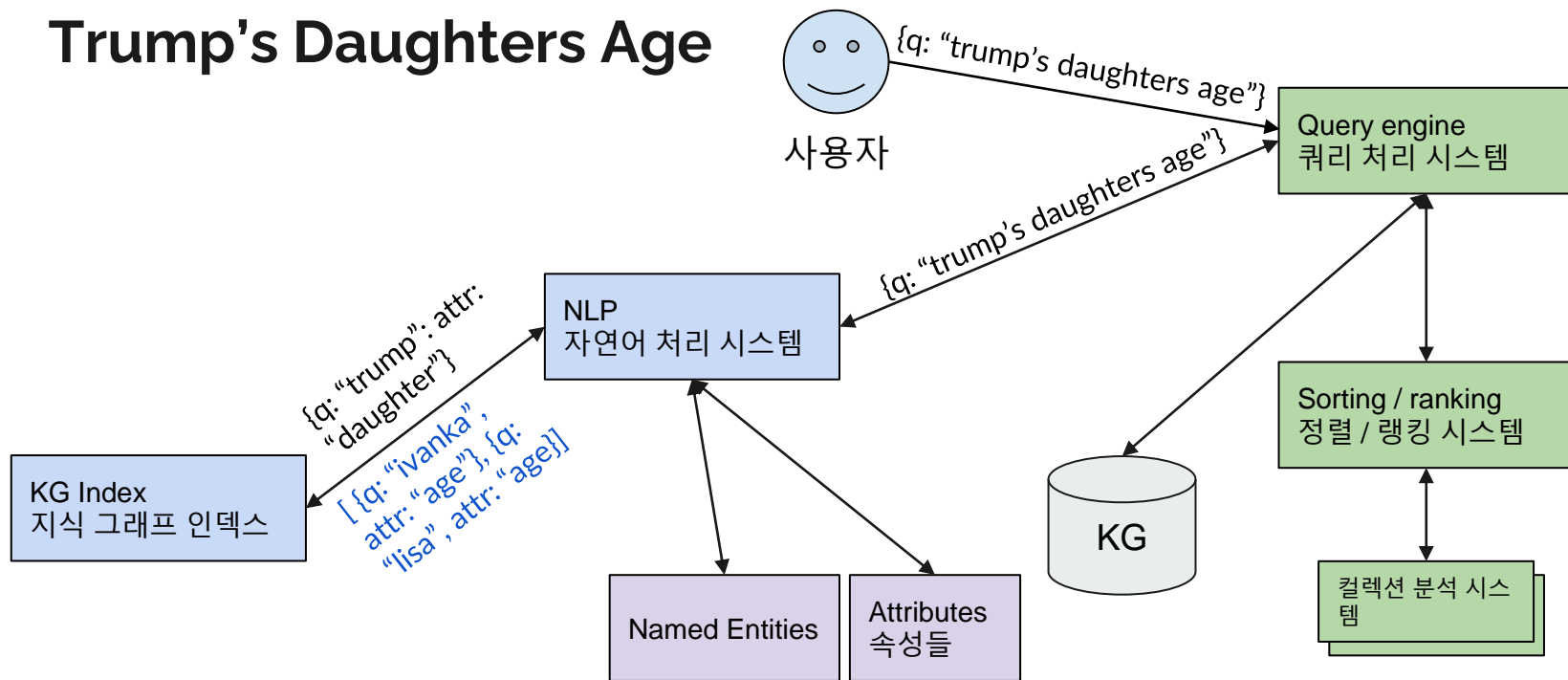
Ivanka Trump's Age



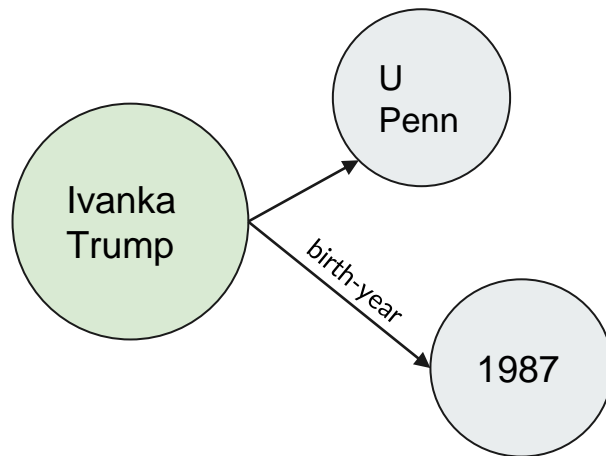
Black Pink Rose Gift



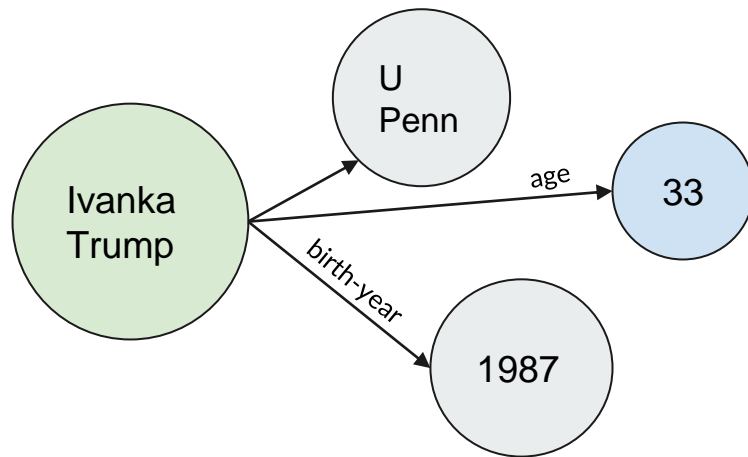
Trump's Daughters Age



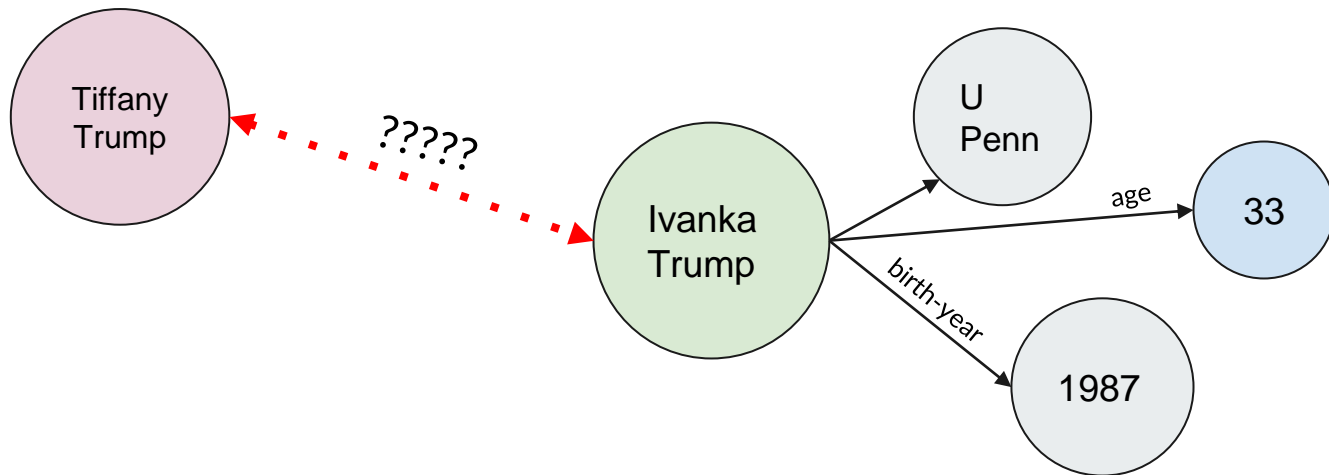
오프라인 속성생성 [Offline attribute generation]



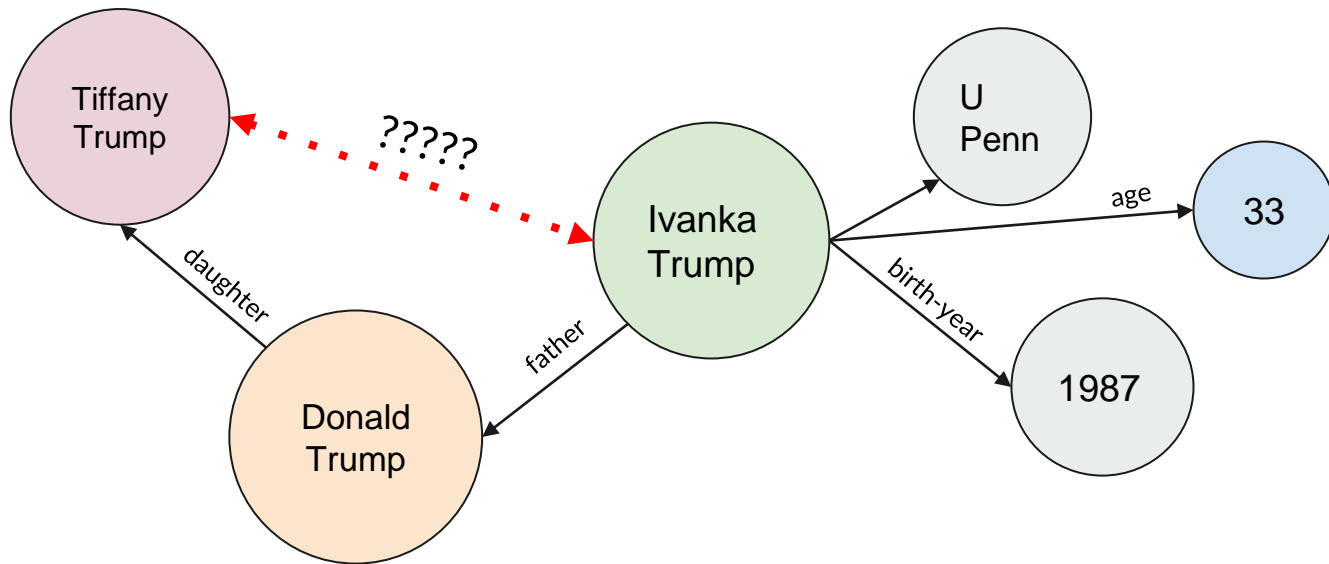
오프라인 속성생성 [Offline attribute generation]



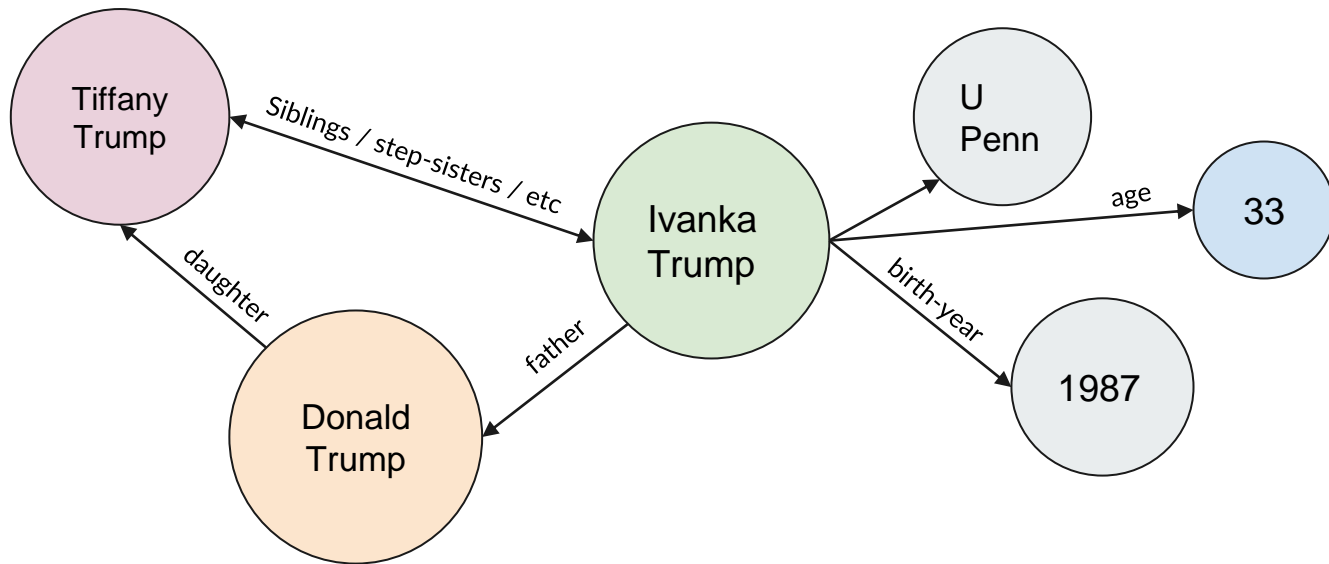
오프라인 속성생성 [Offline attribute generation]



오프라인 속성생성 [Offline attribute generation]



오프라인 속성생성 [Offline attribute generation]



Trump's daughter's age

