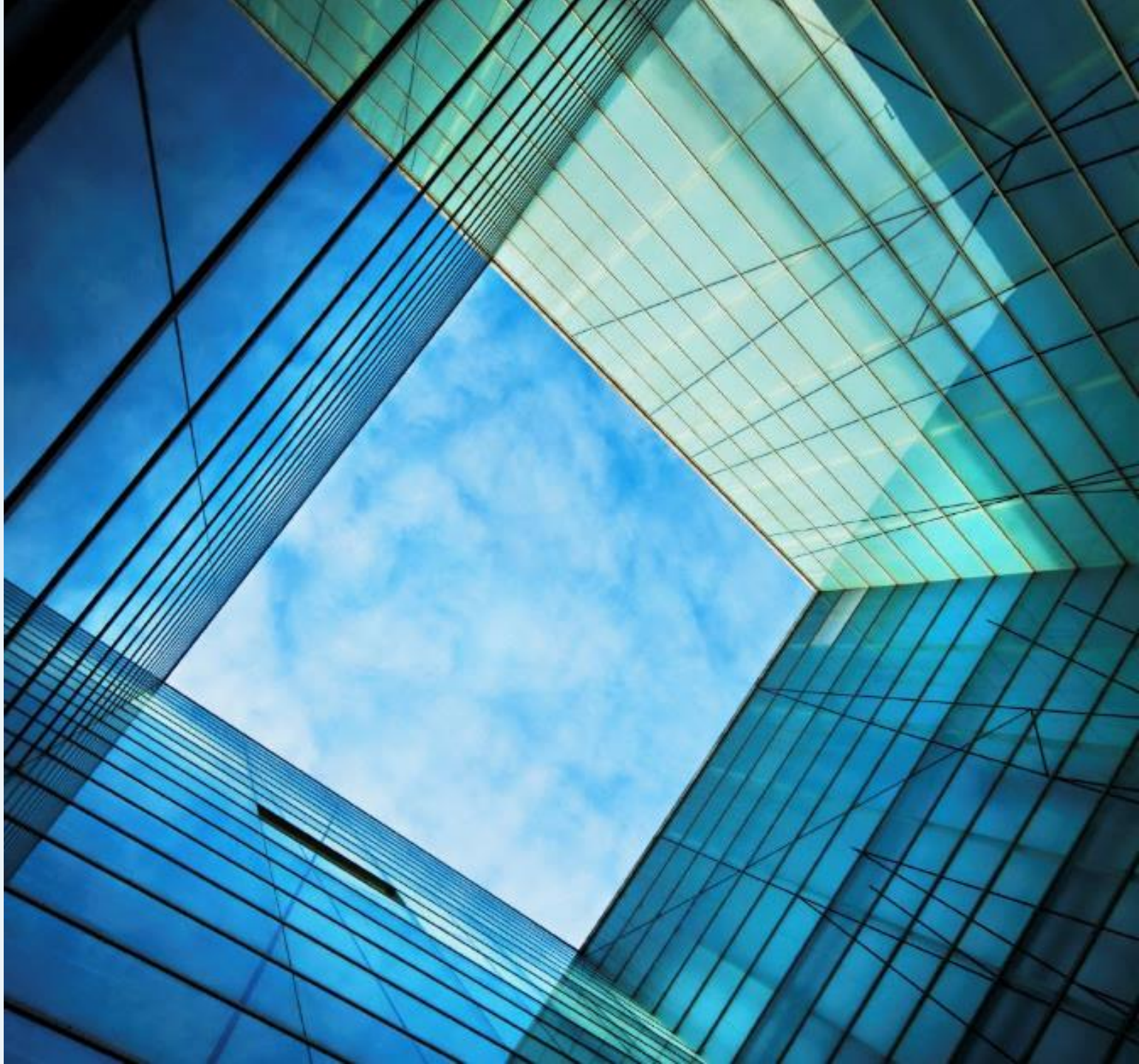
Information Retrieval and Extraction

Term Project 1

NTU CSIE, Fall 2017
授課教師: 陳信希 教授
助教: 顏安孜 陳重吉
{d04922005, d05922016}@ntu.edu.tw



Data

- (Subset of SIGIR paper) Download from https://drive.google.com/file/d/0B5Uu9BOINP_ldzI2UlBkTXZoSHM/view?usp=sharing
- Paper: DBpedia-Entity v2: A Test Collection for Entity Search

  http://hasibi.com/files/sigir2017-dbpedia_entity.pdf

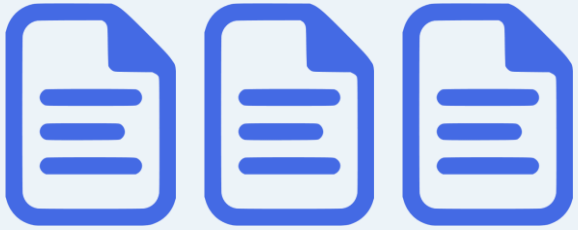
Project 1

Retrieve entity for each query

467 queries

45668 DBpedia Documents

Relevant entity



.....

Data Format

- queries-v2.txt

ID	**Ex:** INEX_LD-20120512
query	**Ex:** south korean girl groups

- qrels-v2.txt

ID	**Ex:** INEX_LD-20120512
Q0	(useless)
entity	**Ex:** <dbpedia:**Girls'_Generation**>
relevance	**Ex: 0,1,2**

(Both txt file are separated by tab)

- DBdoc.json

entity	**Ex:** Girls'_Generation
abstract	**Ex:** Girls' Generation (Hangul: 소녀시대; RR: Sonyeo Sidae), also known as SNSD, is a South Korean girl group formed by S.M. Entertainment. The group is composed of eight members: Taeyeon, Sunny, Tiffany, Hyoyeon, Yuri, Sooyoung, Yoona, and Seohyun……

Query categories

Category	Description	Examples
SemSearch_ES	Named entity queries	"brooklyn bridge", "08 toyota tundra"
INEX-LD	IR-style keyword queries	"electronic music genres"
QALD2	Natural language questions	"Who is the mayor of Berlin?"
ListSearch	Queries that seek a particular list of entities	"Professional sports teams in Philadelphia"

Report – Please use the provided template

- Written in Chinese or English (depend on your native language) with readable font size
- **No more than 6 pages**
- Must include
  - Name and university ID of every teammate
  - Division of work
  - Explore and compare at least **three** models (not limited to the models learned from IR&IE class)(90%)
    - Introduction
    - Methodology
    - Evaluation
    - Discussion
  - Conclusion (10%)
- Bonus (20%)
  - If you adopt the same dataset as the SIGIR 2017 paper, and compare your models with this paper. (10%)
  - If your performance beat the best result in SIGIR 2017 paper. (10%)

Code

- Describe your code
  - write the proper comment for each part and function

Uncommented Code

```
city=raw_input("Enter a city: ")
while city[-1]==" ":
    city = city[:-1]
temp=raw_input("Enter a temperature in Farenheit: ")
temp = float(temp)
temp = (temp - 32.0)*(100.0/180.0)
temp = round(temp,3)
temp = str(temp)
print "In "+city+" it is "+temp+" degrees Celcius!"
```

Commented Code

```
#Alyssa P. Hacker
#fah_to_celsius.py

#collect a city name from user
city=raw_input("Enter a city: ")

#trunacte whitespace
while city[-1]==" ":
    city = city[:-1]

#collect a temp from user
temp=raw_input("Enter a temperature in Farenheit: ")

#convert string to float
temp = float(temp)

#convert Farenheit temp to Celsius temp
temp = (temp - 32.0)*(100.0/180.0)

#truncate to 3 decimal places
temp = round(temp,3)

#recast as string so we can concatenate
temp = str(temp)

#print result!
print "In "+city+" it is "+temp+" degrees Celcius!"
```

Evaluation

- **Mean average precision (MAP)**: the mean of the average precision scores for each query (@100)
- **Normalized discounted cumulative gain (nDCG)**: The nDCG values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm (@10)

- reference: Lecture 5. Retrieval Evaluation

Evaluation Toolkit – (1/2)

- trec_eval

  https://github.com/usnistgov/trec_eval
- Installation: Should be as easy as typing "make" in the **source directory**.
- MAP:

  ./trec_eval –m map qrels-v2.txt <result_file>
- nDCG:

  ./trec_eval –m ndcg_cut qrels-v2.txt <result_file>

Evaluation Toolkit – (2/2)

- result_file format (separated by tab)

query_ID	Q0	<dbpedia:**entity**>	**ranking**	**score**	STANDARD

- Ex:

INEX_LD-2009022 Q0 <dbpedia:Afghan_cuisine> 5 0.3 STANDARD

INEX_LD-2009022 Q0 <dbpedia:Akan_cuisine> 3 0.5 STANDARD

INEX_LD-2009022 Q0 <dbpedia:Ambuyat> 4 0.4 STANDARD

INEX_LD-2009022 Q0 <dbpedia:American_Chinese_cuisine> 2 1 STANDARD

INEX_LD-2009022 Q0 <dbpedia:Ants_climbing_a_tree> 1 2 STANDARD

Submit format

- Project1_team_<team number>.zip
  - Report_team_<team number>.pdf
  - Code_team_<team number>(file)
    - readme.txt (description of each script)
    - $script_1$
    - …
    - $script_n$

Ex:

- Project1_team_0.zip
  - Report_team_0.pdf
  - Code_team_0
    - readme.txt
    - Vector_model.py
    - Probabilistic_model.py
    - main.py

Project 1 presentation

- Date: 11/23
- Please submit your report and presentation slides to CEIBA before 11/21 23:59
- **5 minutes** per group
- Judging Criteria
  - Content
  - State your idea, methodology, evaluation and conclusion clearly and logically

Grading Policy

- Report 70%
- Presentation 30%

Project 1 Schedule

- 10/12 Project 1 release
- 11/17 23:59 Submit code and report to CEIBA
- 11/21 23:59 Submit presentation slides to CEIBA
- 11/23 23:59 Presentation in class

Rule

- We will ask you demo, if
  - TAs could not get the same(similar) result with your code
  - Unclear comment code
- You Can:
  - Use any toolkit
  - Use any library
  - Use any open-source (github)
- **Don’t**
  - **Retrieve documents by yourself**

Some Resources

- List of information retrieval libraries ▫ https://en.wikipedia.org/wiki/List_of_information_retrieval_libraries
- Tf–idf term weighting scikit-learn.org/stable/modules/feature_extraction.html