Overview of the NTCIR-16 Data Search 2 Task

Makoto P. Kato (University of Tsukuba), Hiroaki Ohshima (University of Hyogo), Ying-Hsang Liu (Oslo Metropolitan University), Hsin-Liang Chen (Philadelphia College of Osteopathic Medicine), Yu Nakano (University of Tsukuba)





- **The open data movement is now being accelerated by the expectations for open science and citizen science**
  - Each country strongly encourages the open data movement:
    - Data.gov (United States)
    - Data.gov.uk (United Kingdom)
    - Data.gov.au (Australia)
    - e-Stat (Japan)
- **Besides the governmental portals, there are also thousands of data repositories on the Web**

Demand for a better data search engine

(e.g. Google Dataset Search)

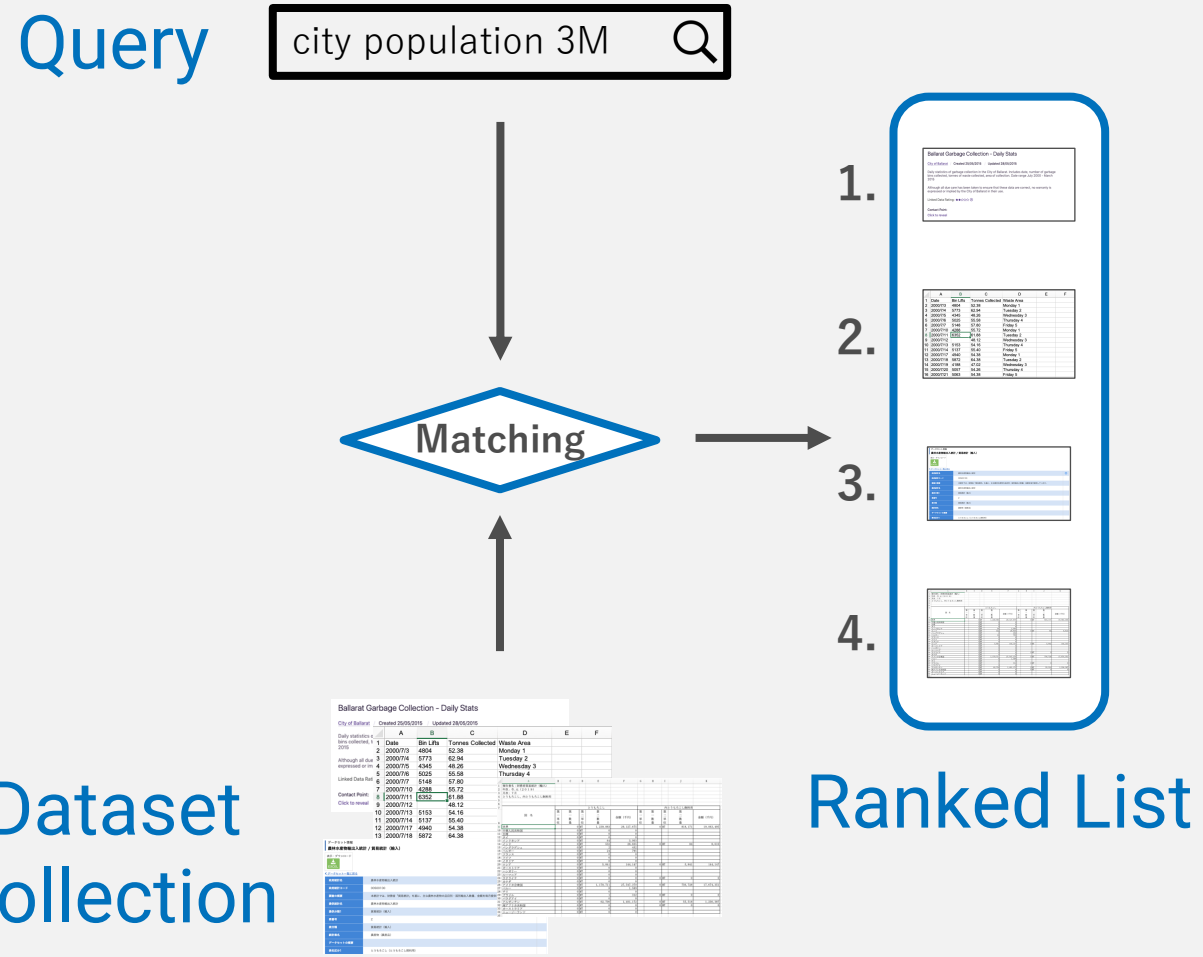
The very first IR evaluation campaign for data search

English	**Documents (or *datasets*)**	46,615
	Training queries	96
	Test queries	96
	Relevance judgments for training queries	2,008
	Relevance judgments for test queries	6,240

Japanese	**Documents (or *datasets*)**	1,338,402
	Training queries	96
	Test queries	96
	Relevance judgments for training queries	2,035
	Relevance judgments for test queries	5,719

Ad-hoc retrieval for statistical data

- **Subtasks**
  - English and Japanese
- **Input**
  - 96 queries for each of the subtasks
- **Document (or Dataset) collection**
  - Data.gov for English
  - e-Stats for Japanese
- **Output**
  - Ranked list of datasets for each query



Q1. What techniques were potentially effective?

- Not very conclusive, but possibly neural models and table understanding

Q2. What queries were difficult in dataset retrieval?

- Time-related queries are especially difficult

Q3. Is the topic variability large?

- Yes, it is much larger than the system variability

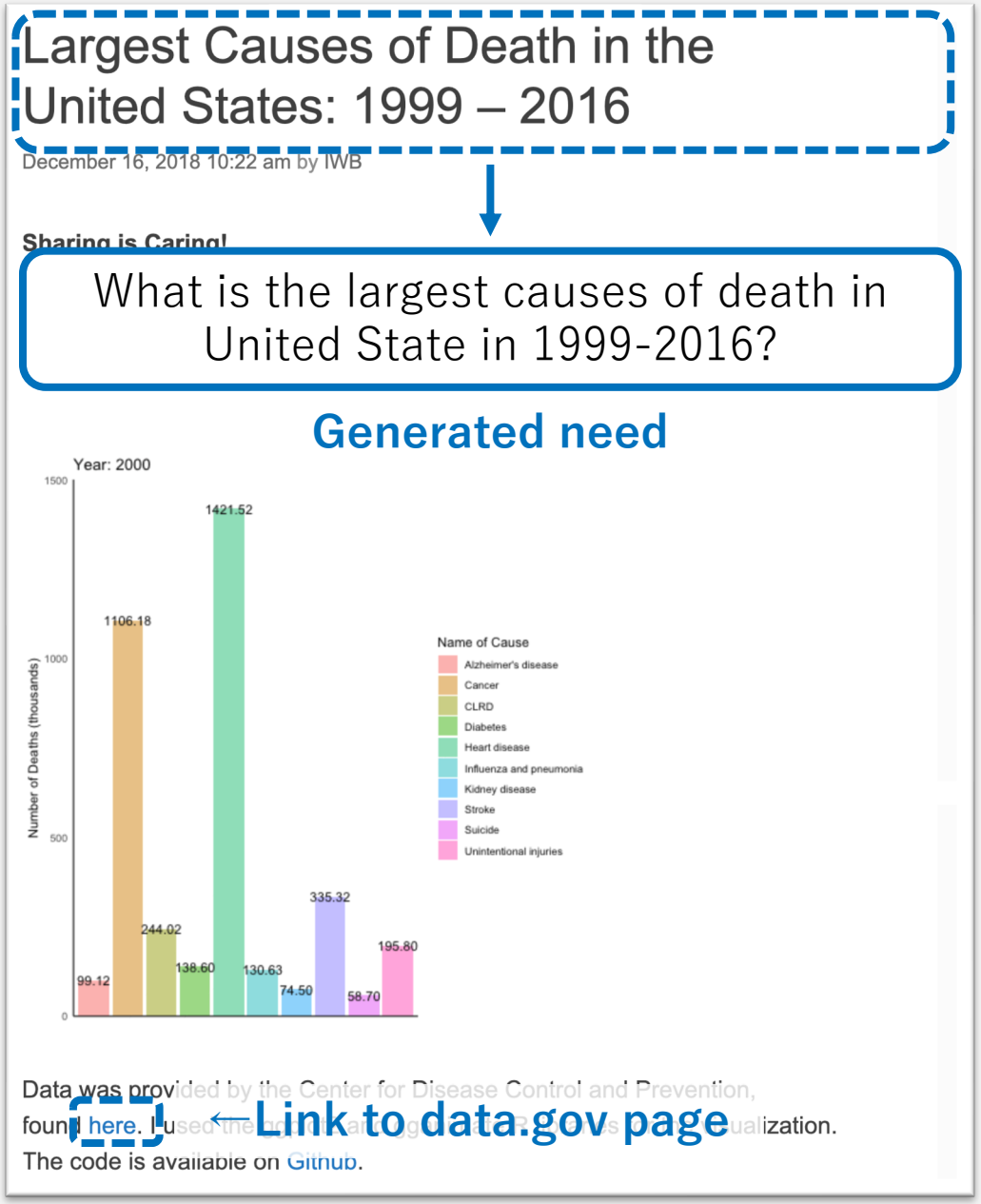
The second round of the "Data Search" task with more queries

English	**Documents (or *datasets*)**	46,615
	Training queries	192
	Test queries	**58**
	Relevance judgments for training queries	8,248
	Relevance judgments for test queries	**6,550**

Japanese	**Documents (or *datasets*)**	1,338,402
	Training queries	192
	Test queries	**72**
	Relevance judgments for training queries	7,754
	Relevance judgments for test queries	**4,035**

- **IR Subtask**
  - Given a query and a dataset collection, a system is expected to generate a ranked list of datasets.
- ~~**QA Subtask**~~
  - ~~Given a question and a dataset, a system is expected to generate an answer to the question, mainly by extracting a part of the dataset.~~
- ~~**UI Subtask**~~
  - ~~Participants are expected to develop a search system with an effective search interface for dataset search tasks.~~

- **NTCIR-15 Data Search**
  - Information needs were derived from questions in cQA
- **NTCIR-16 Data Search 2**
  - Information needs were derived from webpages referring to a dataset
    - Parsed Common Crawl webpages (~ 25B)
    - Identified 47,242 URLs including “data.gov” and 137,388 URLs including “stat.go.jp”.
  - Manually extracted a potential need from the webpages



- **Used crowd-sourcing services to convert information needs to queries**
  - Showed a need and asked workers to input a query without looking at the need
    - Tried to simulate a more realistic situation
  - Selected the most "probable" query from 10 workers' queries
    - Built a unigram language model from those queries, and selected the one with the highest perplexity with respect to the language model

Examples of Need and Queries

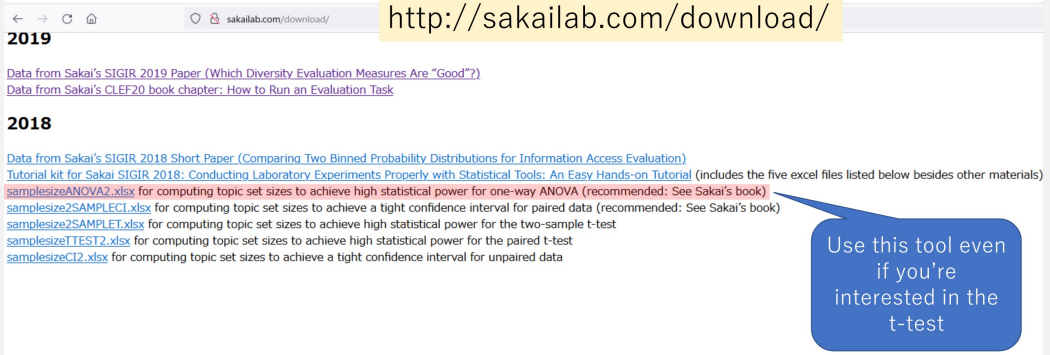
ID	Need	Query
DS2-E-0001	What is the largest causes of death in United State in 1999-2016?	causes of death us 1999-2016
DS2-E-0004	How much is the tuition fee at a private elementary school?	tuition fee private elementary school
DS2-E-0005	Are there hospital differences across the US states?	us states hospital differences

- **See the tutorial given by Prof. Sakai†**
  - Computed the residual variance of nDCG@10 based on the NTCIR-15 Data Search results
  - The minimum topic size* is
    - $n = 36$ for English runs → **58**
    - $n = 68$ for Japanese runs → **72**

queries used in NTCIR-16

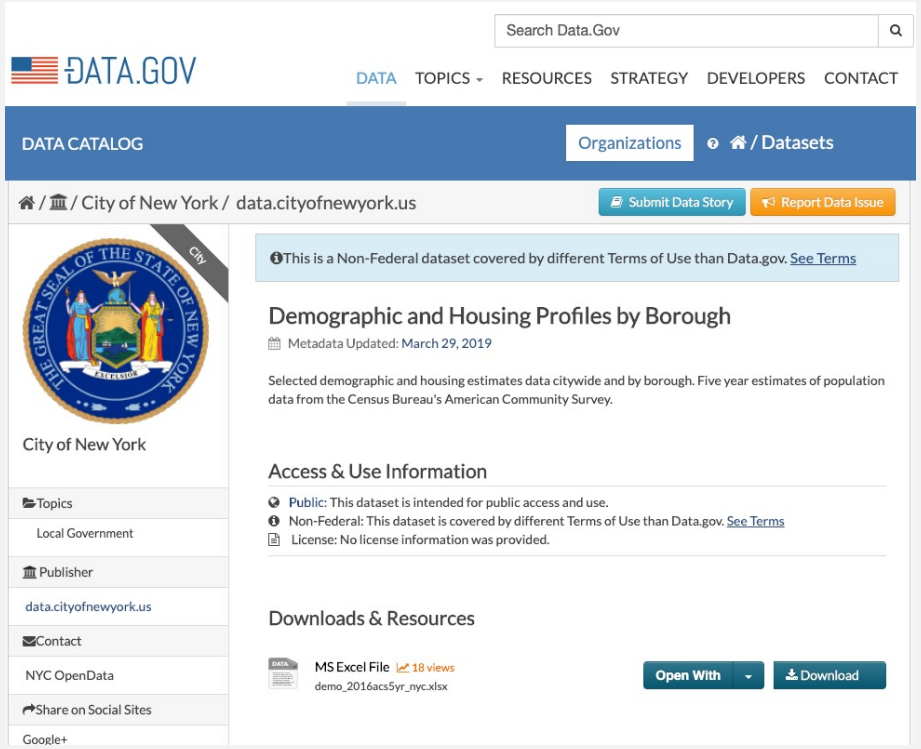
Topic set size design based on the statistical power of the (paired) t-test (1)

http://sakailab.com/download/

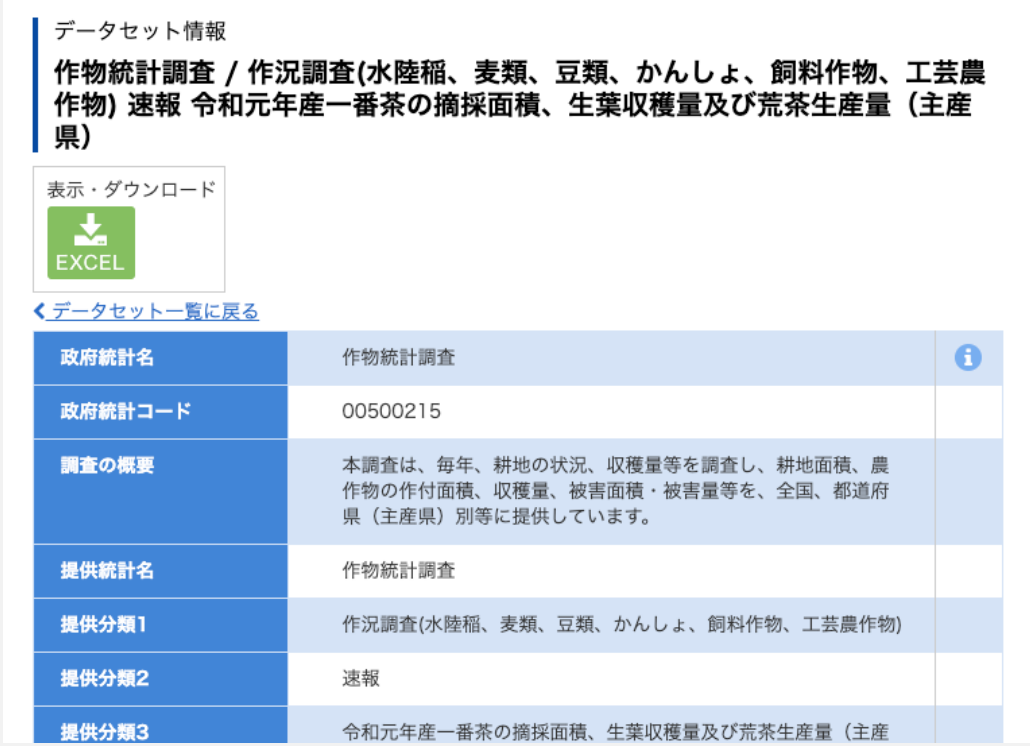


† Sakai. Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power. 2018.
* minD = 0.05, $m = 10$, $\alpha = 0.05$, and $\beta = 0.20$

- **English**
  - Data.gov
    - https://www.data.gov/
    - 46,615 (~445GB)



- **Japanese**
  - e-Stat
    - https://www.e-stat.go.jp/
    - 1,338,402 (~100GB)



Examples of Datasets

Ballarat Garbage Collection - Daily Stats

City of Ballarat / Created 25/05/2015 / Updated 28/05/2015

Daily statistics of garbage collection in the City of Ballarat. Includes date, number of garbage bins collected, tonnes of waste collected, area of collection. Date range July 2000 - March 2015

Although all due care has been taken to ensure that these data are correct, no warranty is expressed or implied by the City of Ballarat in their use.

Linked Data Rating: ★★☆☆☆

Contact Point:

Click to reveal

	A	B	C	D	E	F
1	Date	Bin Lifts	Tonnes Collected	Waste Area		
2	2000/7/3	4804	52.38	Monday 1		
3	2000/7/4	5773	62.94	Tuesday 2		
4	2000/7/5	4345	48.26	Wednesday 3		
5	2000/7/6	5025	55.58	Thursday 4		
6	2000/7/7	5148	57.80	Friday 5		
7	2000/7/10	4288	55.72	Monday 1		
8	2000/7/11	6352	61.88	Tuesday 2		
9	2000/7/12		48.12	Wednesday 3		
10	2000/7/13	5153	54.16	Thursday 4		
11	2000/7/14	5137	55.40	Friday 5		
12	2000/7/17	4940	54.38	Monday 1		
13	2000/7/18	5872	64.38	Tuesday 2		
14	2000/7/19	4188	47.02	Wednesday 3		
15	2000/7/20	5057	54.26	Thursday 4		
16	2000/7/21	5063	54.38	Friday 5		

データセット情報

農林水産物輸出入統計 / 貿易統計（輸入）

表示・ダウンロード

EXCEL

データセット一覧に戻る

政府統計名	農林水産物輸出入統計
政府統計コード	00500100
調査の概要	本統計では、財務省「貿易統計」を基に、主な農林水産物の品目別・国別輸出入数量、金額を毎月提供しています。
提供統計名	農林水産物輸出入統計
提供分類1	貿易統計（輸入）
表番号	2
表分類	貿易統計（輸入）
統計表名	農産物（農産品）
データセットの概要	
表名区分1	とうもろこし（とうもろこし飼料用）

	A	B	C	D	E	F	G	H	I	J	K
1	報告書名：財務省貿易統計（輸入）										
2	年次：令.元（２０１９）										
3	月次：７月										
4	とうもろこし、内とうもろこし飼料用										
5											
6											
7		とうもろこし					内とうもろこし飼料用				
8	国　名	第一単位	第一数量	第二単位	第二数量	金額（千円）	第一単位	第一数量	第二単位	第二数量	金額（千円）
9	世界		0	MT	1,239,883	29,127,675		0	MT	818,171	19,063,495
10	中華人民共和国		0	MT	0	0					
11	台湾		0	MT	0	0					
12	タイ		0	MT	0	0					
13	インドネシア		0	MT	64	3,003					
14	インド		0	MT	532	28,835		0	MT	84	8,610
15	バングラデシュ		0	MT	2	452					
16	ベルギー		0	MT	24	795					
17	フランス		0	MT	0	0					
18	ドイツ		0	MT	0	0					
19	イタリア		0	MT	0	0					
20	ロシア		0	MT	5,841	144,147		0	MT	5,841	144,147
21	オーストリア		0	MT	0	0					
22	ハンガリー		0	MT	0	0					
23	ルーマニア		0	MT	0	0					
24	ウクライナ		0	MT	0	0		0	MT	0	0
25	カナダ		0	MT	0	0					
26	アメリカ合衆国		0	MT	1,170,711	27,547,370		0	MT	756,728	17,674,351
27	ペルー		0	MT	8	1,589					
28	チリ		0	MT	0	0					
29	ブラジル		0	MT	1	312		0	MT	0	0
30	パラグアイ		0	MT	0	0					
31	アルゼンチン		0	MT	62,700	1,401,172		0	MT	55,518	1,236,387
32	南アフリカ共和国		0	MT	0	0		0	MT	0	0
33	オーストラリア		0	MT	0	0					
34	ニュージーランド		0	MT	0	0					
35											

• "Dataset" is a unit of retrieval in Data Search

– Consists of metadata and multiple data files

• Data.gov

– Multiple data files for each metadata

Metadata

Ballarat Garbage Collection - Daily Stats

City of Ballarat / Created 25/05/2015 / Updated 28/05/2015

Daily statistics of garbage collection in the City of Ballarat. Includes date, number of garbage bins collected, tonnes of waste collected, area of collection. Date range July 2000 - March 2015

Although all due care has been taken to ensure that these data are correct, no warranty is expressed or implied by the City of Ballarat in their use.

Linked Data Rating: ★★☆☆☆

Contact Point:

Click to reveal

Data files

	A	B	C	D	E	F
1	Date	Bin Lifts	Tonnes Collected	Waste Area		
2	2000/7/3	4804	52.38	Monday 1		
3	2000/7/4	5773	62.94	Tuesday 2		
4	2000/7/5	4345	48.26	Wednesday 3		
5	2000/7/6	5025	55.58	Thursday 4		
6	2000/7/7	5148	57.80	Friday 5		
7	2000/7/10	4288	55.72	Monday 1		
8	2000/7/11	6362	61.88	Tuesday 2		
9	2000/7/12		48.12	Wednesday 3		
10	2000/7/13	5153	54.16	Thursday 4		
11	2000/7/14	5137	55.40	Friday 5		
12	2000/7/17	4940	54.38	Monday 1		
13	2000/7/18	5872	64.38			
14	2000/7/19	4188	47.02			
15	2000/7/20	5057	54.26			
16	2000/7/21	5063	54.38			

• e-Stat

– A data file for each metadata

Metadata

データセット情報

食料需給表 / 確報 平成16年度食料需給表

表示・ダウンロード

EXCEL

データセット一覧に戻る

政府統計名	食料需給表
政府統計コード	00500300
調査の概要	本統計では、穀類等の品目別の国内生産量、輸出入量、国内消費仕向量、１人当たり供給純食料、食料自給率等を毎年提供しています。
提供統計名	食料需給表
提供分類1	確報
提供分類2	平成16年度食料需給表

Data file

(1) 自給率の推移

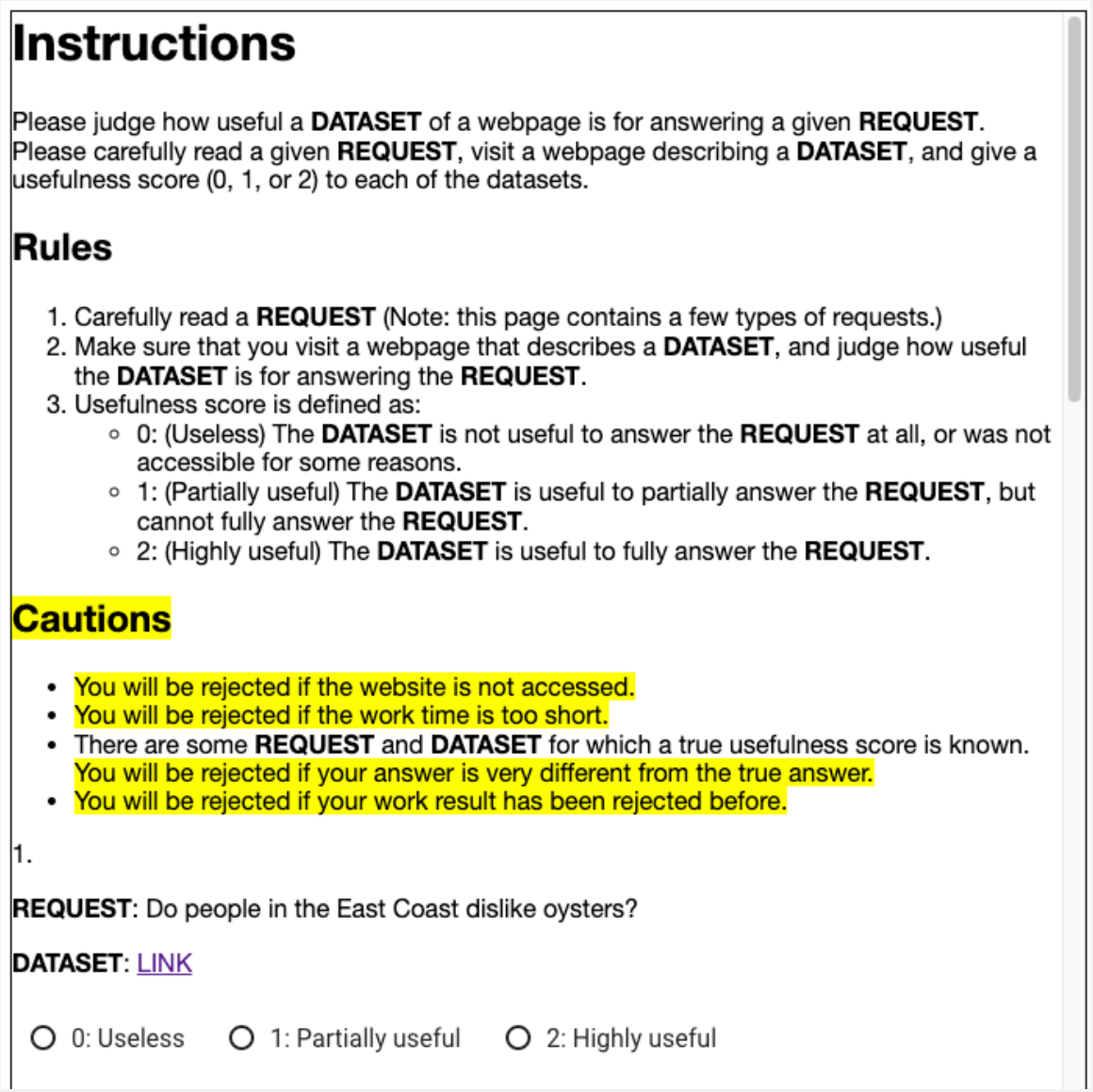
① 品目別自給率の推移

品目	昭和35年度	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	5
米	102	95	98	95	94	95	101	116	118	117	106	100	100	101	102	110	100	1
うち主食用																		
小麦	39	43	38	17	28	28	21	20	20	14	9	8	5	4	4	4	4	
大麦	104	89	85	60	58	57	56	50	46	38	28	23	14	8	9	8	8	
裸麦	112	89	90	28	119	123	90	92	115	98	73	73	64	87	111	98	87	
大・裸麦計	107	89	87	51	70	73	65	59	60	48	34	29	18	10	11	10	9	
雑穀	21	15	11	9	6	5	3	3	2	2	1	1	1	1	1	1	0	
いも類	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	99	98	
かんしょ	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	1
ばれいしょ	101	101	101	100	100	100	100	100	100	100	100	100	100	100	100	99	98	
でんぷん	76	76	71	85	72	67	56	55	56	44	41	36	36	30	23	24	29	
豆類	44	42	35	31	23	25	19	20	17	14	13	11	13	11	11	9	8	
大豆	28	25	21	17	13	11	9	8	7	5	4	4	4	3	4	4	3	
その他の豆類	90	90	76	74	58	70	57	72	61	56	65	49	67	60	59	45	42	
野菜	100	100	100	100	100	100	100	100	100	100	99	99	99	98	98	99	98	
果実	100	98	97	92	90	90	89	89	88	85	84	81	81	83	83	84	81	
みかん	111	113	111	109	109	109	107	109	105	106	105	106	103	104	104	102	103	1
りんご	102	102	102	101	102	102	102	102	103	102	102	101	101	101	100	100	100	1
畜産	93	95	95	90	92	93	91	86	85	85	89	84	82	80	84	76	75	

- **The relevance of each dataset for a given query is judged by crowd-sourcing workers**
  - 0: Not-relevant
  - 1: Partially relevant
  - 2: Highly relevant

- **Inter-rater agreement**
  (measured by Krippendorff's α)
  - English: **0.444**
  - Japanese: **0.474**

    (Fairly consistent with those of NTCIR-15)



Instructions

Please judge how useful a **DATASET** of a webpage is for answering a given **REQUEST**. Please carefully read a given **REQUEST**, visit a webpage describing a **DATASET**, and give a usefulness score (0, 1, or 2) to each of the datasets.

Rules

1. Carefully read a **REQUEST** (Note: this page contains a few types of requests.)
2. Make sure that you visit a webpage that describes a **DATASET**, and judge how useful the **DATASET** is for answering the **REQUEST**.
3. Usefulness score is defined as:
   - 0: (Useless) The **DATASET** is not useful to answer the **REQUEST** at all, or was not accessible for some reasons.
   - 1: (Partially useful) The **DATASET** is useful to partially answer the **REQUEST**, but cannot fully answer the **REQUEST**.
   - 2: (Highly useful) The **DATASET** is useful to fully answer the **REQUEST**.

Cautions

- You will be rejected if the website is not accessed.
- You will be rejected if the work time is too short.
- There are some **REQUEST** and **DATASET** for which a true usefulness score is known. You will be rejected if your answer is very different from the true answer.
- You will be rejected if your work result has been rejected before.

1.

REQUEST: Do people in the East Coast dislike oysters?

DATASET: LINK

○ 0: Useless ○ 1: Partially useful ○ 2: Highly useful

- **Applied standard retrieval models to only the metadata**

- **Baseline retrieval models**
  - BM25, BM25 + RM3, BM25 + SDM, BM25 + BM25PRF
  - Query Likelihood, Query Likelihood + RM3, Query Likelihood + SDM

- **NTCIR-16 Data Search attracted seven research groups and received 48 systems' results in total**
  - Including the organizers' team for providing baseline runs
  - 25 for English and 23 for Japanese
- **Participants**
  - UHGSIS: University of Hyogo
  - STIS: Politeknik Statistika STIS
  - WUT21: Wuhan University of Technology
  - KSU: Kyoto Sangyo University
  - NYUCIN: Universidade Federal de Pernambuco
  - OUHCIR: The University of Oklahoma

- **Data (14/48 runs):**
  - Whether the data files are used
- **Neural (22/48 runs):**
  - Whether neural language models (e.g., BERT) are used
- **Entity (10/48 runs):**
  - Whether entities are treated differently from the other tokens
- **Num. (0/48 runs):**
  - Whether numbers are treated differently from the other tokens

Run ID	Description	Data	Neural	Entity	Num.
KSU-E-1	Category+Table Clipping+Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-3	Category+Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-5	Table Clipping+Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-7	Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-9	Category+Table Header+BM25	Y	N	Y	N
NYUCIN-E-1	BM25 and BERT	Y	Y	N	N
OIRGE-E-1	bm25prf+bm25	N	N	N	N
OIRGE-E-2	bm25	N	N	N	N
OIRGE-E-3	bm25.accurate	N	N	N	N
OIRGE-E-4	sdm+qld	N	N	N	N
OIRGE-E-5	rm3+bm25	N	N	N	N
OIRGE-E-6	qld	N	N	N	N
OIRGE-E-7	sdm+bm25	N	N	N	N
OIRGE-E-8	rm3+qld	N	N	N	N
OIUHCIR-E-1	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N	Y	N	N
OIUHCIR-E-2	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N	Y	N	N
OIUHCIR-E-3	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OIUHCIR-E-4	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OIUHCIR-E-5	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OIUHCIR-E-6	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OIUHCIR-E-7	DOC2VEC	N	N	N	N
OIUHCIR-E-8	DOC2VEC	N	N	N	N
STIS-E-1	prop+bert_score+bm25	Y	Y	N	N
STIS-E-2	prop+bert_score	Y	Y	N	N
wut21-E-1	LM Jelinek Mercer	Y	N	N	N

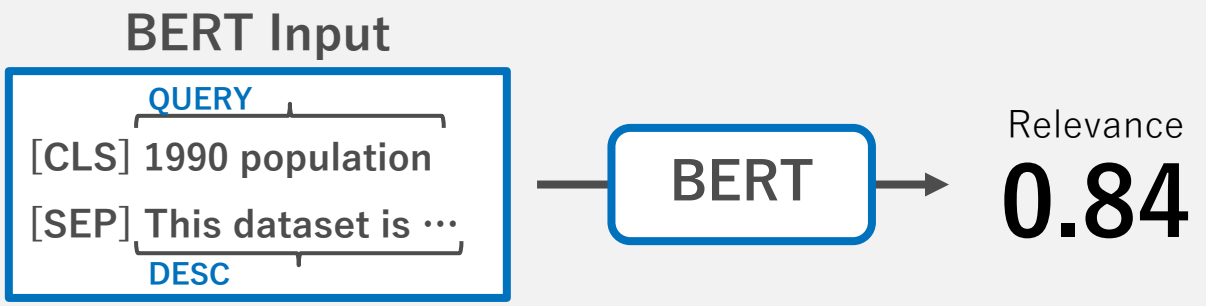
Participants' Runs (English subtask)

Run ID	Description	Data	Neural	Entity	Num.
KSU-E-1	Category+Table Clipping+Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-3	Category+Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-5	Table Clipping+Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-7	Table Header+BERT+MLP	Y	Y	Y	N
KSU-E-9	Category+Table Header+BM25	Y	N	Y	N
NYUCIN-E-1	BM25 and BERT	Y	Y	N	N
ORGE-E-1	bm25prf+bm25	N	N	N	N
ORGE-E-2	bm25	N	N	N	N
ORGE-E-3	bm25.accurate	N	N	N	N
ORGE-E-4	sdm+qld	N	N	N	N
ORGE-E-5	rm3+bm25	N	N	N	N
ORGE-E-6	qld	N	N	N	N
ORGE-E-7	sdm+bm25	N	N	N	N
ORGE-E-8	rm3+qld	N	N	N	N
OUHCIR-E-1	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N	Y	N	N
OUHCIR-E-2	BM25 and TFIDF WEIGHT ADJUSTED and Sentence Transformer	N	Y	N	N
OUHCIR-E-3	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OUHCIR-E-4	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OUHCIR-E-5	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OUHCIR-E-6	BM25 and TFIDF WEIGHT ADJUSTED	N	N	N	N
OUHCIR-E-7	DOC2VEC	N	N	N	N
OUHCIR-E-8	DOC2VEC	N	N	N	N
STIS-E-1	prop+bert_score+bm25	Y	Y	N	N
STIS-E-2	prop+bert_score	Y	Y	N	N
wut21-E-1	LM Jelinek Mercer	Y	N	N	N

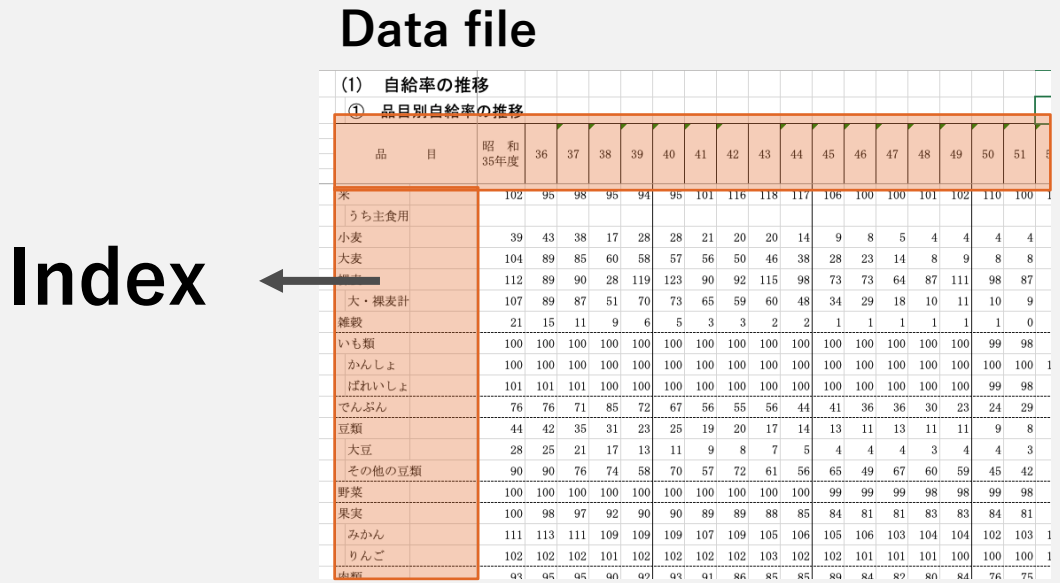
Participants' Runs (Japanese subtask)

Run ID	Description	Data	Neural	Entity	Num.
KSU-J-10	Category+Table Header+BM25	Y	N	Y	N
KSU-J-2	Category+Table Header+BERT+MLP	Y	Y	Y	N
KSU-J-4	Category+Table Header+BERT+MLP	Y	Y	Y	N
KSU-J-6	Table Clipping+Table Header+BERT+MLP	Y	Y	Y	N
KSU-J-8	Table Header+BERT+MLP	Y	Y	Y	N
ORGJ-J-6	qld	N	N	N	N
ORGJ-J-1	bm25prf+bm25	N	N	N	N
ORGJ-J-8	rm3+qld	N	N	N	N
ORGJ-J-7	sdm+bm25	N	N	N	N
ORGJ-J-2	bm25	N	N	N	N
ORGJ-J-5	rm3+bm25	N	N	N	N
ORGJ-J-4	sdm+qld	N	N	N	N
ORGJ-J-3	bm25.accurate	N	N	N	N
UHGSIS-J-9	BM25, BERT, query modification, target 1000	N	Y	N	N
UHGSIS-J-7	BM25, BERT, query modification, target 2000	N	Y	N	N
UHGSIS-J-6	BM25, BERT, query original, target 3000	N	Y	N	N
UHGSIS-J-1	BM25, BERT, query modification, target all	N	Y	N	N
UHGSIS-J-8	BM25, BERT, query original, target 2000	N	Y	N	N
UHGSIS-J-4	BM25, query original, target all	N	N	N	N
UHGSIS-J-3	BM25, query modification, target all	N	Y	N	N
UHGSIS-J-2	BM25, BERT, query original, target all	N	Y	N	N
UHGSIS-J-5	BM25, BERT, query modification, target 3000	N	Y	N	N
UHGSIS-J-10	BM25, BERT, query original, target 1000	N	Y	N	N

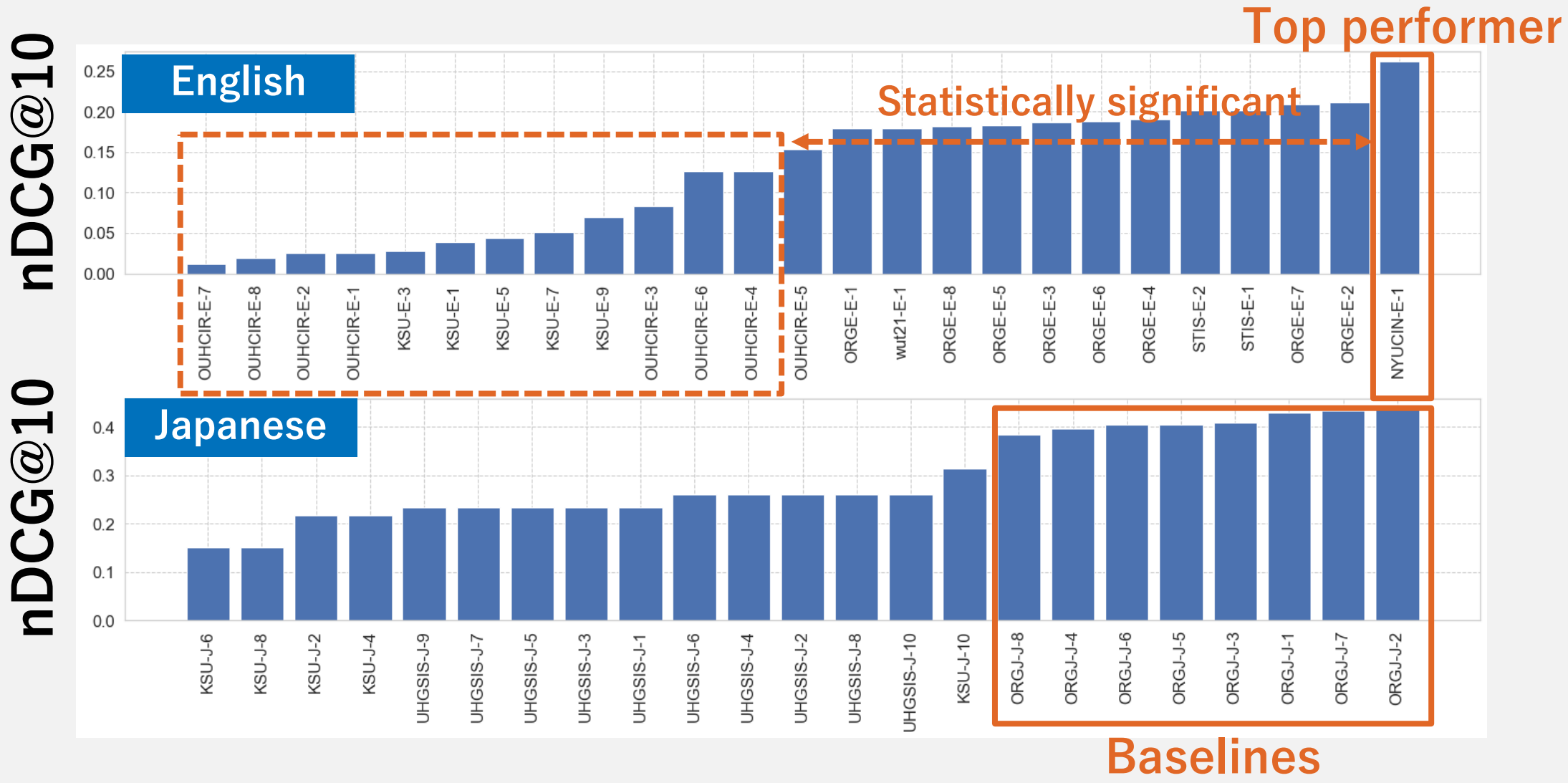
- **Neural : BERT-based relevance estimation**
  - Input a query and a description of a dataset into BERT for estimating the query-dataset relevance



- **Data: Table header extraction**
  - Index terms in the table header together with the description of a dataset



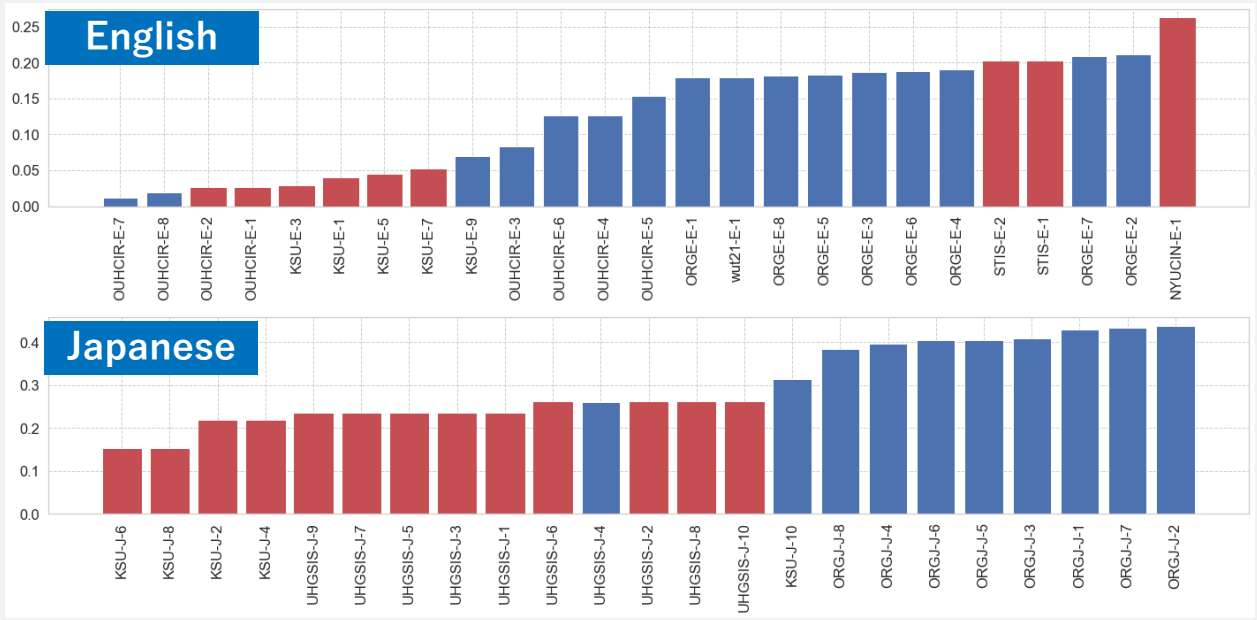
Overall Results



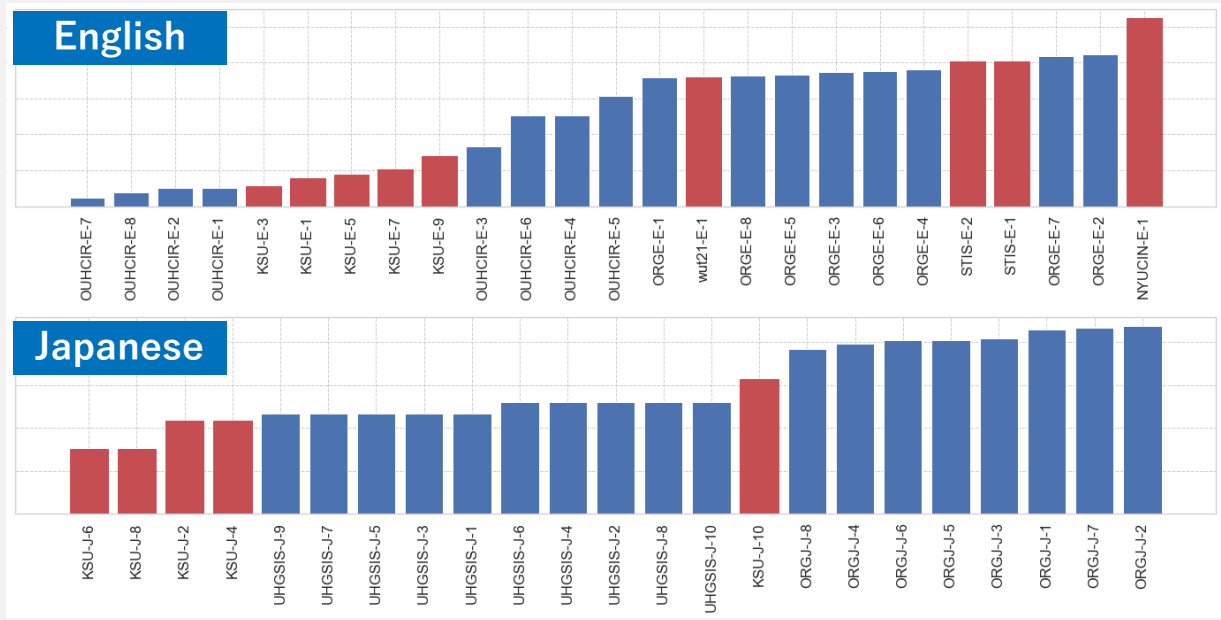
NYUCIN-E-1 is the top English run,
though it is not significantly different from the other top runs including baselines

Q1. What techniques were potentially effective?

"Neural" runs



"Data" runs

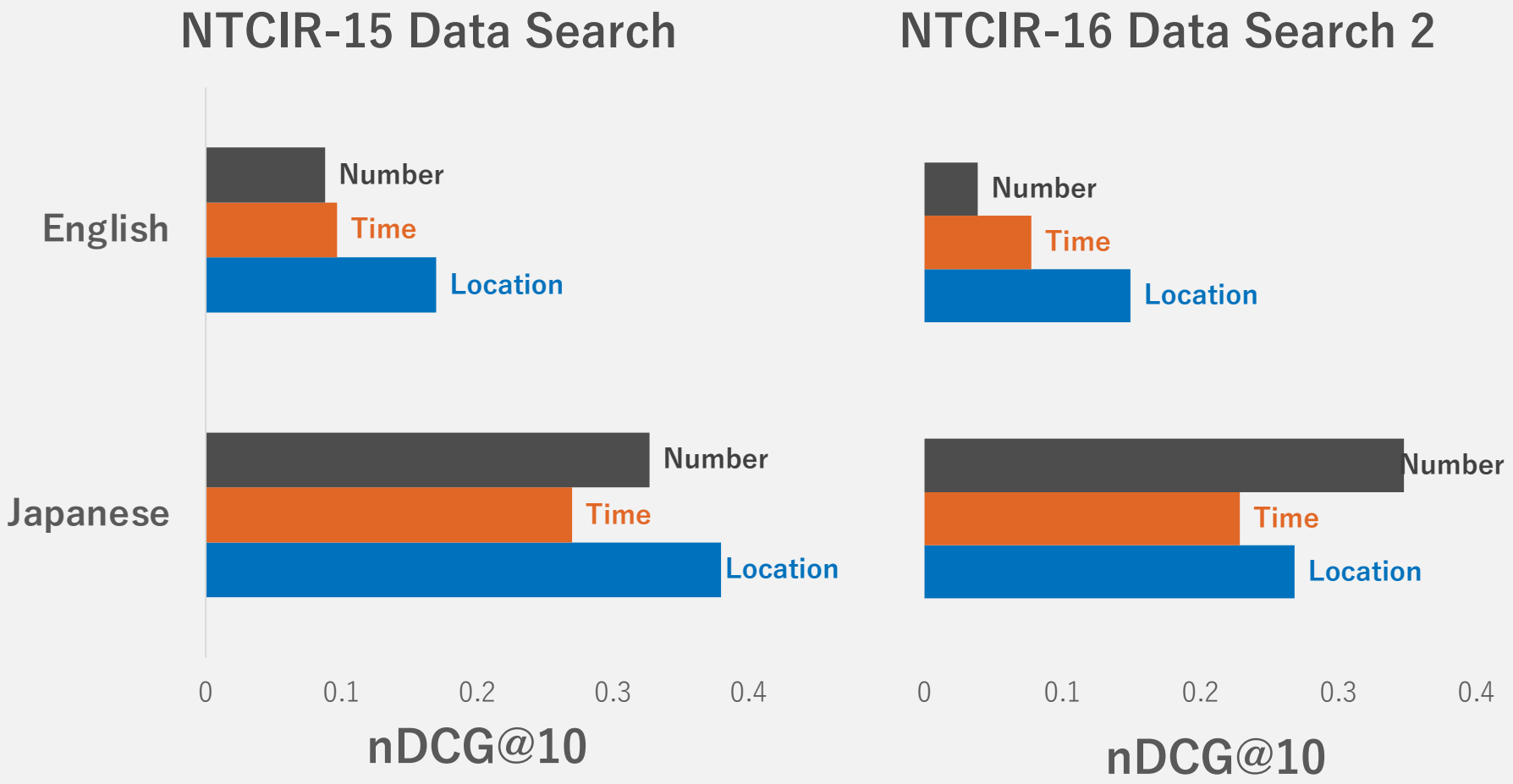


Again, not conclusive, but the top English runs incorporated "Neural" and "Data"

Q2. What queries were difficult in dataset retrieval?

Queries were classified into three types based on what entities are contained

- **Location** (e.g. tokyo population)
- **Time** (e.g. 1990 population)
- **Number** (e.g. 2m population city)



Processing "Time" in queries seems the most difficult

(Fairly consistent with the finding in the previous round)

- **Data Search 2 addressed an ad-hoc retrieval task for datasets**
- **Details of the runs will be discussed at Task Session**

  [JST] 10:30 ~ 11:30 on July 17; [UTC] 01:30 ~ 02:30 on July 17; [EDT] 21:30 ~ 22:30 on July 16

  - 1. NYUCIN at the NTCIR-16 Dataset Search 2 Task
  - 2. KSU Systems at the NTCIR-16 Data Search2 IR Subtask
  - 3. STIS at the NTCIR-16 Data Search Task: Ad-hoc Data Retrieval Ranking with Pretrained Representative Words Prediction

- **Additionally, an invited talk will be given by Prof. Gong Cheng (Nanjing University):**

Towards Content-Based Dataset Search: Test Collections and Beyond