UCLA Samueli
Computer Science

CS145: Introduction to Data Mining (Spring 2024)

Discussion 2: **Project Overview**

Instructor: Dr. Ziniu Hu
Teaching Assistant: Yanqiao Zhu

The Scalable Analytics Institute (ScAI)
Department of Computer Science
University of California, Los Angeles (UCLA)

Announcement

1. HW2 has been released, due next Sunday (EOD)
2. Late due for HW1 at 6PM today with a penalty factor of 50%
3. Reference solution to HW1 will be released under the "Assignments" folder under the "Files" tab after 6PM
4. The reader will start grading HW1 next week
5. Feedback form for the two HWs: https://forms.gle/BvXWz66BzNPw8ASB6
6. Baseline code for all three course projects will be released soon

Open academic graph challenge

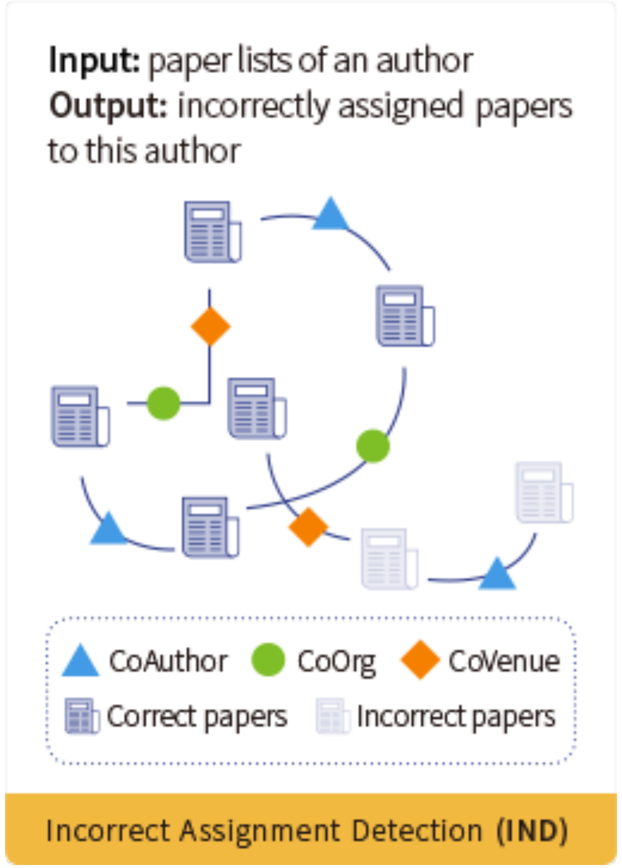
KDD2024
OAG-Challenge

- Academic data mining aims to deepen our understanding of science’s development, nature, and trends
- It offers the potential to unlock enormous scientific, technological, and educational value
- Mining academic data can assist in:
  - Government scientific policy-making
  - Company talent discovery
  - Researchers acquiring new knowledge more efficiently

- Homepage: https://www.biendata.xyz/kdd2024/#overview

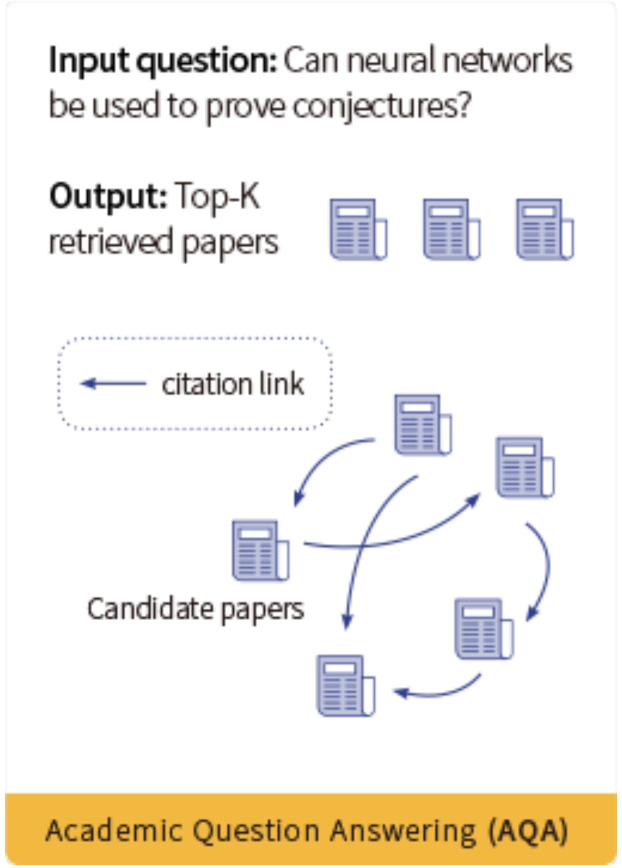
Open academic graph challenge

KDD2024
OAG-Challenge



Input: paper lists of an author
Output: incorrectly assigned papers to this author

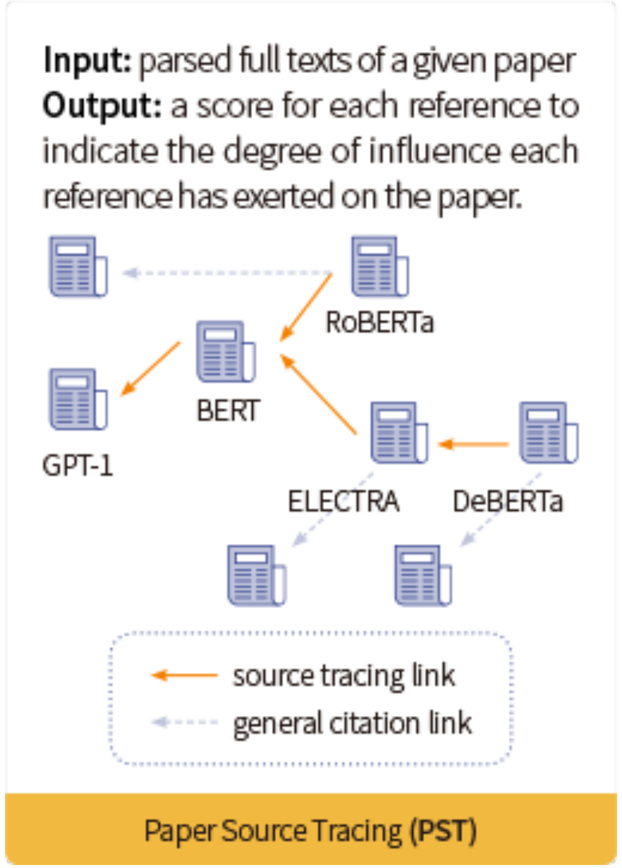
Incorrect Assignment Detection **(IND)**



Input question: Can neural networks be used to prove conjectures?

Output: Top-K retrieved papers

Academic Question Answering **(AQA)**



Input: parsed full texts of a given paper
Output: a score for each reference to indicate the degree of influence each reference has exerted on the paper.

Paper Source Tracing **(PST)**

Open academic graph challenge

- Large language model policy:
  - For all tracks, pre-trained models that have been open-sourced before the end of the competition are allowed to be used
  - WhoIsWho-IND and PST allow the use of APIs
  - After a valid submission to the validation set, participating teams can obtain a free quota of 1 million tokens for the GLM-4 API
- Important deadlines:

Week 2	April 12	Team formation
Week 3	April 19	Dummy submission
Week 10	June 3 & 5	In-class project presentation
Week 10	June 5	Report & code submission

Open academic graph challenge

- Team formation:
  - Team sign-up form: https://1drv.ms/x/s!AsVRzCssZoYOiZ8UBRIpK8g-i-A5sA?e=JqRjHb
  - Deadline: April 14, 11:59PM (extended!)

- Looking for teammates?
  - Use Piazza to collaborate with other classmates and form your teams
  - Check the sign-up form and email the team leader to see if you are a good fit
  - Do NOT email TA regarding your team assignment

Dummy submission

- All teams need to submit a dummy submission to the contest portal for their chosen task
- Deadline: April 19, 11:59PM
- Steps
  - Run the baseline code provided for your chosen task
  - Prepare the dummy submission file according to the specified format
  - Submit the dummy file to the contest portal
  - Verify that the submission was successful and meets the requirements

WhoIsWho-IND: Task

- Background
  - Increasing online publications make name ambiguity more complex
  - Inaccurate disambiguation results lead to invalid author rankings and award cheating
  - Competition aims to develop models to discover paper assignment errors for given authors

WhoIsWho-IND: Task

- Task
  - Given each author’s profile (name and published papers)
  - Develop a model to detect incorrect paper assignments

- Dataset
  - Paper attributes provided:
    - Title, abstract, authors, keywords, venue, publication year
  - Participants not allowed to use disambiguation results of existing academic search systems

WhoIsWho-IND: Dataset

train_author.json

- Data organized into a dictionary
- Key: author ID
- Value contains:
  - "name": name of the author
  - "normal_data": paper IDs owned by the author
  - "outliers": paper IDs incorrectly assigned to the author

pid_to_info_all.json

- Contains specific paper information for all papers used in the competition
- Data organized into a dictionary
- Key: paper’s ID
- Value: specific paper information

ind_valid_author.json

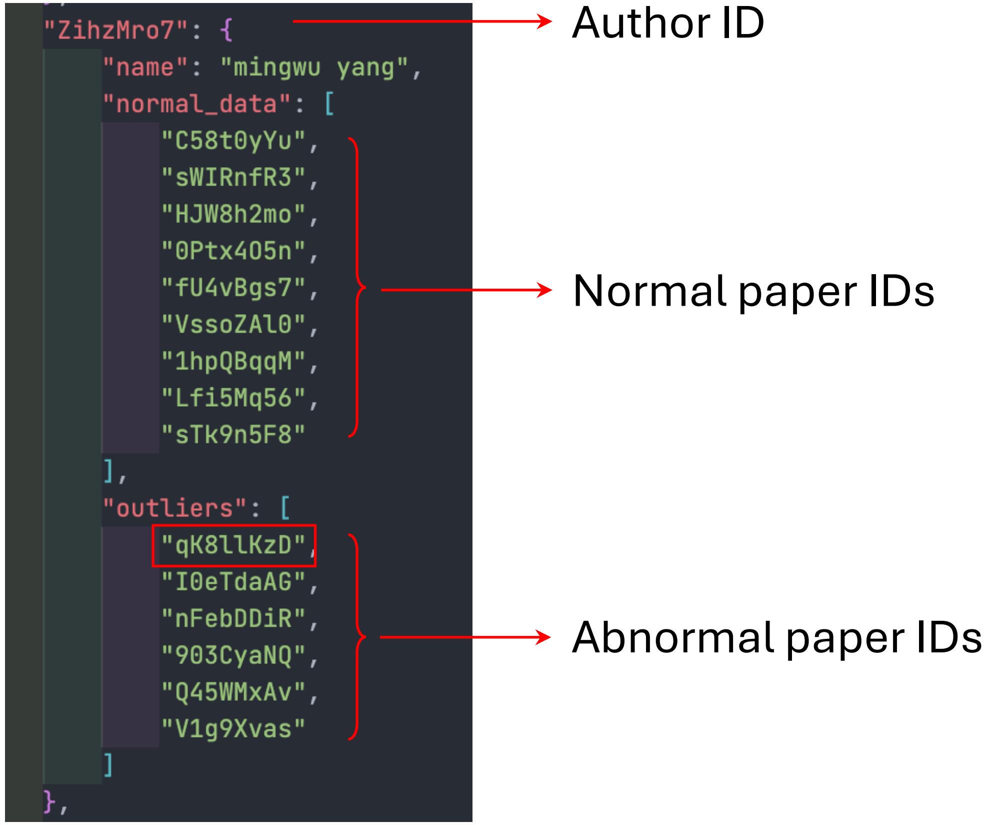
- Organized in a similar format as train_author.json
- "papers" field of each author contains all associated papers of the author

ind_valid_author_submit.json

- Validation set submission example

WhoIsWho-IND: Dataset

train_author.json



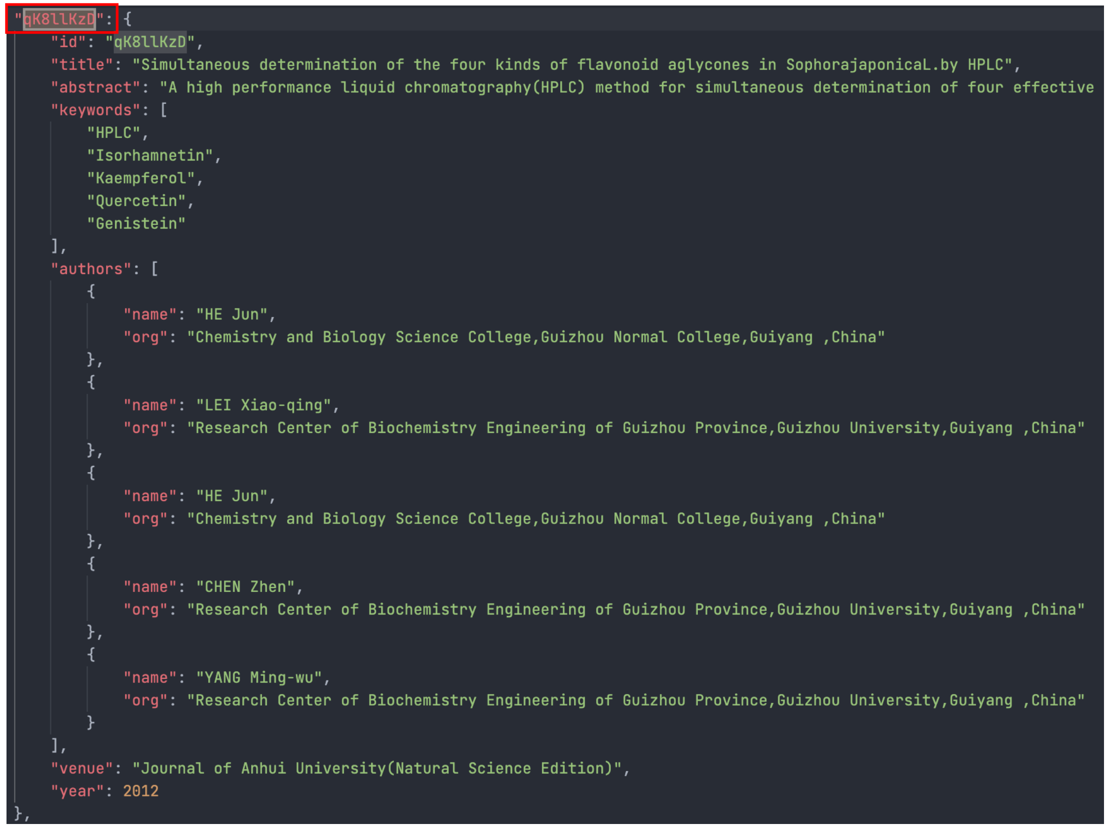
```
"ZihzMro7": {
    "name": "mingwu yang",
    "normal_data": [
        "C58t0yYu",
        "sWIRnfR3",
        "HJW8h2mo",
        "0Ptx405n",
        "fU4vBgs7",
        "VssoZAl0",
        "1hpQBqqM",
        "Lfi5Mq56",
        "sTk9n5F8"
    ],
    "outliers": [
        "qK8llKzD",
        "I0eTdaAG",
        "nFebDDiR",
        "903CyaNQ",
        "Q45WMxAv",
        "V1g9Xvas"
    ]
},
```

Author ID

Normal paper IDs

Abnormal paper IDs

pid_to_info_all.json



```
"qK8llKzD": {
    "id": "qK8llKzD",
    "title": "Simultaneous determination of the four kinds of flavonoid aglycones in SophorajaponicaL.by HPLC",
    "abstract": "A high performance liquid chromatography(HPLC) method for simultaneous determination of four effective
    "keywords": [
        "HPLC",
        "Isorhamnetin",
        "Kaempferol",
        "Quercetin",
        "Genistein"
    ],
    "authors": [
        {
            "name": "HE Jun",
            "org": "Chemistry and Biology Science College,Guizhou Normal College,Guiyang ,China"
        },
        {
            "name": "LEI Xiao-qing",
            "org": "Research Center of Biochemistry Engineering of Guizhou Province,Guizhou University,Guiyang ,China"
        },
        {
            "name": "HE Jun",
            "org": "Chemistry and Biology Science College,Guizhou Normal College,Guiyang ,China"
        },
        {
            "name": "CHEN Zhen",
            "org": "Research Center of Biochemistry Engineering of Guizhou Province,Guizhou University,Guiyang ,China"
        },
        {
            "name": "YANG Ming-wu",
            "org": "Research Center of Biochemistry Engineering of Guizhou Province,Guizhou University,Guiyang ,China"
        }
    ],
    "venue": "Journal of Anhui University(Natural Science Edition)",
    "year": 2012
},
```

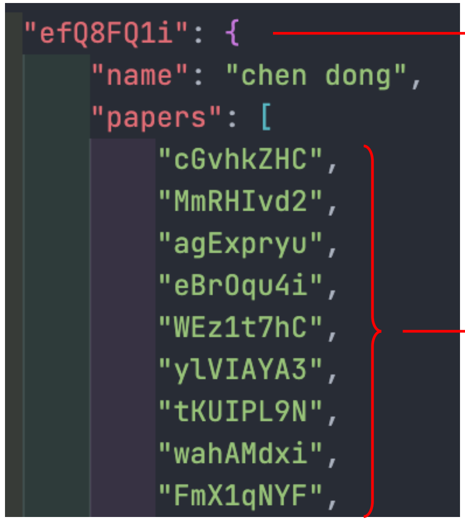
WhoIsWho-IND: Dataset

Column	Type	Description	Example
ID	string	Paper ID	53e9ab9eb7602d970354a97e
title	string	Paper title	Data mining: concepts and techniques
authors.name	string	Author’s name	Jiawei Han
author.org	string	Author’s organization	department of computer science University of Illinois at Urbana Champaign
venue	string	Conference or Journal	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	Publication year	2000
keywords	list of strings	Key words	[“data mining”, “structured data”, “world wide web”, “social network”, “relational data”]
abstract	string	Abstract of a paper	Our ability to generate…

WhoIsWho-IND: Dataset

Your task: predict the abnormal scores of each papers in the validation set

ind_valid_author.json

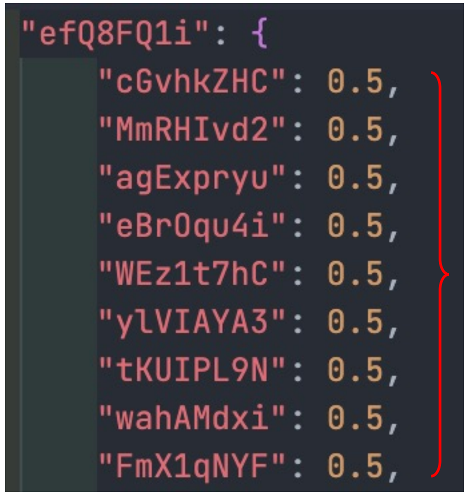


```
"efQ8FQ1i": {
    "name": "chen dong",
    "papers": [
        "cGvhkZHC",
        "MmRHIvd2",
        "agExpryu",
        "eBrOqu4i",
        "WEz1t7hC",
        "ylVIAYA3",
        "tKUIPL9N",
        "wahAMdxi",
        "FmX1qNYF",
```

Author ID

Paper IDs

ind_valid_author_submit.json



```
"efQ8FQ1i": {
    "cGvhkZHC": 0.5,
    "MmRHIvd2": 0.5,
    "agExpryu": 0.5,
    "eBrOqu4i": 0.5,
    "WEz1t7hC": 0.5,
    "ylVIAYA3": 0.5,
    "tKUIPL9N": 0.5,
    "wahAMdxi": 0.5,
    "FmX1qNYF": 0.5,
```

Predict an abnormal score in [0-1] for each paper

Your submission should look like this

WhoIsWho-IND: Evaluation

- Weighted Area Under ROC Curve (AUC) is the evaluation metric
- For each author i,

$$w_i = \frac{\text{Total number of errors}}{\text{Number of errors for author } i}$$

- For all authors:

$$\text{WeightedAUC} = \sum_{i=1}^{M} w_i \cdot \text{AUC}_i$$

  - M is the number of authors

WhoIsWho-IND: Baselines

- Graph-based anomaly detection methods
  - Construct a paper similarity graph based on attribute similarity (e.g., Co-authorship, co-organization)
  - Detect anomalies in the graph
    - Logistic regression (LR): uses top eigenvectors of each graph as features for node classification
    - GCN: employs graph convolutional networks as encoder, followed by fully-connected layers for normal/abnormal node classification
    - GCCAD: leverages graph contrastive learning, contrasting abnormal nodes with normal ones based on distances to global context
- LLM-based methods
  - Fine-tune an existing open-source LLM model
  - Input each author’s paper list and ask the model to identify anomalous papers

AQA: Task

- Background
  - Imperative to provide high-quality and professional knowledge for technical questions
  - Competition challenges participants to develop a model to retrieve the most relevant papers to answer questions from various domains
- Problem formulation
  - Train a model using the OAG-QA derived dataset
    - Contains questions and papers mentioned in the answers
  - OAG-QA retrieves question posts from StackExchange and Zhihu
    - Extracts paper URL mentioned in the answer
    - Matches it with the paper in OAG
  - Participants provided with question datasets
  - Required to find papers that best match these questions

AQA: Dataset

pid_to_title_abs_new.json

- Maps unique identifiers (pids) to papers
- Each entry follows the format:
  - pid: {"title": "abstract":}
- Total of 352,651 papers

qa_train.txt

- Training dataset consisting of dictionaries of questions
- Contains 8,757 entries
- Each entry follows the format:
  - {"question": the general question, "body": specifications on the general question, "pids": ground-truth paper IDs provided for model training}

qa_valid_wo_ans.txt

- Validation dataset with the same format as the training dataset
- Contains 2,919 entries
- Each entry formatted as:
  - {"question": the general question, "body": specifications on the general question}

qa_test_wo_ans.txt

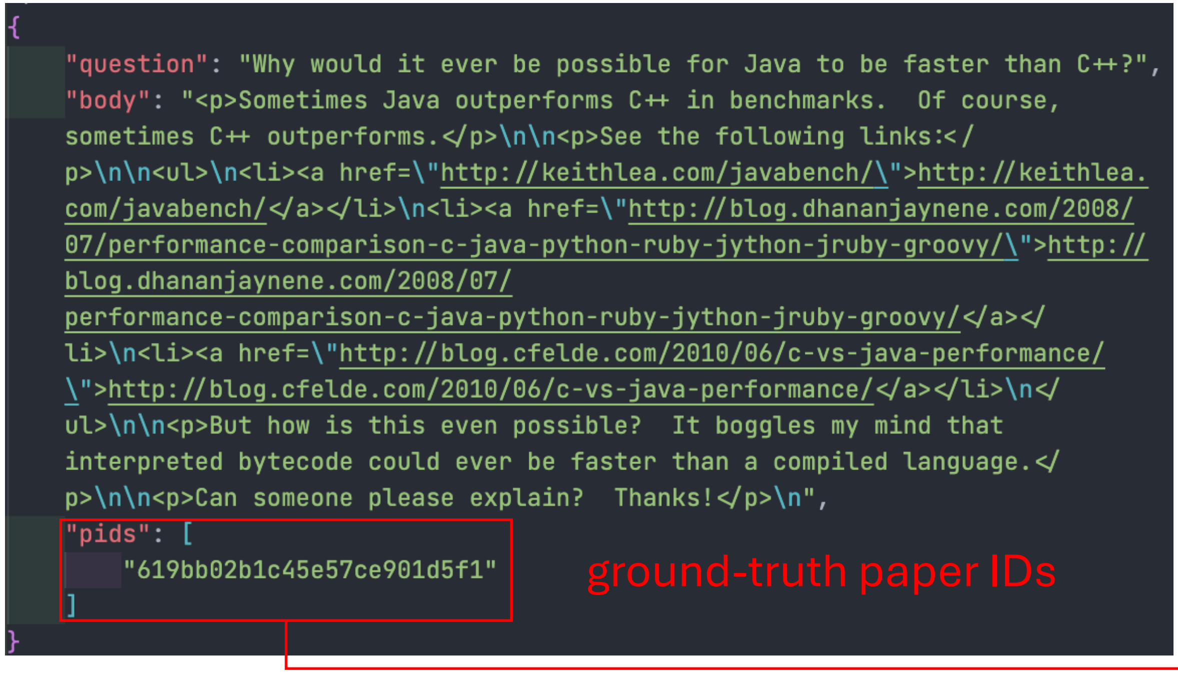
- Additional unlabeled data as a supplement
- Corresponding paper ID for the answer not provided
- Format same as the validation set data
- Participants can decide how to use it

result.txt

- Submission example file for validation
- Participants required to upload a text file returning pids of papers matching questions in the validation set
- Each line should contain the top 20 pids that best match the specific question provided in that line
- File should be in .txt format
- Split pids in each line with English commas

AQA: Dataset

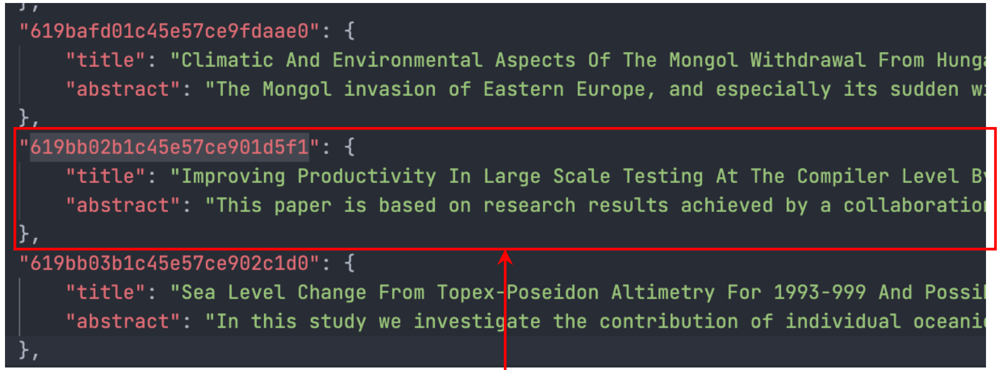
qa_train.txt



```
{
    "question": "Why would it ever be possible for Java to be faster than C++?",
    "body": "<p>Sometimes Java outperforms C++ in benchmarks.  Of course,
    sometimes C++ outperforms.</p>\n\n<p>See the following links:</
    p>\n\n<ul>\n<li><a href=\"http://keithlea.com/javabench/\">http://keithlea.
    com/javabench/</a></li>\n<li><a href=\"http://blog.dhananjaynene.com/2008/
    07/performance-comparison-c-java-python-ruby-jython-jruby-groovy/\">http://
    blog.dhananjaynene.com/2008/07/
    performance-comparison-c-java-python-ruby-jython-jruby-groovy/</a></
    li>\n<li><a href=\"http://blog.cfelde.com/2010/06/c-vs-java-performance/
    \">http://blog.cfelde.com/2010/06/c-vs-java-performance/</a></li>\n</
    ul>\n\n<p>But how is this even possible?  It boggles my mind that
    interpreted bytecode could ever be faster than a compiled language.</
    p>\n\n<p>Can someone please explain?  Thanks!</p>\n",
    "pids": [
        "619bb02b1c45e57ce901d5f1"
    ]
}
```

ground-truth paper IDs

pid_to_title_abs_new.json

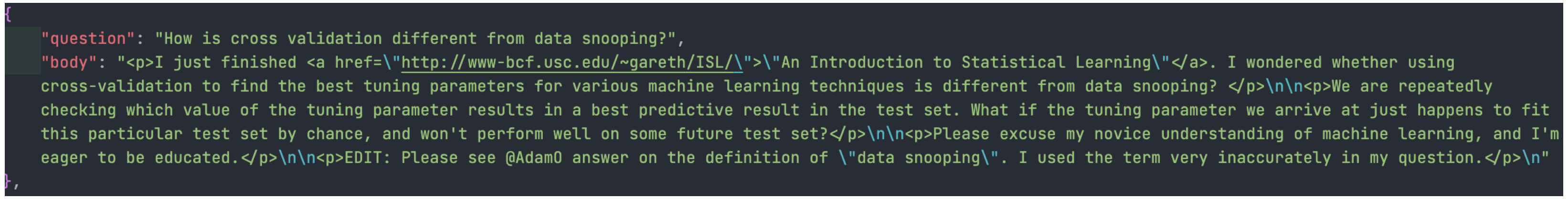


```
"619bafd01c45e57ce9fdaae0": {
    "title": "Climatic And Environmental Aspects Of The Mongol Withdrawal From Hunga
    "abstract": "The Mongol invasion of Eastern Europe, and especially its sudden wi
},
"619bb02b1c45e57ce901d5f1": {
    "title": "Improving Productivity In Large Scale Testing At The Compiler Level By
    "abstract": "This paper is based on research results achieved by a collaboratio
},
"619bb03b1c45e57ce902c1d0": {
    "title": "Sea Level Change From Topex-Poseidon Altimetry For 1993-999 And Possi
    "abstract": "In this study we investigate the contribution of individual oceani
},
```

AQA: Dataset

Your task: find papers that match the questions in the validation set

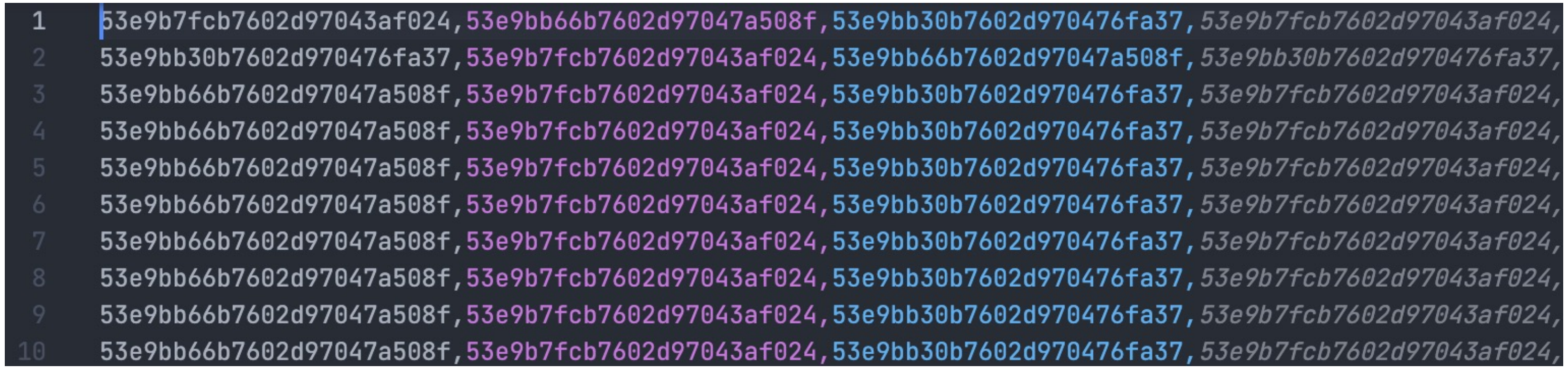
qa_valid_wo_ans.txt



```
{
    "question": "How is cross validation different from data snooping?",
    "body": "<p>I just finished <a href=\"http://www-bcf.usc.edu/~gareth/ISL/\">\"An Introduction to Statistical Learning\"</a>. I wondered whether using
    cross-validation to find the best tuning parameters for various machine learning techniques is different from data snooping? </p>\n\n<p>We are repeatedly
    checking which value of the tuning parameter results in a best predictive result in the test set. What if the tuning parameter we arrive at just happens to fit
    this particular test set by chance, and won't perform well on some future test set?</p>\n\n<p>Please excuse my novice understanding of machine learning, and I'm
    eager to be educated.</p>\n\n<p>EDIT: Please see @AdamO answer on the definition of \"data snooping\". I used the term very inaccurately in my question.</p>\n"
},
```

result.txt

Your submission should look like this



```
53e9b7fcb7602d97043af024,53e9bb66b7602d97047a508f,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,53e9bb66b7602d97047a508f,53e9bb30b7602d970476fa37,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
53e9bb66b7602d97047a508f,53e9b7fcb7602d97043af024,53e9bb30b7602d970476fa37,53e9b7fcb7602d97043af024,
```

Each line matches one question, containing top 20 paper IDs that best match that specific question

AQA: Evaluation

- Evaluation metrics: Mean Average Precision (MAP) and top-K MAP
  - For each question V_q, $\mathrm{AP}(V_q) = \frac{1}{R_q} \sum_{k=1}^{M} P_q(k) 1_k$
    - R_q: number of labeled positive paper ID
    - M: number of papers in the database
    - $P_q(k)$: precision at cut-off k in the ranked list for question V_q
    - 1_k: indicator function (equals 1 if k-th returned paper ID is the groundtruth of the question, otherwise 0)
  - Given n questions, $\mathrm{MAP} = \frac{1}{n} \sum_{q=1}^{n} \mathrm{AP}(V_q)$
  - Top-K MAP calculated similarly by setting M = K in the AP equation

AQA: Baselines

- Sparse retrieval
  - BM25: traditional text retrieval method based on term frequency and inverse document frequency

- Dense retrieval
  - DPR-FT: full fine-tuning of Dense Passage Retriever (DPR)
  - DPR-PT2: parameter-efficient fine-tuning of DPR with P-Tuning v2
  - ColBERT-FT: full fine-tuning of ColBERT
  - ColBERT-PT2: parameter-efficient fine-tuning of ColBERT with P-Tuning v2

PST: Task

- Background
  - Exponential increase in research paper volume due to swift technology advancement
  - Millions of papers published globally every year
  - Difficult for researchers to grasp ins and outs of technological development from numerous sources

PST: Task

- Task
  - Identify reference sources (ref-sources) from full texts of a given paper
  - “Ref-source”: Most important reference (source paper) that greatly inspires the paper
  - Each paper can have one or more ref-sources, or none
  - Rate importance of each reference within [0, 1]
  - Source paper definition:
    - Is the main idea inspired by the reference?
    - Is the core method derived from the reference?
    - Is the reference essential for the paper?
  - Input: XML format file of the paper generated using Grobid API
  - Output: Importance score of each reference to the paper

PST: Dataset

- Training set: paper_source_trace_train_ans.json (788 labeled papers)
- Validation set: paper_source_trace_valid_wo_ans.json (394 labeled papers)
- Data format
  - Training and validation sets are lists of dictionaries, each corresponding to a paper
    - "_id": unique ID value of the paper
    - "title": paper title
    - "refs_trace": list of source papers
    - "authors.name": author name of the paper
    - "authors.org": organization of the author
    - "venue": published journal or conference
    - "year": year of publication
    - "referenced_serial_number": serial number of the important reference in the reference list of the paper
    - "references": all references in the paper

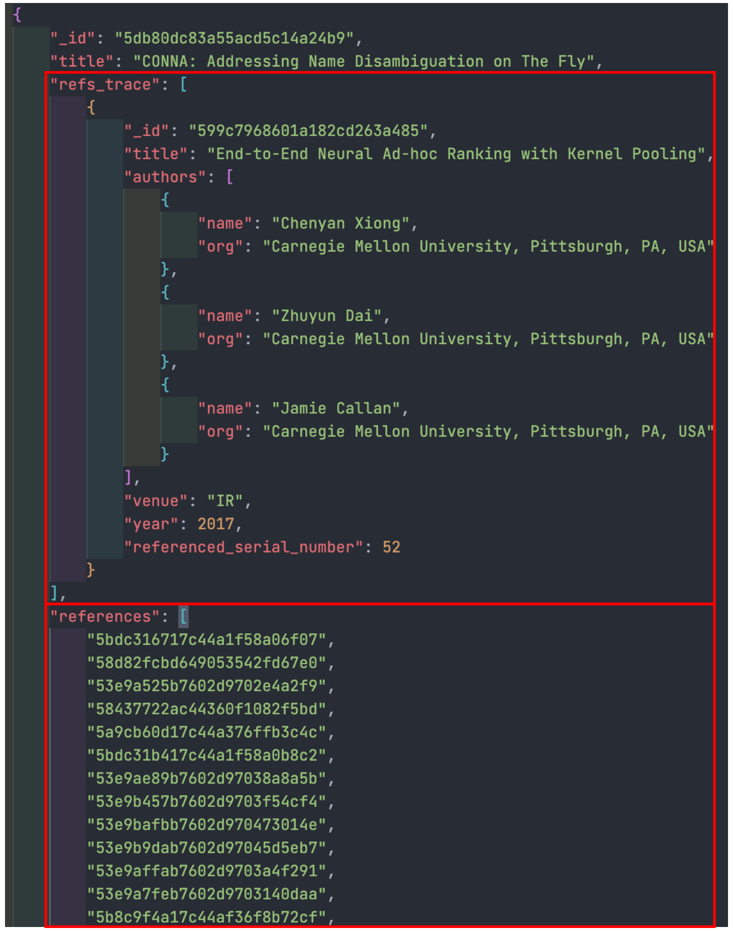
PST: Dataset

- Submission example: submission_example_valid.json
- Full text of papers: Located in the paper-xml folder
  - Each paper's XML file is named {paper ID}.xml
- Additional data: paper_source_gen_by_rule.json (4,854 papers)
  - Key: paper ID
  - Value: dictionary of its source papers
    - Key: serial number of the reference in the paper reference list
    - Value: corresponding paper's title
  - Collected using a rule-based approach (not annotated by experts, correctness not guaranteed)

PST: Dataset

Your task: predict an importance score for each reference to the papers in the validation set

paper_source_trace_train_ans.json



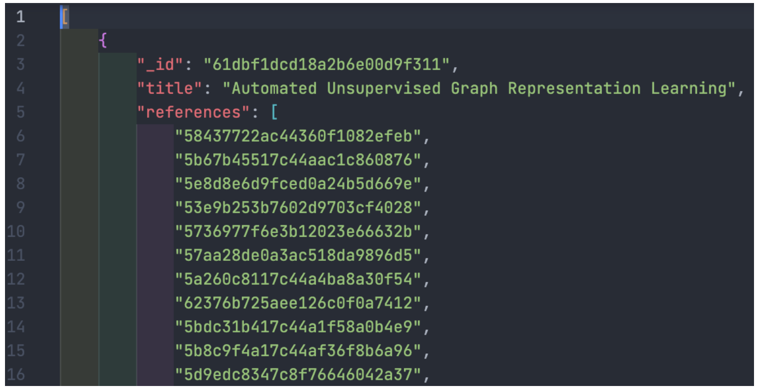
```
{
    "_id": "5db80dc83a55acd5c14a24b9",
    "title": "CONNA: Addressing Name Disambiguation on The Fly",
    "refs_trace": [
        {
            "_id": "599c7968601a182cd263a485",
            "title": "End-to-End Neural Ad-hoc Ranking with Kernel Pooling",
            "authors": [
                {
                    "name": "Chenyan Xiong",
                    "org": "Carnegie Mellon University, Pittsburgh, PA, USA"
                },
                {
                    "name": "Zhuyun Dai",
                    "org": "Carnegie Mellon University, Pittsburgh, PA, USA"
                },
                {
                    "name": "Jamie Callan",
                    "org": "Carnegie Mellon University, Pittsburgh, PA, USA"
                }
            ],
            "venue": "IR",
            "year": 2017,
            "referenced_serial_number": 52
        }
    ],
    "references": [
        "5bdc316717c44a1f58a06f07",
        "58d82fcbd649053542fd67e0",
        "53e9a525b7602d9702e4a2f9",
        "58437722ac44360f1082f5bd",
        "5a9cb60d17c44a376ffb3c4c",
        "5bdc31b417c44a1f58a0b8c2",
        "53e9ae89b7602d97038a8a5b",
        "53e9b457b7602d9703f54cf4",
        "53e9bafbb7602d970473014e",
        "53e9b9dab7602d97045d5eb7",
        "53e9affab7602d9703a4f291",
        "53e9a7feb7602d9703140daa",
        "5b8c9f4a17c44af36f8b72cf",
```

Most important reference

Note: Occasionally these papers might not be included in the references below

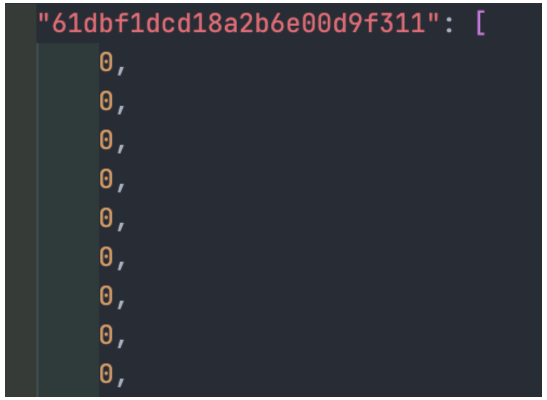
List of all references

paper_source_trace_valid_wo_ans.json



```
[
    {
        "_id": "61dbf1dcd18a2b6e00d9f311",
        "title": "Automated Unsupervised Graph Representation Learning",
        "references": [
            "58437722ac44360f1082efeb",
            "5b67b45517c44aac1c860876",
            "5e8d8e6d9fced0a24b5d669e",
            "53e9b253b7602d9703cf4028",
            "5736977f6e3b12023e66632b",
            "57aa28de0a3ac518da9896d5",
            "5a260c8117c44a4ba8a30f54",
            "62376b725aee126c0f0a7412",
            "5bdc31b417c44a1f58a0b4e9",
            "5b8c9f4a17c44af36f8b6a96",
            "5d9edc8347c8f76646042a37",
```

submission_example_valid.json



```
"61dbf1dcd18a2b6e00d9f311": [
    0,
    0,
    0,
    0,
    0,
    0,
    0,
    0,
    0,
```

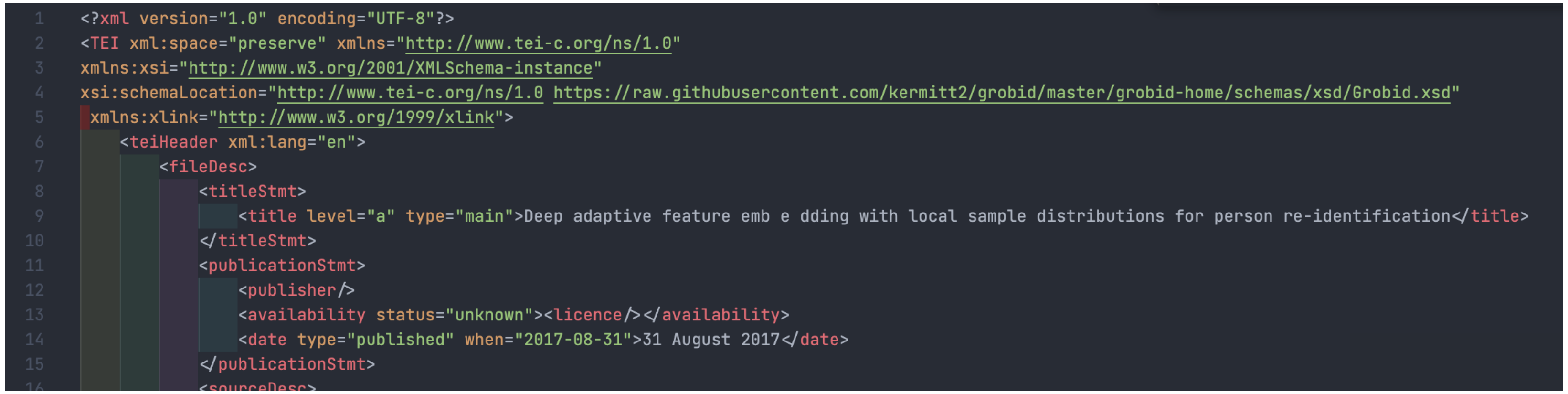
Your submission should look like this
All scores are normalized into [0, 1]

PST: Dataset

The full text is given in the Electronic Text Encoding and Interchange (TEI) XML format

- The <text> section, especially the <body>, which contains the main content of the paper. This is where the model will look for references to the paper’s sources.
- The <ref> tags within the <body>, which indicate references to bibliography entries. These tags can help identify the location of source references in the text.
- The <listBibl> section in the <back>, which contains the bibliography entries.
  Each <biblStruct> represents a referenced work and provides details that can be used to match with the source papers.

paper_xml/599c795b601a182cd26343e7.xml



```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xml:space="preserve" xmlns="http://www.tei-c.org/ns/1.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.tei-c.org/ns/1.0 https://raw.githubusercontent.com/kermitt2/grobid/master/grobid-home/schemas/xsd/Grobid.xsd"
xmlns:xlink="http://www.w3.org/1999/xlink">
    <teiHeader xml:lang="en">
        <fileDesc>
            <titleStmt>
                <title level="a" type="main">Deep adaptive feature emb e dding with local sample distributions for person re-identification</title>
            </titleStmt>
            <publicationStmt>
                <publisher/>
                <availability status="unknown"><licence/></availability>
                <date type="published" when="2017-08-31">31 August 2017</date>
            </publicationStmt>
            <sourceDesc>
```

PST: Evaluation

- For each paper in the test set, AP will be calculated first

$$\mathrm{AP}(V_q) = \frac{1}{R_q} \sum_{k=1}^{M} P_q(k) 1_k$$

  - Rq is the number of positive examples (important references)
  - Pq(k) is the precision at cut-off k in the ranked list
  - 1k is the actual annotation result, with values of 0 or 1
    - 0 is a negative example (non-important reference)
    - 1 is a positive example (important reference)
  - M represents the number of references for the paper
- MAP is then calculated by taking the mean of AP for all papers

PST: Baselines

- Statistical approaches
  - Rule: employs regular expressions to extract references appearing near signal words like “motivated by” or “inspired by”
  - Random forest (RF): extracts statistical features about citations, citing positions, text similarity, etc., and uses RF to predict reference importance
- Graph-based approaches
  - LINE and NetSMF: train paper embeddings in citation networks, calculate cosine similarity between paper and reference embeddings to measure reference importance
- Pre-training methods
  - Extract contextual text where each reference appears in full texts
  - Encode text with pre-trained models (BERT, SciBERT, Galactica-standard, GLM)
  - Fine-tune using reference annotation results in training set
  - Can also adopt closed-source models: GPT-3.5, GPT-4, Claude-instant