WUT21 at the NTCIR-16 Data Search 2 Task

Lin Li
Wuhan University of Technology
cathylilin@whut.edu.cn

Xinyu Chen
Wuhan University of Technology
268595@whut.edu.cn

Sijie Long
Wuhan University of Technology
sijie.long@whut.edu.cn

Introduction

- A common strategy for text-based information retrieval is to use a ranking function to rank all texts according to search terms and select the top n.
- This report describes and discusses our results using different textual similarities for topics in the IR subtask to calculate how well topics match documents and returning a sorted list.


```
[
    "id": "0063664a-d0d7-4ce2-9462-0463a89fc274",
    "url": "https://catalog.data.gov/dataset/0063664a-
        d0d7-4ce2-9462-0463a89fc274",
    "attribution": "CRED REA Fish Team Stationary Point
        Count Surveys at Sarigan, Marianas Archipelago,
        2005 (https://catalog.data.gov/dataset/0063664a-
        d0d7-4ce2-9462-0463a89fc274) is licensed under U
        .S. Government Work (http://www.usa.gov/
        publicdomain/label/1.0/)"
    "title": "CRED REA Fish Team Stationary Point Count
        Surveys at Sarigan, Marianas Archipelago, 2005",
    "description": "Stationary Point Counts at 4 stations
         at each survey site were surveyed as part of
        Rapid Ecological Assessments (REA) conducted at
        3 sites around Sarigan in the Marianas
        Archipelago (MA) during 3 September - 1 October
        2005 in the NOAA Oscar Elton Sette (OES 0511)
        Reef Assessment and Monitoring Program (RAMP)
        Cruise. Raw survey data included species level
        abundance estimates.",
    "data": [
        {
            "data_format": "excel",
            "data_organization": "National Oceanic and
                Atmospheric Administration, Department
                of Commerce",
            "data_url": "https://data.nodc.noaa.gov/coris
                /data/NOAA/nmfs/pifsc/cred/REAFish/
                CNMI_2005/CRED_REA_FISH_SAIPAN_2005.xls"
                ,
            "data_filename": "CRED_REA_FISH_SAIPAN_2005.
                xls"
        },
        {
            "data_format": "csv",
            ...
        }
    ],
    "data_fields": {
        "Resource Type": "Dataset",
        "Metadata Date": "June 20, 2018",
        "Metadata Created Date": "February 7, 2018",
        "Metadata Updated Date": "February 27, 2019",
        ...
        "metadata_sources": [
            "https://catalog.data.gov/harvest/object/
                fc5a39b7-4c9f-49b8-af95-2812d9b3264c"
        ]
    }
}
```


Data sample

WUT21 at the NTCIR-16 Data Search 2 Task

Lin Li
Wuhan University of Technology
cathylilin@whut.edu.cn

Xinyu Chen
Wuhan University of Technology
268595@whut.edu.cn

Sijie Long
Wuhan University of Technology
sijie.long@whut.edu.cn

Methods

- LM Jelinek Mercer Similarity algorithm:

  Under the query-likelihood approach, language models for IR try to estimate for each document the probability that the query Q was generated by the underlying language model. If it is assumed that terms occur independently, then the probability becomes the product of the individual query terms given the document mode.

WUT21 at the NTCIR-16 Data Search 2 Task

Lin Li
Wuhan University of Technology
cathylilin@whut.edu.cn

Xinyu Chen
Wuhan University of Technology
268595@whut.edu.cn

Sijie Long
Wuhan University of Technology
sijie.long@whut.edu.cn

Experiments

- Statistical analysing

Table 2: The number of topics with L2 labels.

	Number of topics
L2 label	69
sum	192

Table 1: The number of L2 labels in the training set.

	training set
L2 label	141
sum	10536

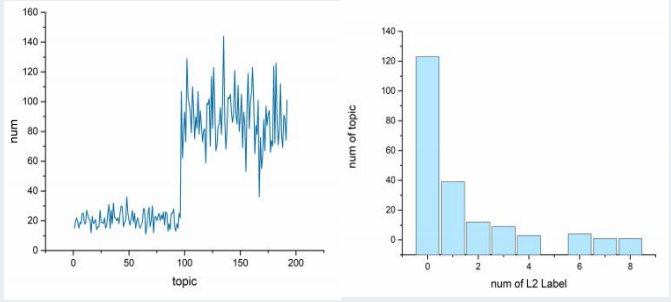

Table 1 records the number of L2 labels in the training set and the size of the training set.
Table 2 records the number of topics with L2 labels and the total number of topics in the training set.
Figure 1 counts the number of document entries in the training set corresponding to each topic.
Figure 2 counts the number distribution of L2 label in topics.

WUT21 at the NTCIR-16 Data Search 2 Task

Lin Li
Wuhan University of Technology
cathylilin@whut.edu.cn

Xinyu Chen
Wuhan University of Technology
268595@whut.edu.cn

Sijie Long
Wuhan University of Technology
sijie.long@whut.edu.cn

Experiments

- Search process

WUT21 at the NTCIR-16 Data Search 2 Task

Lin Li
Wuhan University of Technology
cathylilin@whut.edu.cn

Xinyu Chen
Wuhan University of Technology
268595@whut.edu.cn

Sijie Long
Wuhan University of Technology
sijie.long@whut.edu.cn

Conclusion

- In the final performance results, the effect presented by our team is moderate in the overall performance