STIS POLYTECHNIC OF STATISTICS

For Better Official Statistics

STIS at the NTCIR16-DataSearch 2 Task: Ad-hoc Data Retrieval Ranking with Pretrained Representative Words Prediction

Lya Hulliyyatus Suadaa, Lutfi Rahmatuti Maghfiroh, Muhammad Luqman and Isfan Nur Fauzi

Jakarta, June 17th 2022

Introduction

NTCIR-16 Data Search is a shared task on ad-hoc retrieval for governmental statistical data (more information: https://ntcir.datasearch.jp/)

Information Retrieval Subtask:

- Given queries and data collections, we are expected to generate a ranked list of statistical data for each query.
- Japanese & **English**

Task Introduction: Information Retrieval

Users

Query

causes of death us 1996-2016

Data Search Engine

Ranked List of Retrieved Data

Deaths in 122 U.S. cities - 1962-2016. 122 Cities Mortality Reporting System
The data include mortality and population counts for all U.S. counties for the years 1968 to 2016…

Associations between environmental quality and mortality in the contiguous United States 2000-2005
Age-adjusted mortality rates for the contiguous United States in 2000–2005 were obtained from the Wide-ranging Online…

Global COVID-19 Viewer: Population Estimates by Age Groups and Sex
A new global mapping tool shows the number of people affected by the by the COVID-19 pandemic down to the U.S…

CDC WONDER: Mortality - Underlying Cause of Death
The CDC WONDER Mortality - Underlying Cause of Death online database is a county-level national mortality and...

1 2 Next

DATA.GOV

Deaths in 122 U.S. cities - 1962-2016. 122 Cities Mortality Reporting System

The data include mortality and population counts for all U.S. counties for the years 1968 to 2016. Counts and rates of death can be obtained by underlying cause of death…

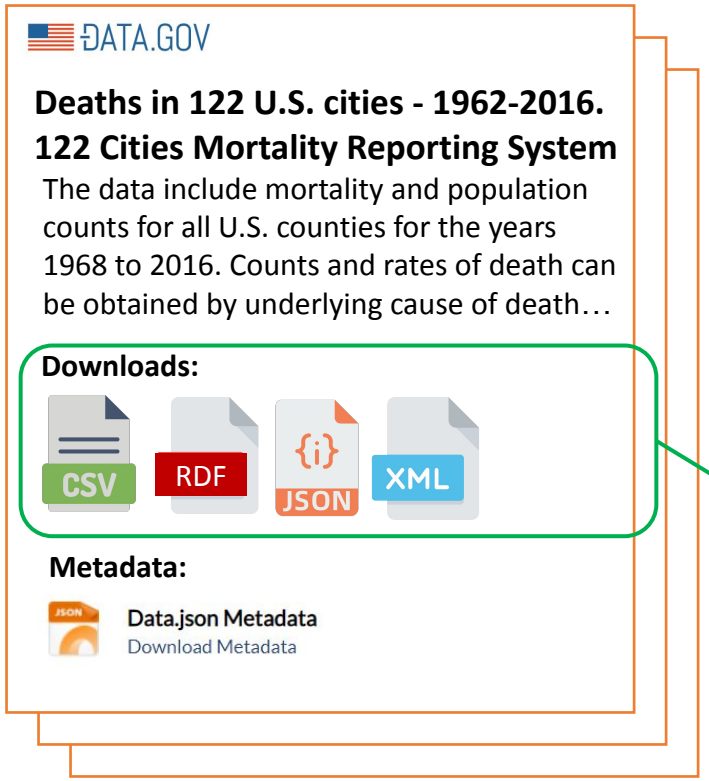
Downloads:

CSV RDF JSON XML

Metadata:

Data.json Metadata
Download Metadata

1. **Dataset collections**: datasets published by the US government (https://data.gov)

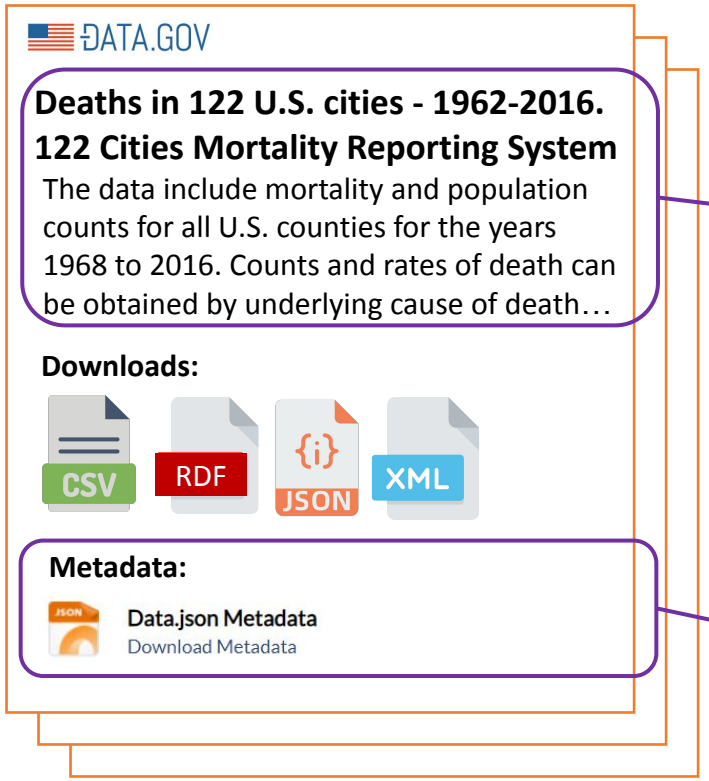


Data

{
"id":"3548b49b-0173-4724-868d-722fe61d39c6",
"title":"Deaths in 122 U.S. cities - 1962-2016. 122 Cities Mortality Reporting System",
"description":"The data include mortality and population counts for all U.S. counties for the years 1968 to 2016...
"data":
[{"data_format":"csv","data_organization":"U.S. Department of Health & Human Services","data_url":"…csv"},
{"data_format":"rdf","data_organization":"U.S. Department of Health & Human Services","data_url":"…rdf","data_filename":"...rdf+xml"},
{"data_format":"json","data_organization":"U.S. Department of Health & Human Services","data_url":"…json","data_filename":"...json"},
{"data_format":"xml","data_organization":"U.S. Department of Health & Human Services","data_url":"...xml","data_filename":"….xml"}],
…
}

Datasets

1. **Dataset collections**: datasets published by the US government (https://data.gov)

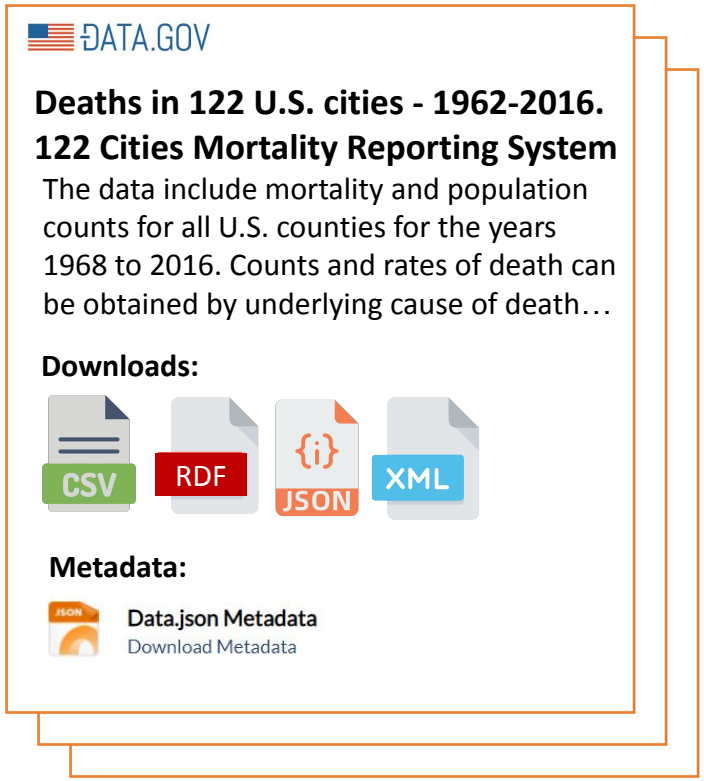


Metadata


```
{
"id":"3548b49b-0173-4724-868d-722fe61d39c6",
"title":"Deaths in 122 U.S. cities - 1962-2016. 122 Cities Mortality Reporting System",
"description":"The data include mortality and population counts for all U.S. counties for the years 1968 to 2016...
…
"data_fields":{
    "Resource Type":"Dataset",
    "Metadata Created Date":"May 8, 2016",
    "Metadata Updated Date":"September 2, 2019",
    …,
    "tags":["122-cities","2016","death","influenza","mortality",
    "pneumonia"],
    "metadata_sources":["https:\/\/catalog.data.gov\/harvest
    \/object\/1145c63b-6d03-4f3c-98a4-d9f16c18f103"]},
    …}
}
```

Datasets

1. **Dataset collections**: datasets published by the US government (https://data.gov)



2. **Queries:**

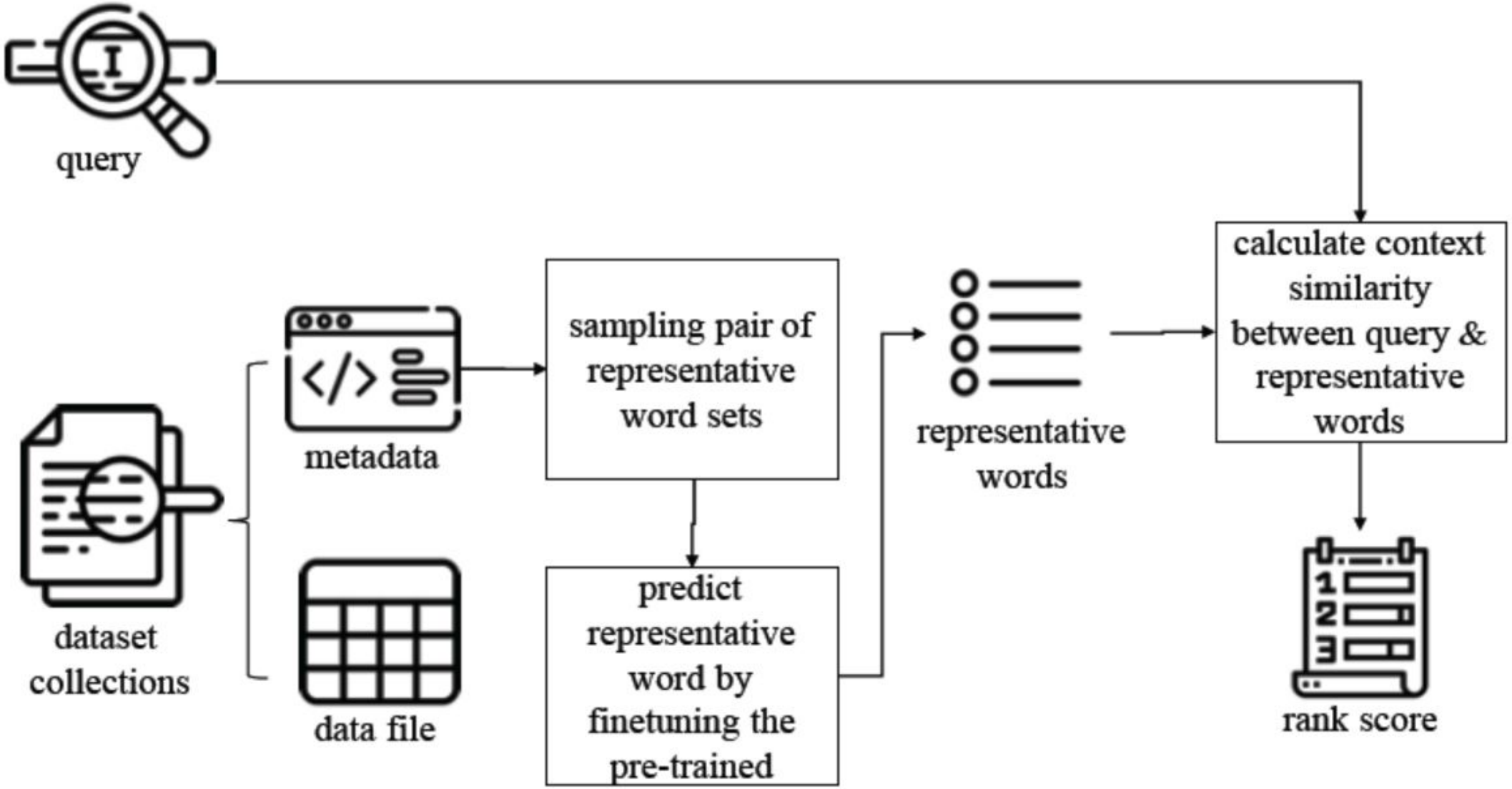
causes of death us 1996-2016

Pretrain Representative Words Prediction

- Implemented Pre-training with Representative wOrds Prediction (PROP) for ad-hoc retrieval (Xinyun Ma, et al., WSDM 2021) in constructing the representative words prediction for each document.

- Representative words: inspired by a query likelihood model that ranks the documents based on the relationship between a query and the document contents.

Pretrain Representative Words Prediction



Representative Word Sets Sampling

- sample a pair of representative word sets from vocabulary $V = \{w_i\}_n^1$ based on document language model following dirichlet distribution with probability $P(w_i|D)$

$$\text{if } QL_u > QL_v \text{ then } w_{pos} = u \text{ and } w_{neg} = v$$
$$\text{else } w_{pos} = v \text{ and } w_{neg} = u$$

Representative Words Prediction

- finetune pre-trained Transformer BERT for representative words prediction task

- hidden state:

$$h_{[CLS]} = \text{Transformer}([CLS] + w_{rep} + [SEP] + D + [SEP])$$

where $w_{rep} = \{w_{pos}, w_{neg}\}$

- probability of word represent to the the document:

$$P(w_{rep}|D) = \text{MLP}(h_{[CLS]})$$

- loss function:

$$\mathcal{L} = \max(0, 1 - P(w_{pos}|D) + P(w_{neg}|D))$$

Rank Score Calculation

- $\{w_{prop}\}_1^k$: k list of representative word prediction from the finetuned BERT

- rank score using prop:

$$S_{prop} = \text{avg}(BERT_{Score}(query, w_{prop_i})), i = 1...k$$

1st Model	2nd Model
$S = S_{prop}$	$S = \alpha S_{base} + (1- \alpha)S_{prop}$

Results

Model	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q-mea
BM25	**0.191**	**0.188**	**0.211**	**0.199**	**0.222**	**0.233**	**0.248**
Ranking by PROP and BertScore	0.172	0.175	0.201	0.191	0.188	0.201	0.218
BM25 + Reranking by PROP and BertScore	0.163	0.173	0.202	0.192	0.183	0.200	0.221

Table 1: Results from NTCIR-16 Data Search 2. The best score is in bold.

Results

Model	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q-mea
BM25	**0.191**	**0.188**	**0.211**	**0.199**	**0.222**	**0.233**	**0.248**
Ranking by PROP and BertScore	0.172	0.175	0.201	0.191	0.188	0.201	0.218
BM25 + Reranking by PROP and BertScore	0.163	0.173	0.202	0.192	0.183	0.200	0.221

Table 1: Results from NTCIR-16 Data Search 2. The best score is in bold.

Results

Query	Representative Word	$Rank_{BM25}$	$Rank_{PROP}$
causes of death us 1999-2016	x county cause numerous rate, new census, update scientifically deaths deaths, death estimating internationally, direction mortality death poisoning low ages rates, base longer meets deaths, drug death computer selected, estimated coded, affect states death poisoning, pending ratings x poisoned	6	1
	deaths finalization deaths, provisional death provisional comparisons, classifications updated, deaths reported death deaths causative categories, specifically cause, drug, counting drugs provisional, drug delay provisional, drug x updates vital, drug numbered pending drug	9	2
	rated rate x nchs differing, rated, rating census causes estimate number baseline, poisoned base acquisition, references death, poisoned drug includes, update cdcs adjusted, www published demographic, x rating ageing, death defined wonder	1	8

Table 2: Samples of representative words prediction for each query and their ranks.

Results

Query	Representative Word	$Rank_{BM25}$	$Rank_{PROP}$
annual turnover care workers	disabled, incoming, benefits statistics, insurance social, socially calculations workers, securing three, three, receiving edition calculate, disability series people disability tables workers, peoples social statistical	56	1
	low researchers, teams v u shorebird shores forage, forest v tennessee foraging group, radio forest primary radio migrate, ponds estimating migratory shorebird, opening vulnerable estimated network rates, calidris, western flight shorebird reservation created migration, establishing rate, build regional valleys	1	35

Table 2: Samples of representative words prediction for each query and their ranks.

Results

Query: annual turnover care workers

https://catalog.data.gov/dataset/annual-statistical-report-on-the-social-security-disability-insurance-program-2005

DATA.GOV

Annual Statistical Report on the Social Security Disability Insurance Program, 2005

This annual report provides program and demographic information on the people who receive Social Security Disability Insurance Program benefits. This edition presents a series of detailed tables on the three categories of beneficiaries: **disabled workers, disabled widowers, and disabled adult children**. Numbers presented in these tables may differ slightly from other published statistics because all tables, except those using data from the Survey of Income and Program Participation, are based on 100 percent data files. Report for 2005.

Downloads:

XLS PDF

Representative Words:
disabled, incoming, benefits statistics, insurance social, socially calculations workers, securing three, three, receiving edition calculate, disability series people disability tables workers, peoples social statistical

Results

Query: annual turnover care workers

https://catalog.data.gov/dataset/annual-statistical-report-on-the-social-security-disability-insurance-program-2005

DATA.GOV

Annual Statistical Report on the Social Security Disability Insurance Program, 2005

This annual report provides program and demographic information on the people who receive Social Security Disability Insurance Program benefits. This edition presents a series of detailed tables on the three categories of beneficiaries: **disabled workers, disabled widowers, and disabled adult children**. Numbers presented in these tables may differ slightly from other published statistics because all tables, except those using data from the Survey of Income and Program Participation, are based on 100 percent data files. Report for 2005.

Downloads:

XLS PDF

Representative Words:
disabled, incoming, benefits statistics, insurance social, socially calculations workers, securing three, three, receiving edition calculate, disability series people disability tables workers, peoples social statistical

Results

Query: annual turnover care workers

https://catalog.data.gov/dataset/annual-statistical-report-on-the-social-security-disability-insurance-program-2005

DATA.GOV

Annual Statistical Report on the Social Security Disability Insurance Program, 2005

This annual report provides pro sability
Insurance Program benefits. Th ries:
disabled workers, disabled widc slightly
from other published statistics l m
Participation, are based on 100

Downloads:

XLS PDF

All Disabled Beneficiaries

Table 4.
Number and average monthly benefit, by sex and age, December 2005

Age	Total		Workers		Widow(er)s		Adult children	
	Number	Average monthly benefit (dollars)	Number	Average monthly benefit (dollars)	Number	Average monthly benefit (dollars)	Number	Average monthly benefit (dollars)
	All disabled beneficiaries							
Total	7.503.405	895,60	6.519.001	938,00	215.866	609,30	768.538	616,30
Under 25	130.921	513,30	54.623	504,30	...	...	76.298	519,80
25–29	210.312	599,80	146.385	614,60	...	...	63.927	565,90
30–34	291.762	670,70	221.138	694,40	...	...	70.624	596,40
35–39	455.489	735,50	368.579	762,20	...	...	86.910	622,20
40–44	732.382	791,90	622.211	819,20			110.171	637,60

Table 1 | Table 2 | Table 3 | **Table 4** | Table 5 | Table 6 | Table 7 | Table 8 | Table ...

Representative Words:
disabled, incoming, benefits statistics, insurance social, socially calculations workers, securing three, three, receiving edition calculate, disability series people disability tables workers, peoples social statistical

Results

Query: annual turnover care workers

http://datadiscoverystudio.org/geoportal/rest/metadata/item/ce61995d7d6b4070b1bdcbe68d710276/html

DATA.GOV

Turnover Rates of Fall Migrating Pectoral Sandpipers Through the Lower Mississippi Alluvial Valley

The Mississippi Alluvial Valley (MA V) is the historic alluvial floodplain of the Lower Mississippi River. Most of the MAV is … The primary objective of this study is to estimate the turnover rates of fall migrating shorebirds in the MA V using 2 target species, the pectoral sandpiper (Calidris metanotos, PESA) and the least sandpiper (Calidris minulilla, LESA). We will estimate turnover rates from PESAs using radio telemetry data from 2001 and 2002 and from LESAs using capture-resight data from 2002.

Downloads:

PDF

Representative Words:
low researchers, teams v u shorebird shores forage, forest v tennessee foraging group, radio forest primary radio migrate, ponds estimating migratory shorebird, opening vulnerable estimated network rates, calidris, western flight shorebird reservation created migration, establishing rate, build regional valleys

Results

Query: annual turnover care workers

http://datadiscoverystudio.org/geoportal/rest/metadata/item/ce61995d7d6b4070b1bdcbe68d710276/html

DATA.GOV

Turnover Rates of Fall Migrating Pectoral Sandpipers Through the Lower Mississippi Alluvial Valley

The Mississippi Alluvial Valley (MA V) is the historic alluvial floodplain of the Lower Mississippi River. Most of the MAV is … The primary objective of this study is to estimate the turnover rates of fall migrating shorebirds in the MA V using 2 target species, the pectoral sandpiper (Calidris metanotos, PESA) and the least sandpiper (Calidris minulilla, LESA). We will estimate turnover rates from PESAs using radio telemetry data from 2001 and 2002 and from LESAs using capture-resight data from 2002.

Downloads:

PDF

Representative Words:
low researchers, teams v u shorebird shores forage, forest v tennessee foraging group, radio forest primary radio migrate, ponds estimating migratory shorebird, opening vulnerable estimated network rates, calidris, western flight shorebird reservation created migration, establishing rate, build regional valleys

Results

Query: annual turnover care workers

http://datadiscoverystudio.org/geoportal/rest/metadata/item/ce61995d7d6b4070b1bdcbe68d710276/html

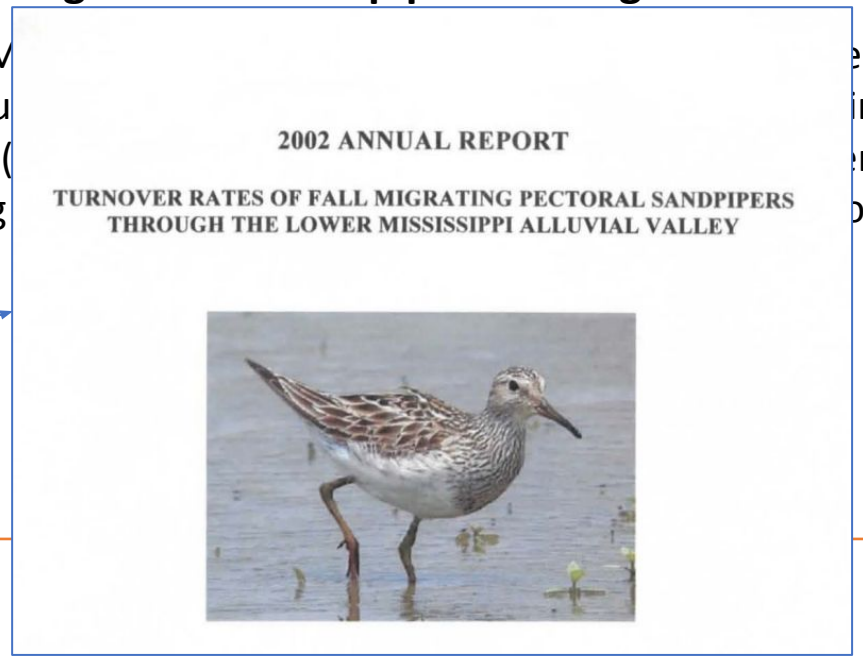
DATA.GOV

Turnover Rates of Fall Migrating Pectoral Sandpipers Through the Lower Mississippi Alluvial Valley

The Mississippi Alluvial Valley (M … er Mississippi River. Most of the MAV is … The primary objective of this stu … ing shorebirds in the MA V using 2 target species, the pectoral sandpiper (… r (Calidris minulilla, LESA). We will estimate turnover rates from PESAs using … om LESAs using capture-resight data from 2002.

Downloads:

PDF


2002 ANNUAL REPORT

TURNOVER RATES OF FALL MIGRATING PECTORAL SANDPIPERS THROUGH THE LOWER MISSISSIPPI ALLUVIAL VALLEY

Representative Words:
low researchers, teams v u shorebird shores forage, forest v tennessee foraging group, radio forest primary radio migrate, ponds estimating migratory shorebird, opening vulnerable estimated network rates, calidris, western flight shorebird reservation created migration, establishing rate, build regional valleys

Conclusion

Addressing ad-hoc retrieval approach for governmental statistical data:

- proposed using a pre-trained model to capture representative words prediction for each document then calculate the similarity between the query and the representative words as a rank score.
- proposed combined the representative similarity score to re-rank candidate documents of BM25 model for each query.

THANK YOU