Automated Report Generation + Natural Language Processing

US Census Bureau | International Programs Center

Kevin Gui, Data Fellow | UC Davis, Data Science and Economics

Keywords:

R Shiny, Automation, Report Generation, NLP, Classification

Summary:

To allow low and middle-income countries to disseminate their census data better, Kevin worked in the International Programs Center and developed an automated report generation feature in the **Open Source Dissemination System (OSDS).** Utilizing R and R Shiny, he created a standalone R package {popreports} to generate the reports, integrating this functionality into OSDS, and allowing users to generate national, subnational, and batch reports to disseminate data efficiently.

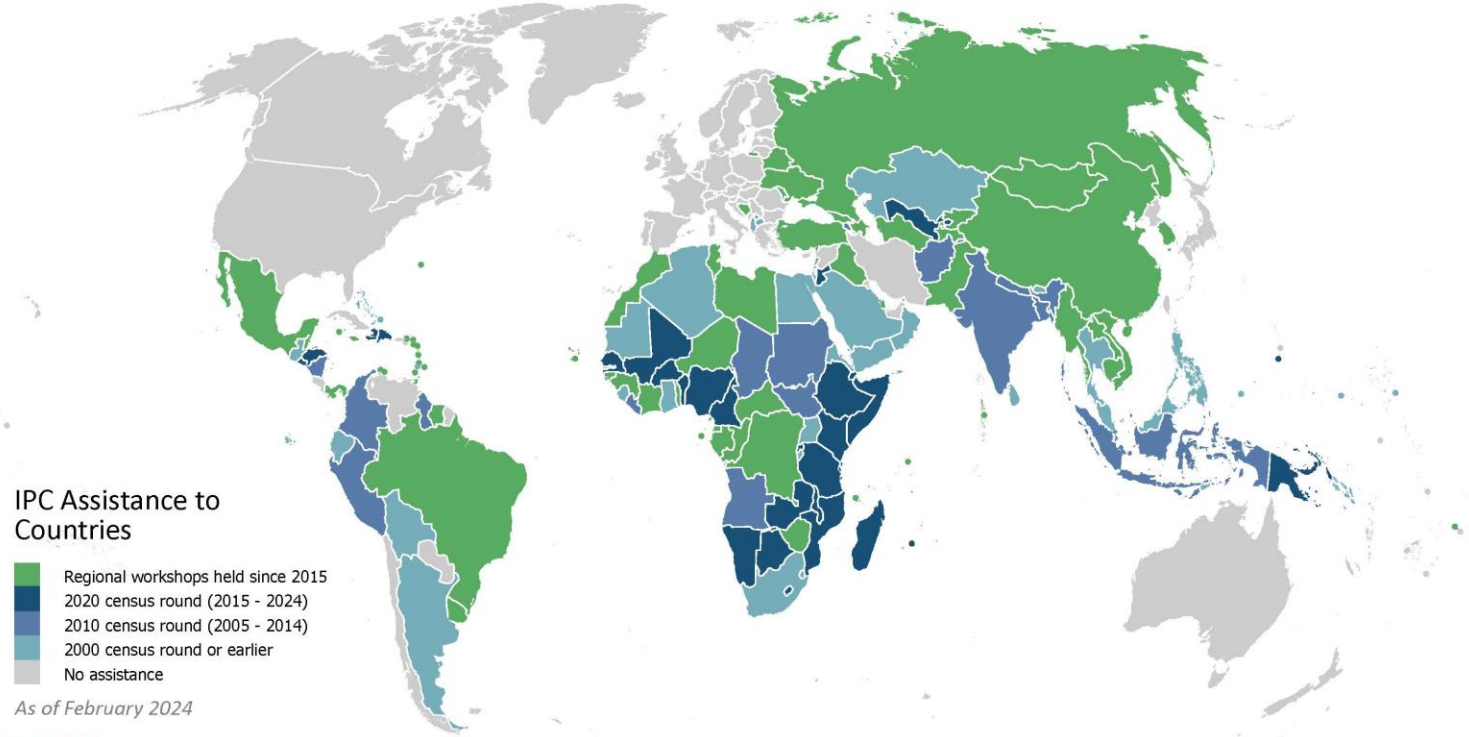
Kevin also worked on the **HIV/AIDS Review & Verification System (HARVEST)**, which is an important dashboard visualizing HIV/AIDS statistics around the world, using **Natural Language Processing** and **Machine Learning** to classify web-scraped documents that contained relevant and non-relevant information. This assists the current humans in separating relevant and non-relevant document, significantly reducing their workload.

Automated Report Generation and Natural Language Processing

Kevin Gui
3rd year Data Science student @ UC Davis

International Programs Center Introduction

The International Programs Center (IPC) advances data-driven decision making through tools, capacity strengthening, and data products for the global statistical community.



United States Census Bureau USAID FROM THE AMERICAN PEOPLE

International Programs Center (census.gov)

Table of Contents

Two Different Projects:

1. Automating Census Reports in OSDS (Open Source Dissemination System)
2. Natural Language Processing and HIV/AIDS Document Classification for HARVEST (HIV/AIDS Review and Verification System)

Note: Both products are created for use by international partners in low- and middle-income countries

1. Automated Report Generation in OSDS

What is OSDS?

United States Census Bureau

The Goal

To build an open-source, sovereign, data dissemination and report-producing platform where ease-of-use is a production requirement.

The Vision

To construct an architecture that accepts and cleans inputs and automatically produces maps, graphs, topic-specific data tables, and **PDF "hot" reports from that material in an interactive web platform**.

Automated Report Generation in OSDS

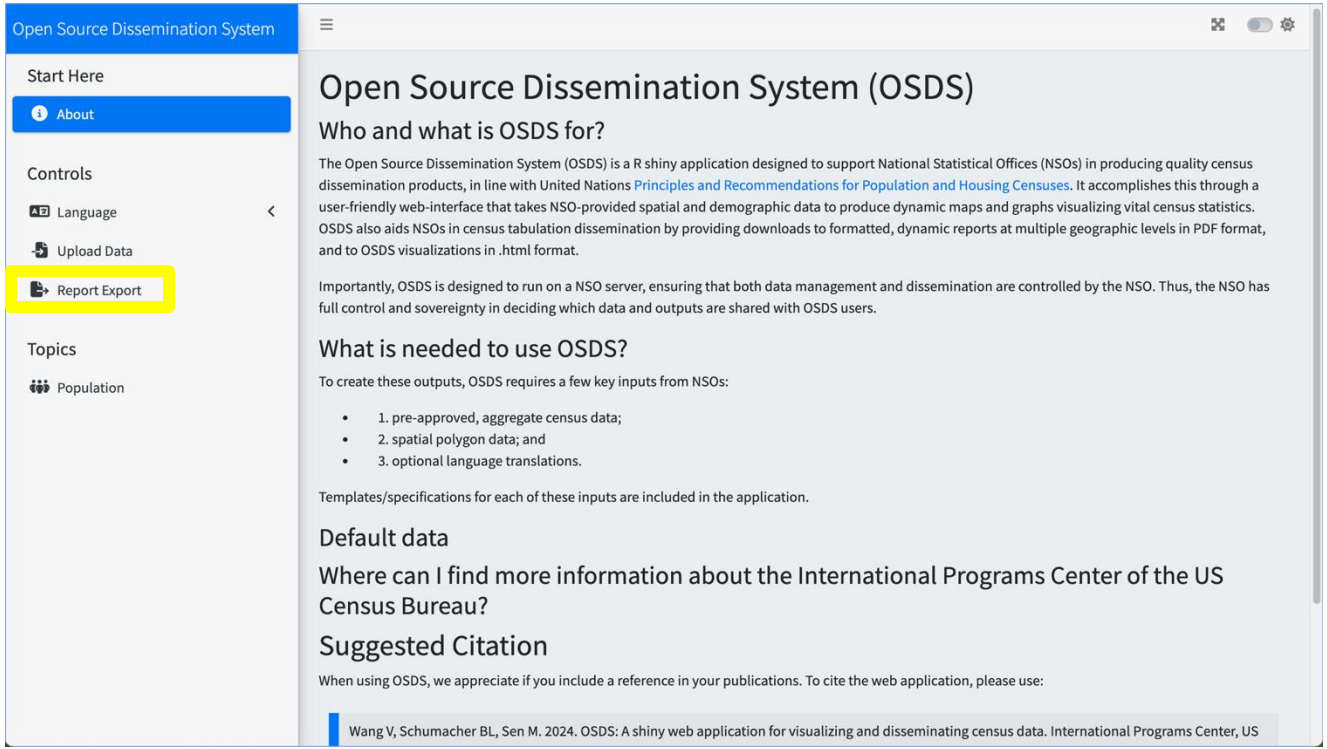
Problem:

- The OSDS application did not allow users to analyze data within the app, with no way of exporting it to a document for dissemination

Fellowship Aims:

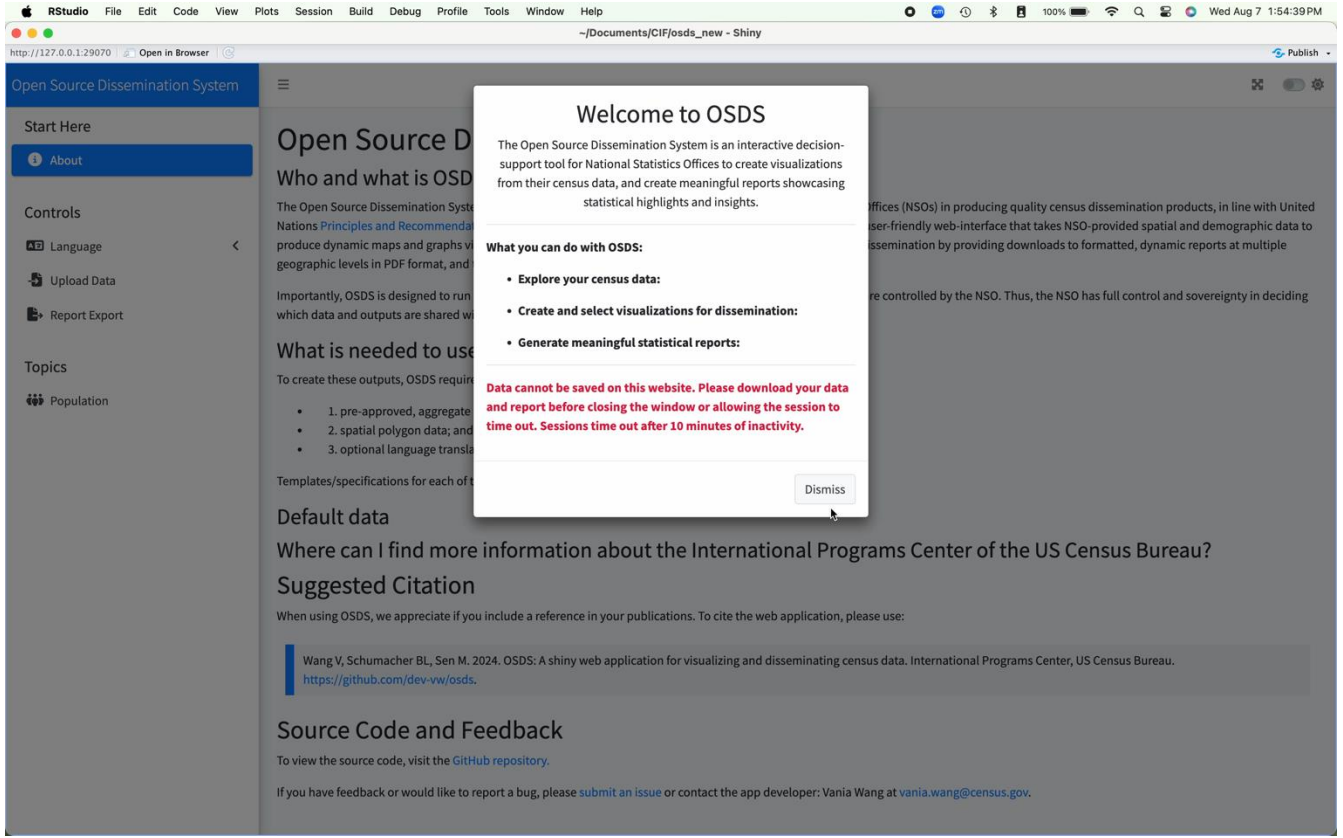
- Develop an export feature to permit simple and quick conversion of complex tabular and geospatial data to downloadable, shareable reports (e.g. PDF, docx).
- Create an R package, {popreports}, which the export feature calls, that works with ingested data in OSDS to create reports on selected census topics.

App Layout



- Before the fellowship, users could only upload and visualize data
- I built out the **report export page**, allowing all important visualized data to be downloaded in a single or many reports

Results/App Demo



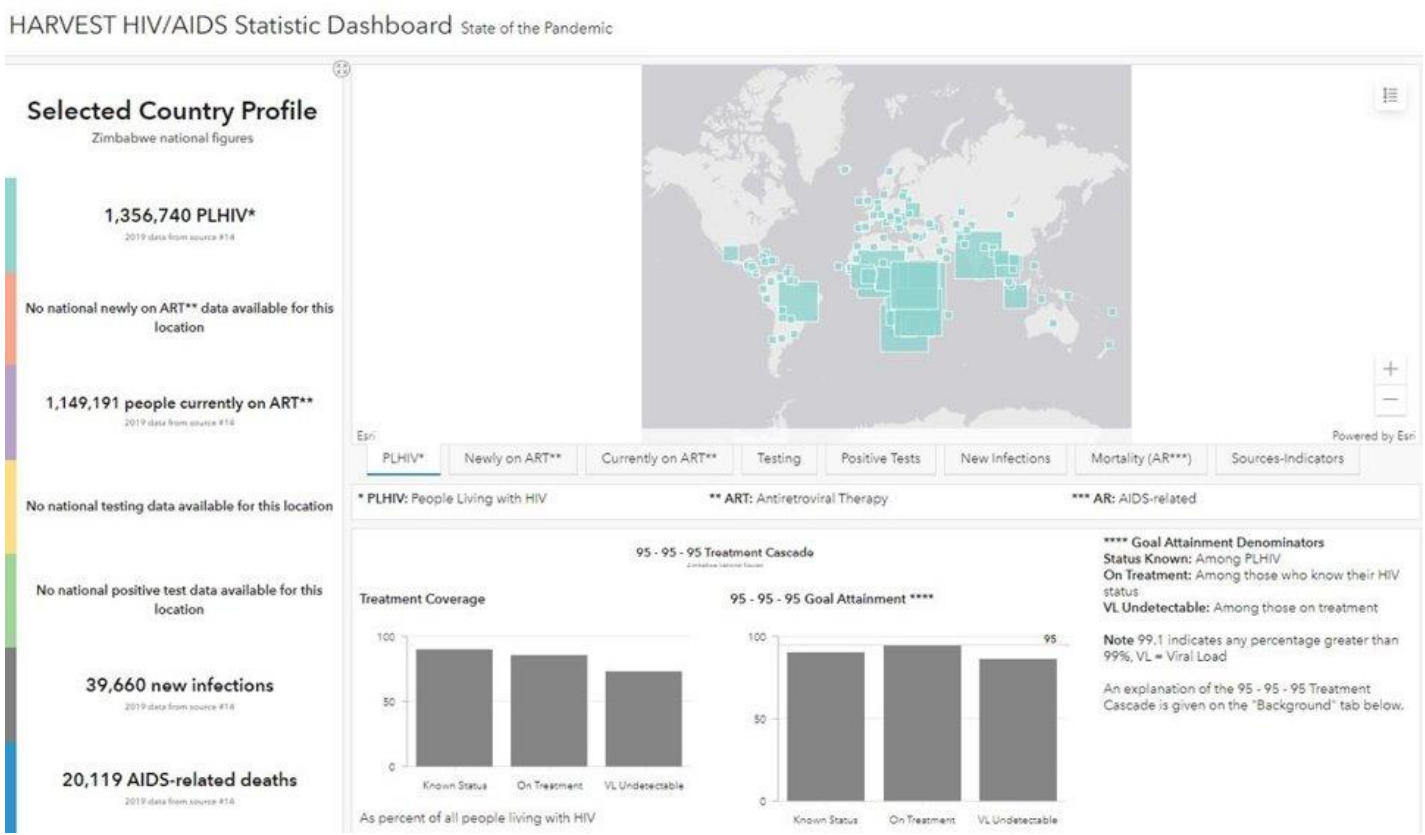
Can be accessed at github.com/dev-vw/osds when it becomes public

Next Steps

- Explore new topics to further disseminate census data
- Allow users to customize the aesthetic and information displayed in their report
- Testing of the application in the low- and middle-income countries

2. Natural Language Processing in HARVEST

HARVEST Dashboard



Natural Language Processing

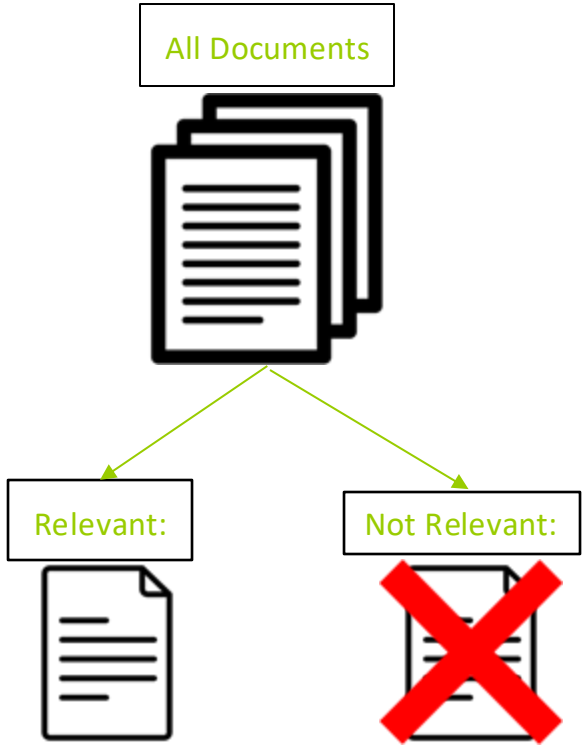
NLP allows computers to understand and speak like humans. Modern NLP applies machine learning tools, techniques, and other algorithms to develop meaningful insights, uncover important patterns, and make informed predictions.

Applications of NLP include:

- sentiment analysis
- text classification
- translation from language to language - text summarization
- and more.

Motivation

- The HARVEST dashboard uses HIV/AIDS statistics scraped from ministries of health webpages. PDFs are scraped and examined by human coders for relevant statistics.
- This project explores whether NLP can help classify these documents into **"relevant"** and **"not relevant"** categories.
- If successful, NLP could help lighten the human coder workload, allowing them to focus on finding statistics in only "relevant" documents.



Pre-processing the Data

United States Census Bureau

Raw text

0	TECHNICAL REPORT\nEffectiveness and cost- \neffectiveness of antenatal screening \nfor HIV, he...
1	SPECIAL REPORT\nImplementing the \nDublin Declaration on \nPartnership to Fight HIV/AIDS \nin E...
2	SPECIAL REPORT\nThematic report: \nSex workers \nMonitoring implementation of the Dublin Decla...
3	TECHNICAL REPORT\nRisk assessment on HIV in Greece\nwww.ecdc.europa.eu \n \n \n \n \n \n ...
4	\nSuggested citation: European Centre for Disease Prevention and Control . Chlamydia . In: ECDC...
	...
67	Suggested citation: HIV testing in Europe and Central Asia . Monitoring implementation of the Du...
68	\n \n \nSuggested citation: European Centre for Disease Prevention and Control. Chlamydia . In:...
69	\n \nEuropean Centre for Disease Prevention and Control, Solna, Sweden \nwww.ecdc.europa.eu \...
70	SPECIAL REPORT\nContinuum of HIV care\n \nMonitoring implementation of the Dublin \nDeclaration ...
71	SPECIAL REPORT\nHIV Continuum of care\nMonitoring implementation of the Dublin \nDeclaration on ...

Pre-processing / Transformations

Remove Punctuation	Remove Unnecessary Symbols	Lowercase all words	Reduce words to their base or root form *(ran, running, runs) -> run*

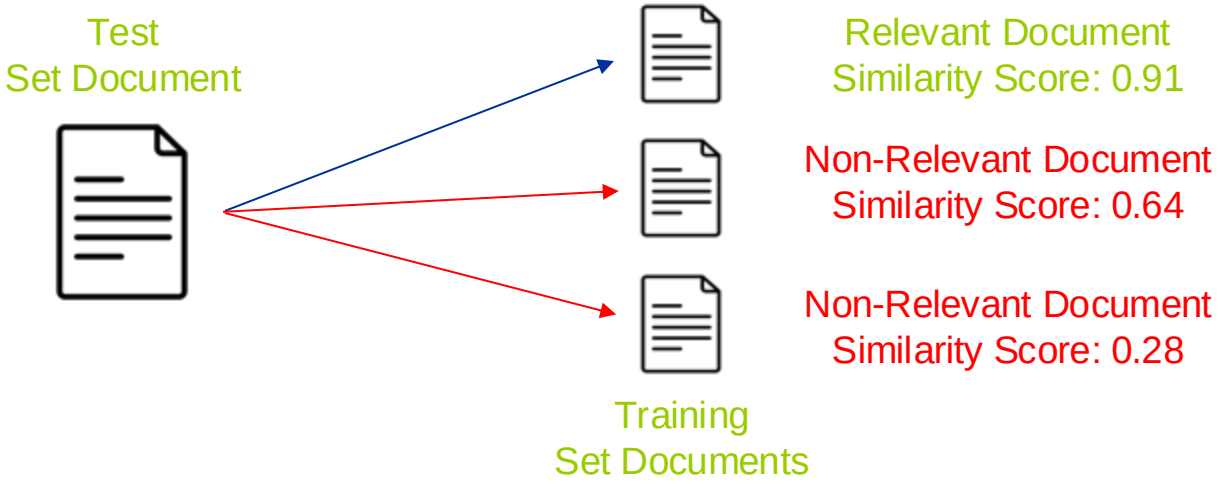
Clean text

0	susceptibility literature susceptibility literature ii institute suggested susceptibility stockh...
1	implementing partnership fight summary implementing partnership fight summary partnership fight ...
2	thematic partnership fight thematic partnership fight ii series thematic submitted partnership f...
3	assessment assessment ii tion van stis like authority particular center organisation assessment ...
4	suggested chlamydia annual stockholm stockholm march fact 216 chlamydia notification 88 notifica...
	...
67	suggested fight stockholm stockholm brief priority largely collected monitor ion liechtenstein o...
68	suggested chlamydia annual epidemiological stockholm stockholm november 201 fact 184 chlamydia n...
69	week march topic overview virus epidemiology sexually transmitted linked middle syndrome summary...
70	partnership fight partnership fight ii agency series thematic submitted partnership fight series...
71	partnership fight partnership fight ii series thematic submitted partnership fight series found ...

Cosine Similarity

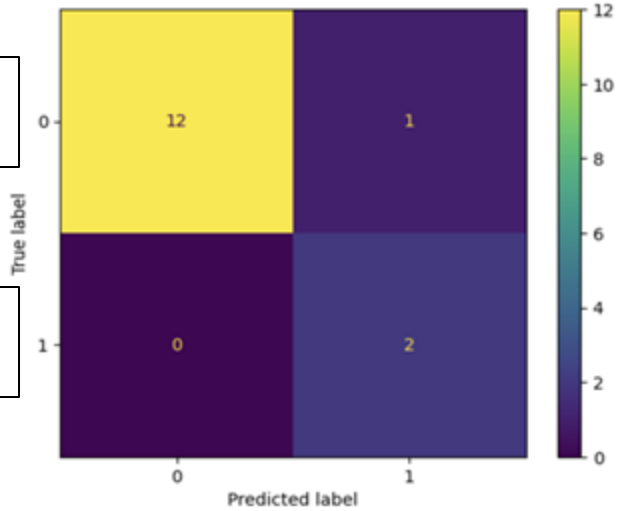
Cosine similarity is a measure to determine how similar two items are to each other by measuring the cosine of the angle their vector representation in a multi-dimensional space. The values of cosine similarity ranges from –1 (completely different) to 1 (exactly the same).

We will use this to classify documents in our test set.



Test Set Document: Classify as a Relevant Document

Cosine Similarity



Actually Not Relevant

Actually Relevant

Predicted Not Relevant

Predicted Relevant

93% Overall Accuracy Score

Next Steps

- Further preprocessing to filter out more unnecessary information
- Adding new features to the model for better performance
- Tweaking parameters to the models
- Trying new machine learning models to see if performance improvements can be made

THANK YOU!

Special thanks to **Britta Schumacher, Vania Wang, and Mitali Sen**!

If you have questions, I can be reached at **kevinguimail@gmail.com**.