LAGOS WiMLDS

WOMEN IN MACHINE LEARNING AND DATASCIENCE.

@wimlds_lagos

Lagos Women In Machine Learning And Data Science Virtual Event 2020

Theme:Linear Modeling And Data Visualization




@wimlds_lagos

Beyond Word Clouds,How to do more with Text Visualization using Scattertext

by
Wuraola Oyewusi
Research and Innovation Lead, Data Science Nigeria, July 2020

@wimlds_lagos

What is Data Visualization ?

Data Visualization is one of the Key Components of Exploratory Data Analysis

@wimlds_lagos

Different Methods of Text Visualization

- Word Highlight
- Word Tree
- TextArc
- Arc Diagrams
- Literature Fingerprinting
- Self Organizing Maps
- Themescape
- Document Cards
- Tile Bars
- Stream Graph
- Tag Clouds or Word Clouds

http://jcsites.juniata.edu/faculty/rhodes/ida/textDocViz.htm

@wimlds_lagos

What is a Tag or Word Cloud?

A tag cloud (word cloud or wordle or weighted list in visual design) is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text.

Tags are usually single words, and the importance of each tag is shown with font size or color.

This format is useful for quickly perceiving the most prominent terms to determine its relative prominence. Bigger term means greater weight

https://en.wikipedia.org/wiki/Tag_cloud

@wimlds_lagos

What Is Scattertext?

Scattertext is an open source tool for visualizing linguistic variation between document categories in a language-independent way. Each axis corresponds to the rank-frequency a term occurs in a category of documents.

Through a tie-breaking strategy, the tool is able to display thousands of visible term-representing points and find space to legibly label hundreds of them. Scattertext also lends itself to a query-based visualization of how the use of terms with similar embeddings differs between document categories, as well as a visualization for comparing the importance scores of bag-of-words features to univariate metrics.

A tool for finding distinguishing terms in small-to-medium-sized corpora, and presenting them in a sexy, interactive scatter plot with non-overlapping term labels. Exploratory data analysis just got more fun.

https://arxiv.org/abs/1703.00565
https://spacy.io/universe/project/scattertext

@wimlds_lagos

Immunization News Dataset Description

This Dataset consists of 2,728 instances of News Headlines, Teaser,Source, Data of News Articles about Immunization.

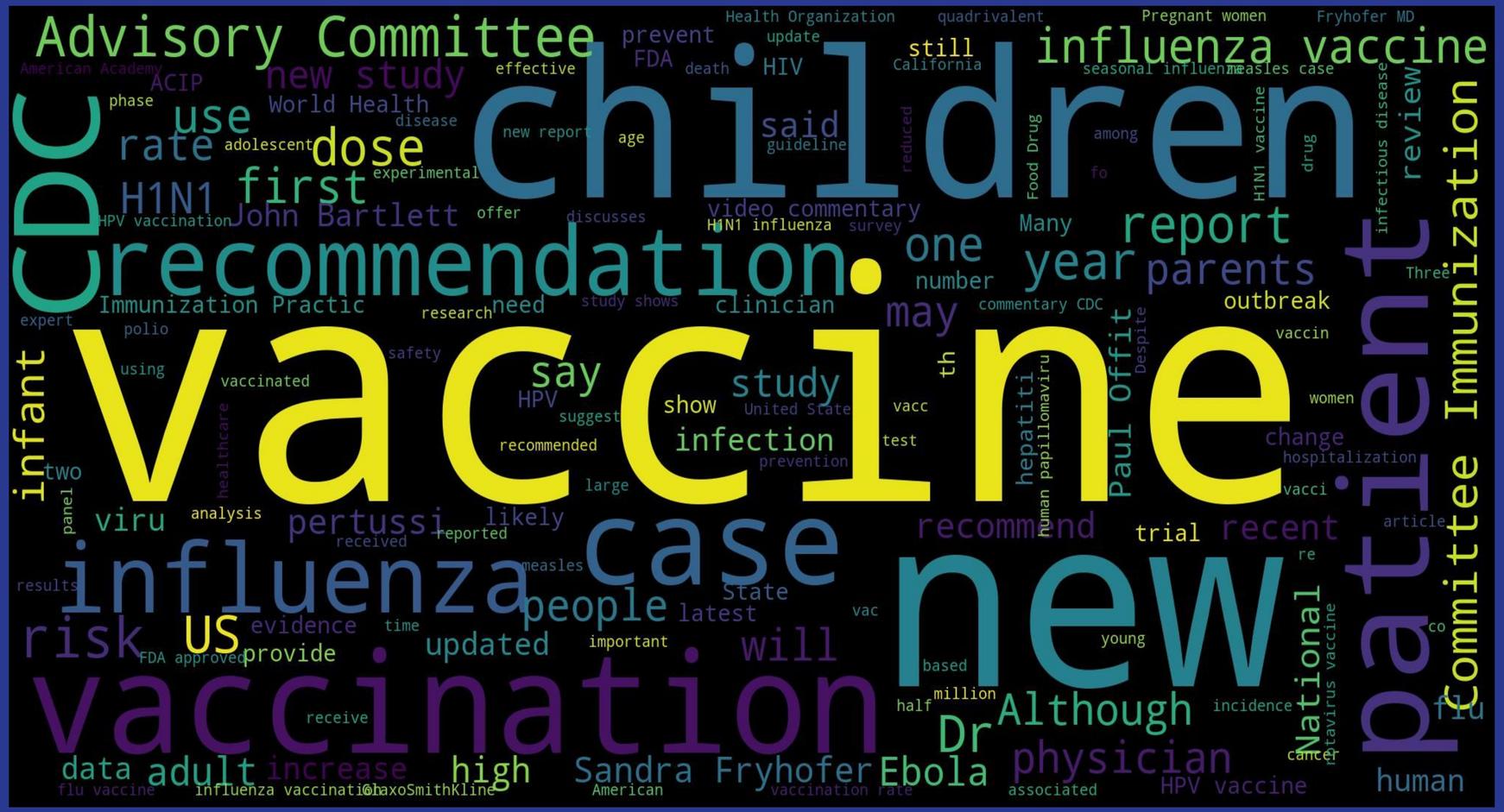
There 1858 instances labelled about Adult Immunization and 881 instances of Children or Peadiatric Immunization

This Colaboratory Notebook shows the full process of Data Preprocessing in python

This Colaboratory Notebook shows the process of Word Cloud and Scattertext Creation using python

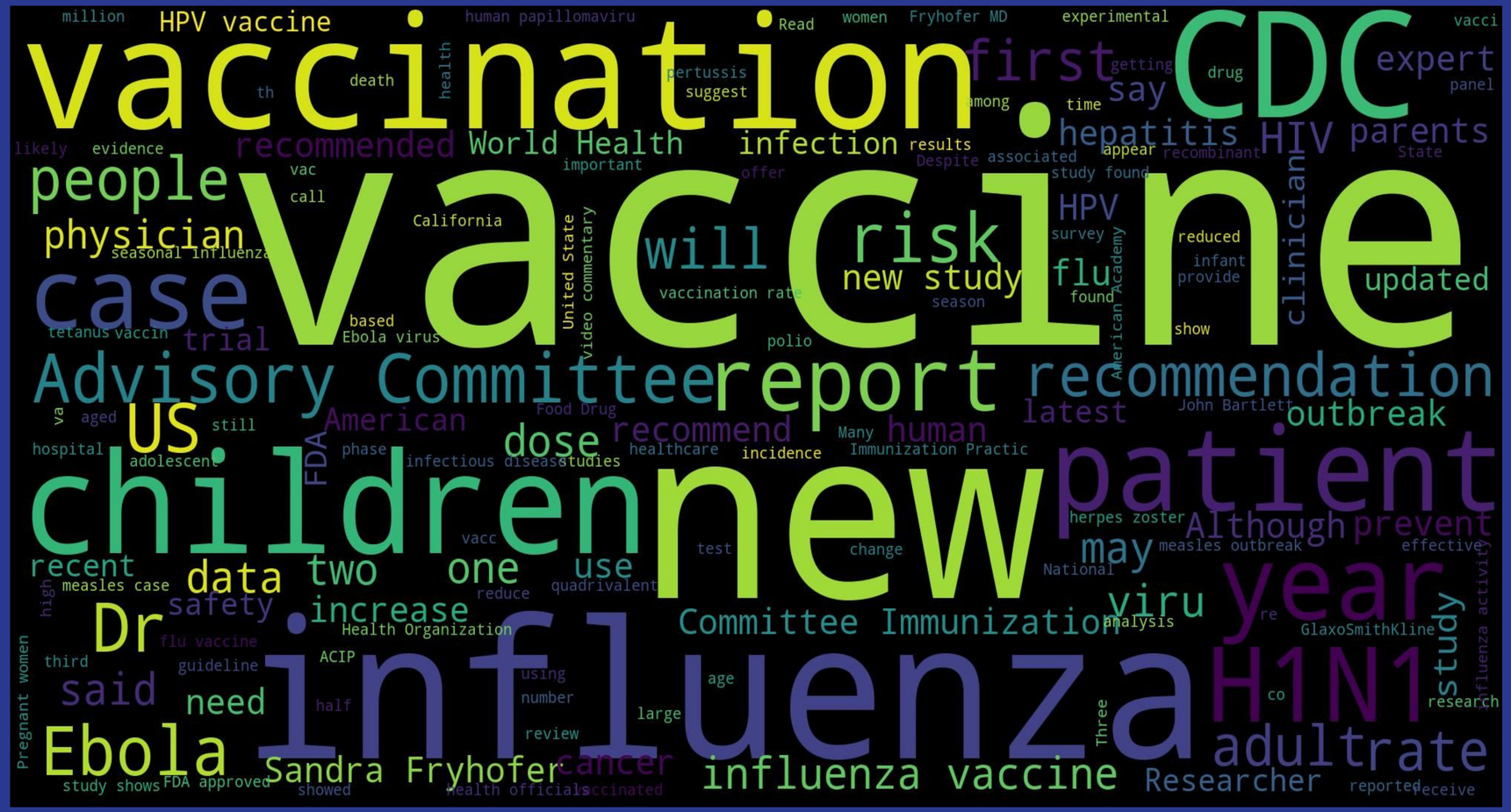
@wimlds_lagos

View of All Immunization Text Word Cloud



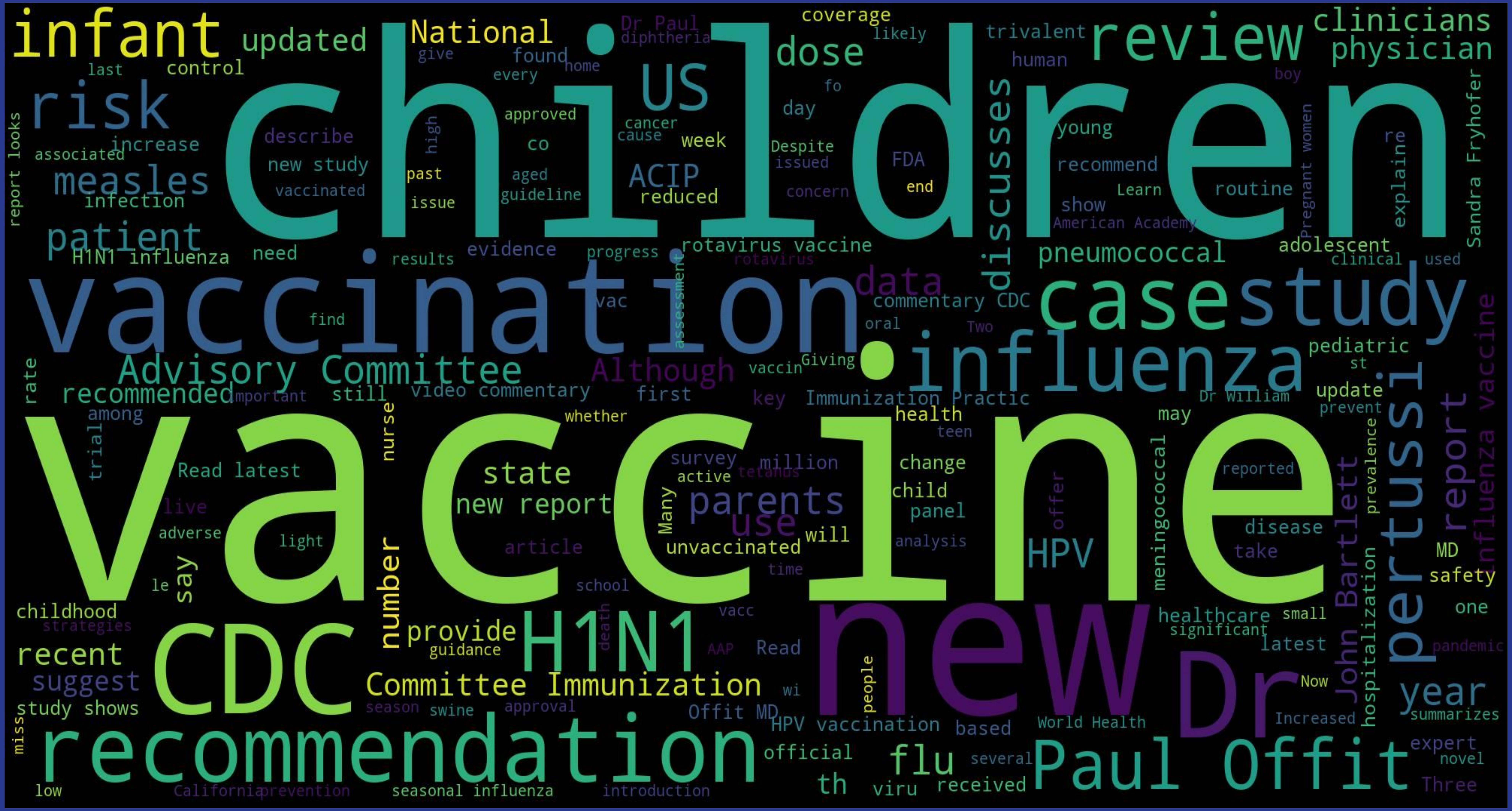
@wimlds_lagos

View of Adult Immunization Text Word Cloud



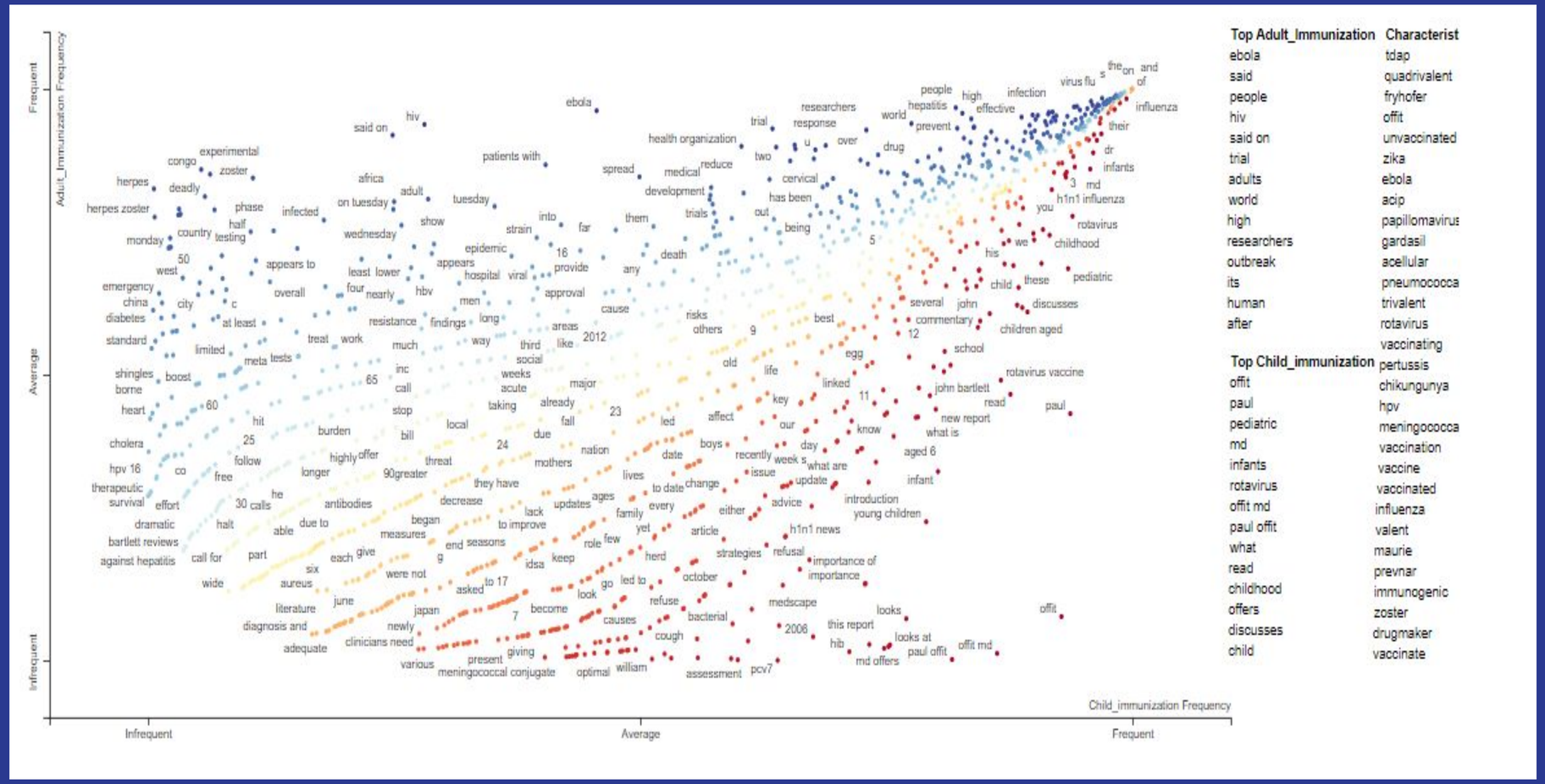
@wimlds_lagos

View of Children Immunization Text Word Cloud



@wimlds_lagos

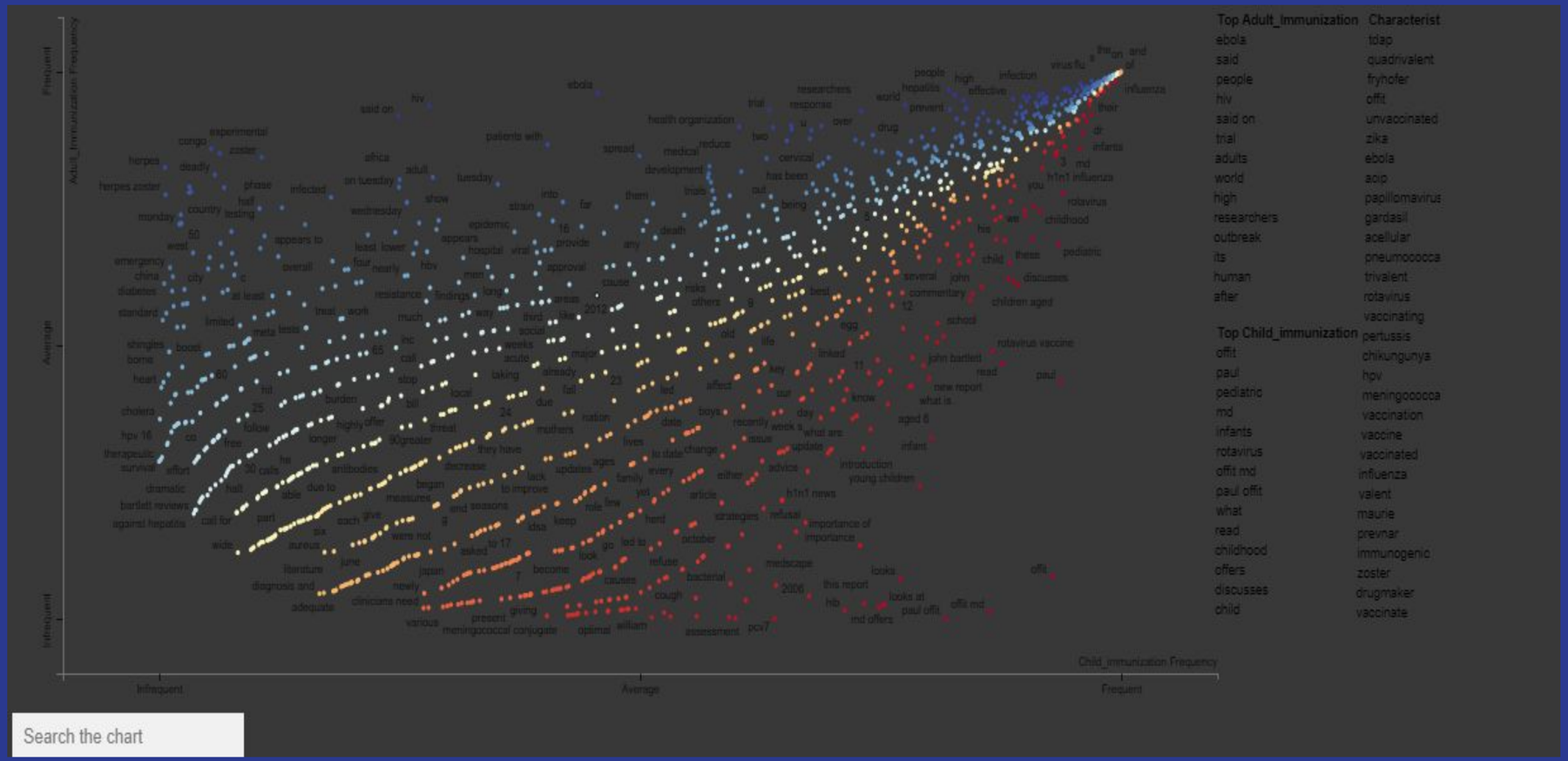
View of Immunization Analysis Scattertext



[Interactive link]

@wimlds_lagos

View of Immunization Analysis Scattertext



Interactive link

@wimlds_lagos

References and Further Reading

Scattertext 0.0.2.66
https://github.com/JasonKessler/scattertext

Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ
https://arxiv.org/abs/1703.00565

Text and document Visualization :
http://jcsites.juniata.edu/faculty/rhodes/ida/textDocViz.html

Exploratory Data analysis and Data Visualization
https://www.creative-wisdom.com/teaching/WBI/EDA.shtml

https://courses.cs.washington.edu/courses/cse512/15sp/lectures/CSE512-Text.pdf

https://www.datacamp.com/community/tutorials/wordcloud-python