HIV Dataset

Context:

The Publication regions include the India, Indonesia, Vietnam, Afghanistan, Bangladesh from around 13 Websites. The data was scraped using BeautifulSoup, Selenium and Newspaper3K. Initially, each publication HIV articles are stored in a separate CSVs and In Final concatenated all CSVs into one file. Data sources that were missing a label were simply assigned a label of "Nan".

The dataset is consisting of 10 columns following parameters

* Headlines
* Description
* Authors
* Published Dates
* Publication
* Category
* Article
* Keywords
* Summary
* URL

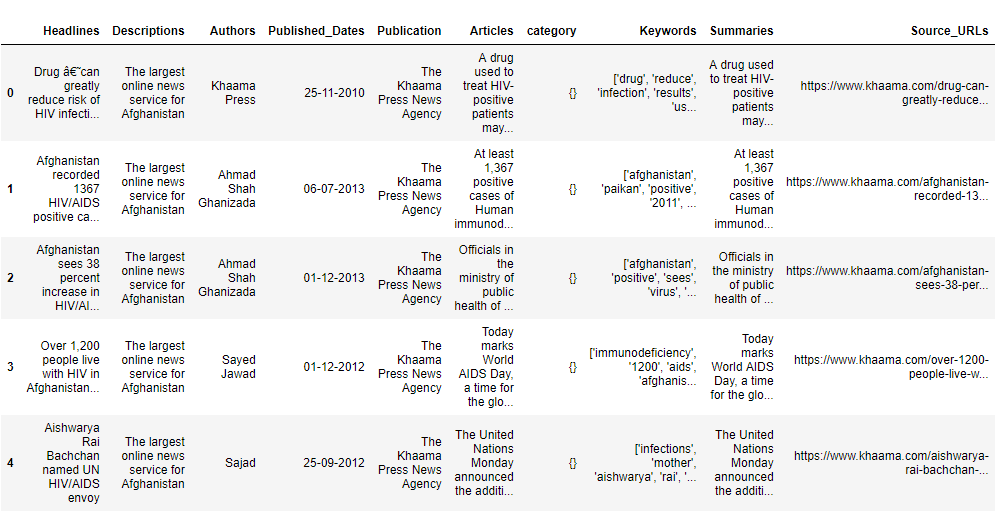
and it represents 2295 HIV posts in total. The data primarily falls between the years of 2000 and 2019, there is possibly insignificant number of articles from 2000.

Content:

Dataset has chopped into two CSVs file by concatenating all CSVs files where the initial dataset was (2295,10) after cleaning the dataset by removing Nan values. The resultant dataset was (580, 10)

Final\_HIV.csv – 2295 news articles (Articles 1- 2295)

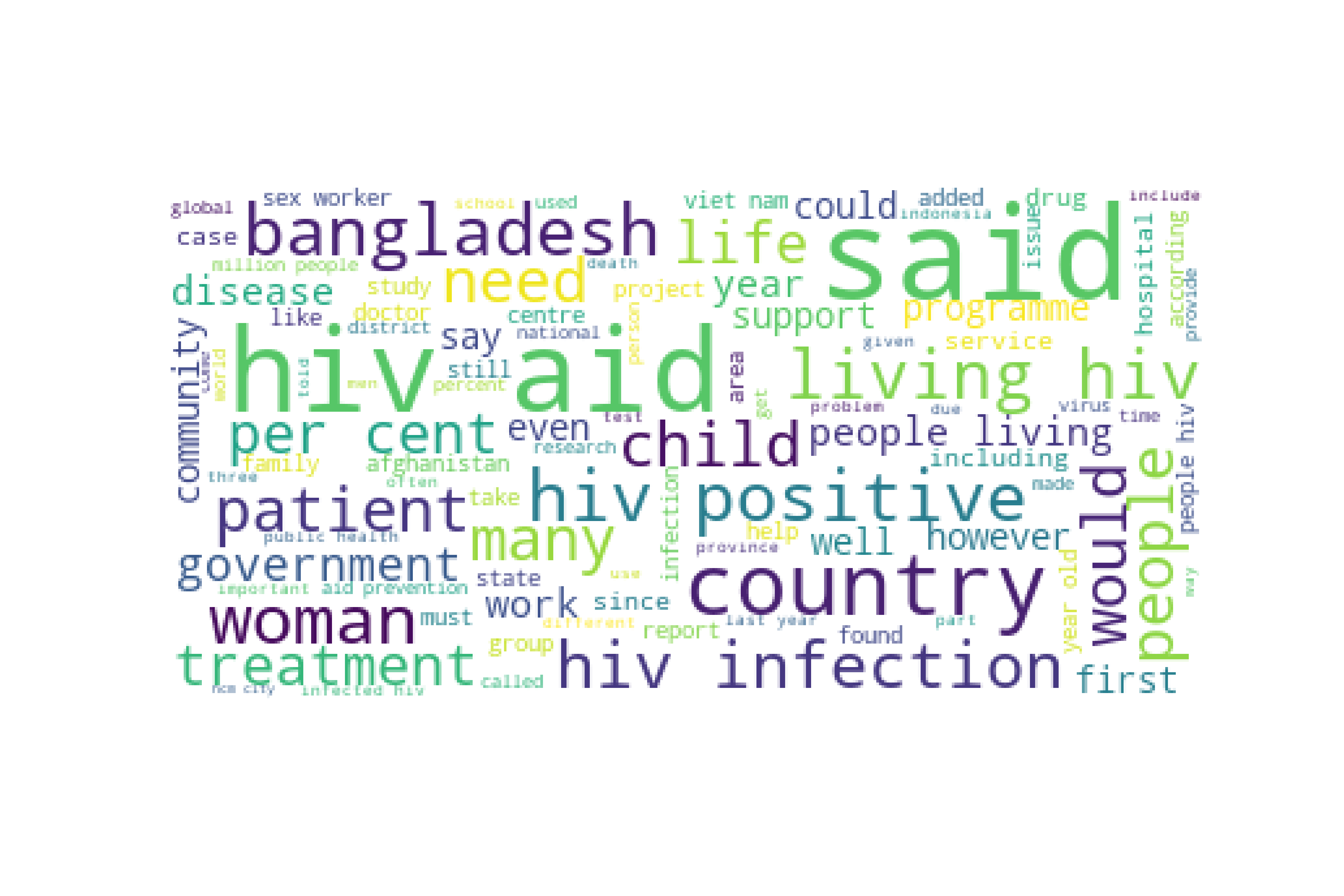
Final\_HIV\_without\_Nan.csv – 580 news articles (Articles 1-580)



Data Extraction:

The Data Extraction consists of above parameters. While scraping the news details, every news article has been cleaned and it doesn’t require further cleaning. While scraping, every article undergoes basic NLP process like Keywords and Summary of the article.

NLP:

The text corpus that we created after pre-processing to get insights on the most frequently used words

Automated Keyword Extraction from HIV Articles using NLP from the collected data.

