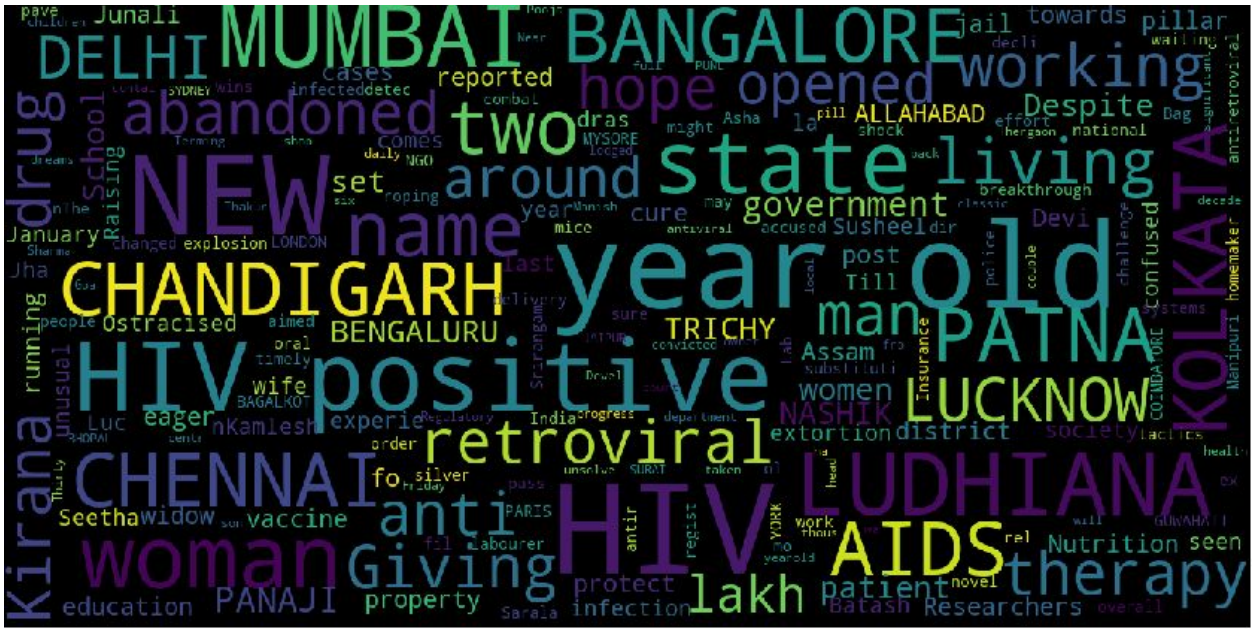
TEXT-ANALYSIS OF NEWS ARTICLES ON HIV

A detailed analysis on HIV related articles on The Times of India (2010-18)



(IMG : Word cloud generated from the news articles' text-corpus)

By : Mr. Rajaswa Ravindra Patil

25.12.2018

BITS Pilani, Goa Campus.

(f20170334@goa.bits-pilani.ac.in | 9834623198)

INDEX

ACKNOWLEDGMENTS

This task covered everything from data collection to data analysis and data visualization. I would like to thank Professor Dr. Sukant Khurana for his continuous support and motivation without which I would have never come across this task and completed it successfully .

DATA :

The data has been web scraped from the **archives** on official website of **The Times of India**. A **total of 1228 news articles** (from **1-1-2010** to **16-12-2018**) related to HIV have been collected using various **Python** packages. The data has been stored in a CSV file for proper analysis. Various features of the news articles like the author name, date posted, keywords etc have been captured simultaneously and the data in its raw form looks like this:

	DATE	AUTHOR	IMAGE	TEXT	KEYWORDS	SUMMARY
0	8-1-2010	[]	https://static.toiimg.com/photo /msid-5423538/5...	Junali Devi, a widow living in Assam, is eager...	[junali, told, hiv, special, life, website, th...	Junali Devi, a widow living in Assam, is eager...
1	11-1-2010	[Karthika Gopalakrishnan]	https://static.toiimg.com/photo /msid-47529300/...	CHENNAI: Researchers working towards a cure fo...	[virus, marrow, working, cure, hiv, molecule, ...	These cells have a second molecule called CCR5...
2	12-1-2010	[]	https://static.toiimg.com/photo /msid-47529300/...	LUCKNOW: After running from the pillar to post...	[job, ngos, state, organisations, hiv, woman, ...	The hapless woman was thrown out by her husban...
3	30-1-2010	[]	https://static.toiimg.com/photo /msid-47529300/...	BANGALORE: Nutrition, education and property r...	[state, hiv, better, education, kids, rights, ...	BANGALORE: Nutrition, education and property r...
4	1-2-2010	[Kounteya Sinha]	https://static.toiimg.com/photo /msid-47529300/...	NEW DELHI: A vaccine to protect HIV patients f...	[trial, hiv, shot, patients, immune, tb, studi...	NEW DELHI: A vaccine to protect HIV patients f...
5	2-2-2010	[Sanjeev Kumar Verma]	https://static.toiimg.com/photo /msid-47529300/...	PATNA: With two new anti-retroviral therapy (A...	[link, functional, state, hiv, medical, centre...	PATNA: With two new anti-retroviral therapy (A...
6	6-2-2010	[Sanjeev Kumar Verma]	https://static.toiimg.com/photo /msid-47529300/...	PATNA: With two new anti-retroviral therapy (A...	[link, functional, state, hiv, medical, centre...	PATNA: With two new anti-retroviral therapy (A...
7	7-2-2010	[]	https://static.toiimg.com/photo /msid-47529300/...	Two new HIV+ cases reported from district jail...	[varanasi, hiv, division, reported, prisons, c...	Two new HIV+ cases reported from district jail...
8	11-2-2010	[]	https://static.toiimg.com/photo /msid-47529300/...	ALLAHABAD: Ostracised by society and abandoned...	[commits, allahabad, hiv, youths, villagers, v...	ALLAHABAD: Ostracised by society and abandoned...
9	14-2-2010	[Ashish Tripathi]	https://static.toiimg.com/photo /msid-47529300/...	LUCKNOW: When it comes to 'extortion', the Luc...	[local, hiv, locate, husband, petrol, woman, a...	LUCKNOW: When it comes to 'extortion', the Luc...

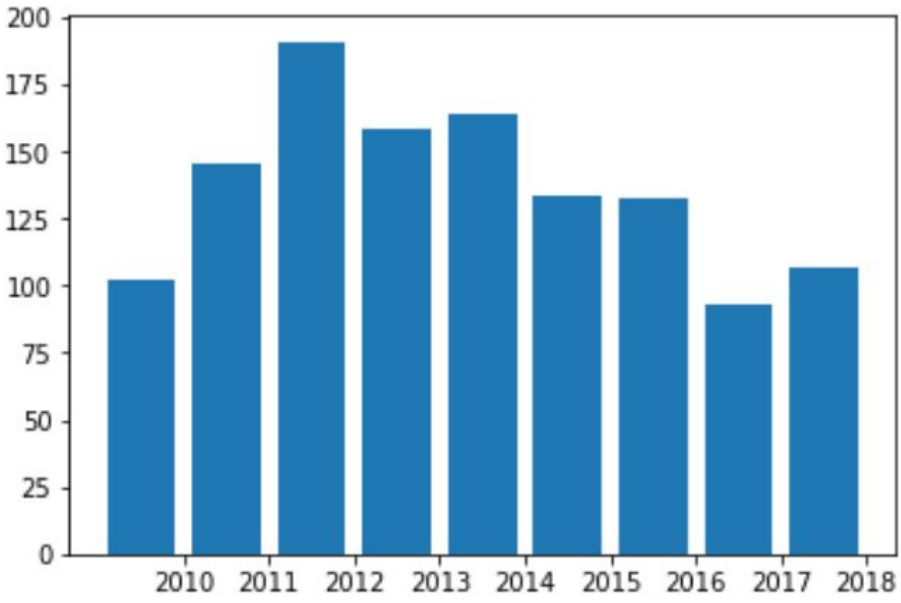
.

ABOUT THE CODE and RELATED FILES :

All the programming for this analysis report has been done in Python3. The code and all the data files can be found here.

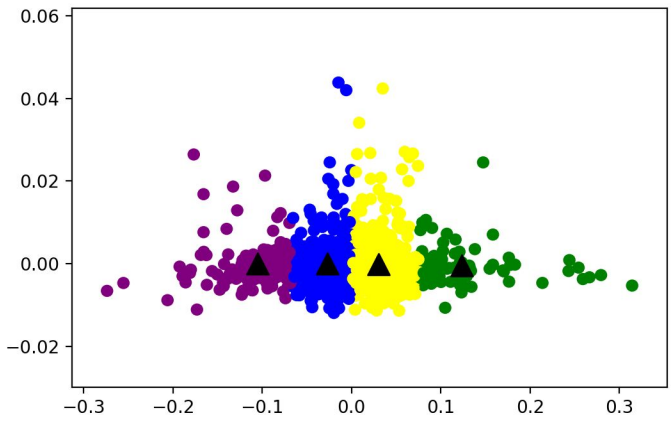
ANALYSIS & MOTIVATION :

Year-wise frequency distribution of the news articles:



The year 2012 had highest number of news articles (**191**), whereas year 2017 had least number of articles (**93**). On an a**verage 134** news articles on HIV have been posted each year.

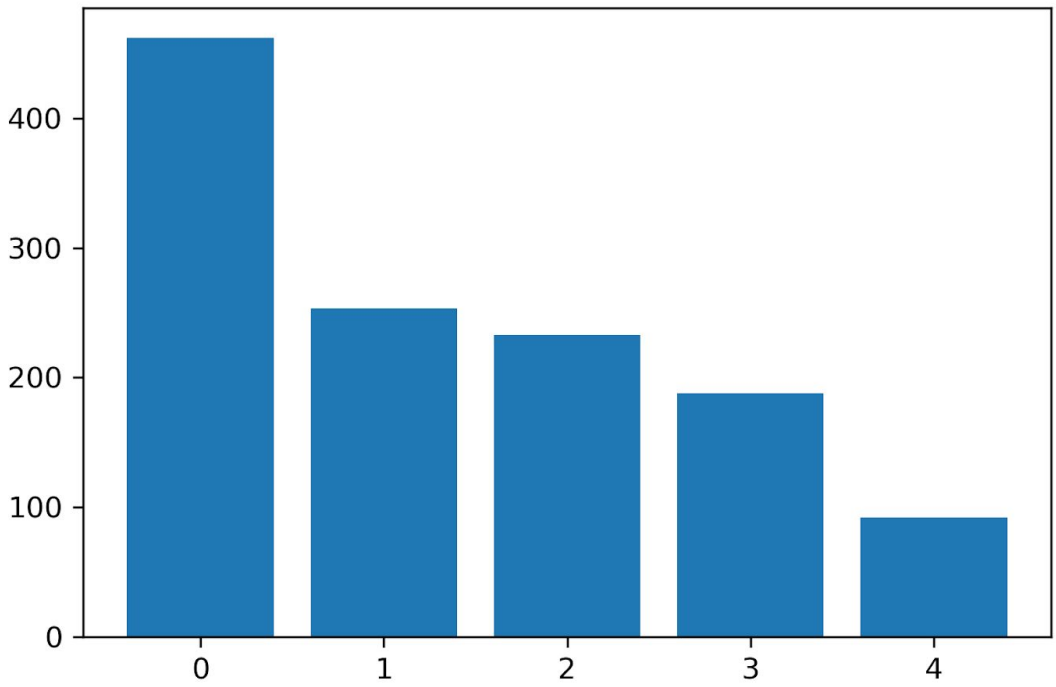
Applying **Principal Component Analysis** with with **doc2vec vectorization** and **K-means clustering** gives clear division of articles around 4 clusters in 2D as follows :



The distinct **visualization by PCA** suggests that the news articles can be perfectly categorized / classified into 4 or more classes (clusters).

1. K-means Clustering :

Using **TF-IDF vectorizer** on the news articles and applying **K-means clustering** with 5 clusters gives us following classification :

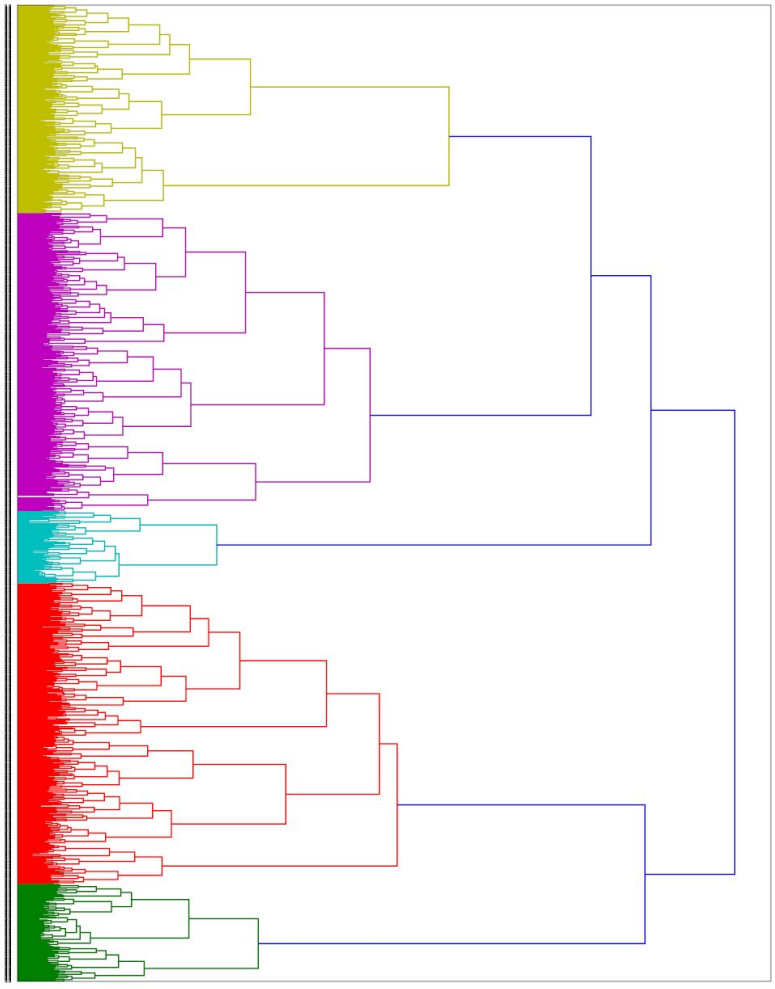


The keywords associated with the above clusters are as follows :

A. CLUSTER 0 : 462 articles ; [hiv, patients, aids, treatment, positive, health]
B. CLUSTER 1 : 188 articles ; [hiv, children, school, parents, students, child, aids]
C. CLUSTER 2 : 253 articles ; [hiv, sex, women, cases, prevalence, aids]
D. CLUSTER 3 : 233 articles ; [court, police, husband, medical, woman, patient]
E. CLUSTER 4 : 92 articles ; [transfusion, cbi, bank, probe, blood, probe]

Taking a different **hierarchical-clustering** approach gives us following **5 clusters** with their respective hierarchy displayed in the tree diagram below :

F. CLUSTER 0 : [hiv, positive, test, women, district]
G. CLUSTER 1 : [children, state, care, live, hiv]
H. CLUSTER 2 : [blood, virus, report, infect. patient]
I. CLUSTER 3 : [virus, health, drug, india, use]
J. CLuSTER 4 : [medic, treatment, drug, doctor]

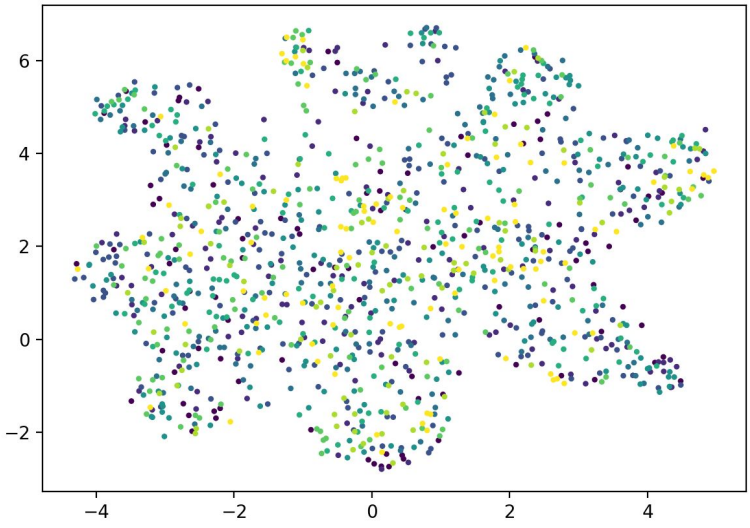


2. Latent Semantic Analysis (LSA) Topic Modeling :

Here, we use **LSA topic modeling** with **TruncatedSVD vectorization** to check whether over the 9 years of 2010-18, the articles related to HIV have shown any significant bias towards any topics. We set the LSA model to get 9 distinct topics for 9 distinct years :

A. TOPIC 0 : [patient, children, hospital, treatment]
B. TOPIC 1 : [police, husband, woman, court]
C. TOPIC 2 : [children, school, education, students]
D. TOPIC 3 : [blood, transfusion, sex, virus]
E. TOPIC 4 : [researchers, scientists, vaccine, cells, virus]
F. TOPIC 5 : [centres, state, blood]
G. TOPIC 6 : [woman, pregnant, mother, child]
H. TOPIC 7 : [patient, doctor, medical, surgery]
I. TOPIC 8 : [court, cbi, women]

The scattered plots of the articles (**9 color codes for 9 years**) show that the articles from all the years are more or less **scattered uniformly** over all the topics :



This shows that **none of the years (2010-18) show any significant bias towards any specific topics**.

3. Latent Dirichlet Allocation (LDA) Topic Modeling :

Here we use **LDA Topic Modeling** to classify the news articles over **5 clusters**. We will also calculate the **weights of the major topic words** in each cluster **to quantify their relevance** to that particular cluster of news articles.

A. **TOPIC 0 :** [('hospital', 0.01672098),
('patient', 0.01566928),
('positive', 0.010242492),
('woman', 0.009367754),
('doctor', 0.009104535),
('medic', 0.009065249),
("'s", 0.008623805),
('police', 0.008275199),
('famili', 0.0070295325),
('husband', 0.0068002935),
('test', 0.0061585666),
('treatment', 0.0061373697),
('also', 0.005716172),
('report', 0.00563402),
('told', 0.005519625),
('alleged', 0.005346895),
('ask', 0.0049371943),
('case', 0.004911922),
('would', 0.004845531),
('two', 0.0048310827)]

B. **TOPIC 1 :** [('people', 0.020363407),
('test', 0.01892579),
('posit', 0.015246991),
('centr', 0.012578322),
('state', 0.012078756),
('patient', 0.0110293375),
('case', 0.0101225),
('district', 0.009700607),
('number', 0.0093120225),
('health', 0.009074151),
('also', 0.00895092),
('year', 0.008839393),
('govern', 0.0080328435),
('women', 0.0075533874),
('infect', 0.007291588),
('awar', 0.0072338944),
('live', 0.006893148),
('treatment', 0.00632943),
('provide', 0.0060324157),
('programm', 0.0060080374)]

C. **TOPIC 2 :** [('children', 0.029168006),
('blood', 0.02565352),
('test', 0.0131369345),
('school', 0.011788974),
('child', 0.011215503),
('infect', 0.0108561),
('posit', 0.010019534),
('mother', 0.009689249),
('parent', 0.009687141),
('transfus', 0.0078100953),
('bank', 0.007448437),
('year', 0.0066817747),
("'s", 0.0063667707),
('girl', 0.006339033),
('babi', 0.006176628),
('also', 0.00594452),
('student', 0.0054827924),
('say', 0.005300665),
('one', 0.0050839265),
('govern', 0.0048792916)]

D. **TOPIC 3 :** [('infect', 0.018805092),
('virus', 0.014761212),
('patient', 0.01463189),
('drug', 0.014611903),
('treatment', 0.010450402),
('new', 0.007876726),
('studi', 0.007876687),
('cell', 0.007419501),
('use', 0.0068551465),
('people', 0.0061746673),
('research', 0.005831726),
('immun', 0.005361981),
('year', 0.0053046755),
("'s", 0.0049177194),
('first', 0.004834051),
('therapi', 0.0046369913),
('countri', 0.004522635),
('medicin', 0.0044802395),
('viral', 0.0043750457),
('prevent', 0.0043450478)]

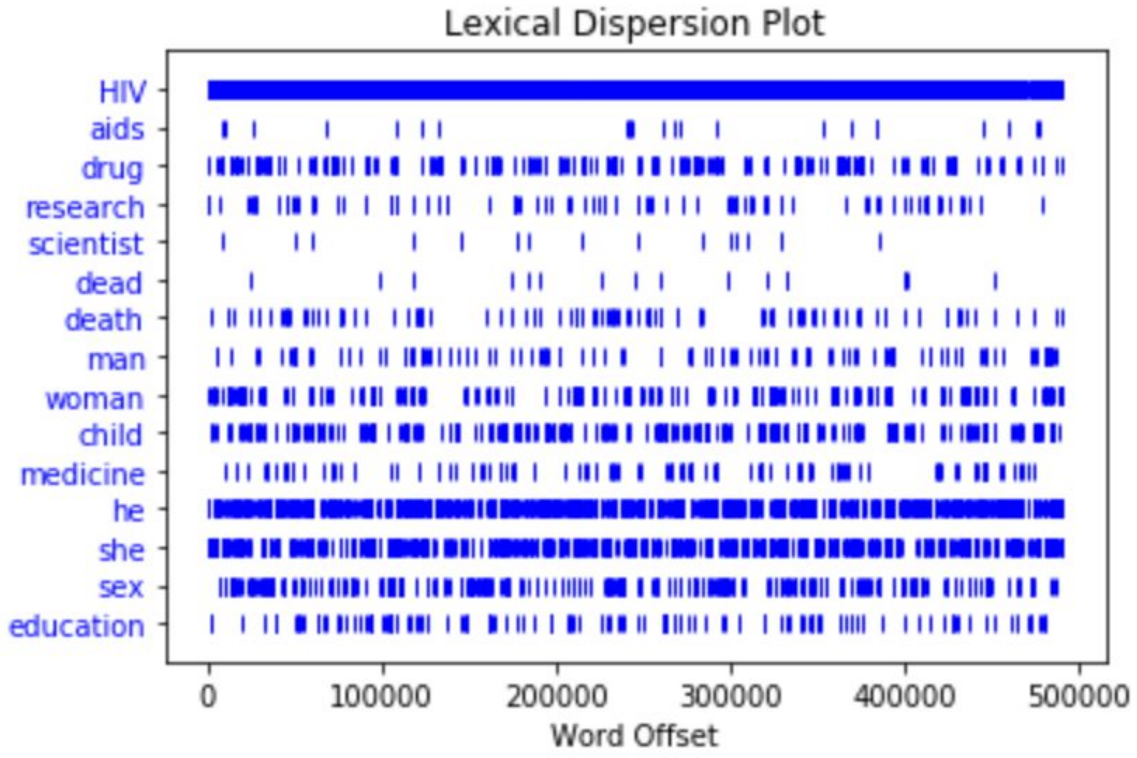
E. **TOPIC 4 :** [('sex', 0.027642984),
('among', 0.019010406),
('worker', 0.015549069),
('preval', 0.014537347),
('men', 0.013994094),
('state', 0.012041239),
('high', 0.011980874),
('year', 0.011428666),
('sexual', 0.008495244),

('infect', 0.007917218),
('popul', 0.007656649),
('case', 0.0073735085),
('group', 0.007176761),
('prison', 0.006840739),
('work', 0.006654726),
('condom', 0.0066277394),
('female', 0.0062295543),
('migrant', 0.006206697),
('risk', 0.0061404444),
('spread', 0.0055407267)]

4. Lexical Dispersion Plot (LDP) :

Though LDP is not a classification tool, it can provide very useful information over any text data chronologically.

Here is a LDP of a few terms of interest :



On the X-axis, the leftmost point (0,0) represents the starting point of our text data (i.e the first article from the year 2010) and the point (500000,0) represents the ending point of our text data (i.e the last article of 2018). On the Y-axis are some of the key terms of

interest. The bar next to each of these terms represents the term's occurrence during that particular time.

Some facts that can be concluded from the LDP :

A. Since the bar next to **HIV** is continuous throughout the timeline, we can conclude that all the articles in our data are HIV related.
B. The term **AIDS** has rarely occurred throughout the timeline, which means that the articles do not put significant stress on the disease.
C. The terms **drug, medicine, research** and **scientist** have occurred a significant number of times. Which means that the articles cover the active research for a drug against HIV. According to the LDP, these terms have had less presence in the news in recent years.
D. The term **dead** and **death** occur a fewer number of times, which means that HIV hasn't been that fatal in the news articles' coverage.
E. The terms **woman** and **child** have occurred more number of times than the term **man**.
F. The terms **sex** and **education** have had their continuous presence in the articles showing their relevance to HIV.

These were some of the key terms manually selected by me. Similarly LDP of any key term / word can be obtained easily.

CONCLUSION :

A. Among the 3 unsupervised classification techniques used, LSA Topic Modeling gives a precise classification. LSA also distinctly separates out the scientific research related articles related to HIV.
B. Whereas, LDA Topic modeling not only classifies the articles, but also provides the relevance of each word to their own group of articles.
C. K-means clustering shows a bit of overlapping during classification.
D. Hierarchical-clustering provides the hierarchy of clusters to further study the pattern in distribution of text.