UIET CSJMUNIVERSITY,KANPUR

COMPUTER SCIENCE & ENGINEERING DEPARTMENT

CONNECTING THE DOTS BETWEEN NEWS ARTICLES

Guide :

Dr. Renu Jain

Presents :

SAURABH SHUKLA

CSJMA16001390053

INDEX :

FIB(FILL IN BLANKS) PROBLEM :

- Given a sentence with a blank.
- We need to fill this blank with an apropriate word.

For example:

1.In this COVID-19 pandemic , to be safe we should wash our hands ……

✓a) frequently
b) after 2 hour
c) after 1 day
d) shouldn't wash

CONT.

For example:

2.I enjoyed …… the film but I didn't like the ending.

a) some of
✓b) most of
c) most
d) some

For example:

3.I have no horse …… I can lend you.

a) who
✓b) that
c) whose
d) which

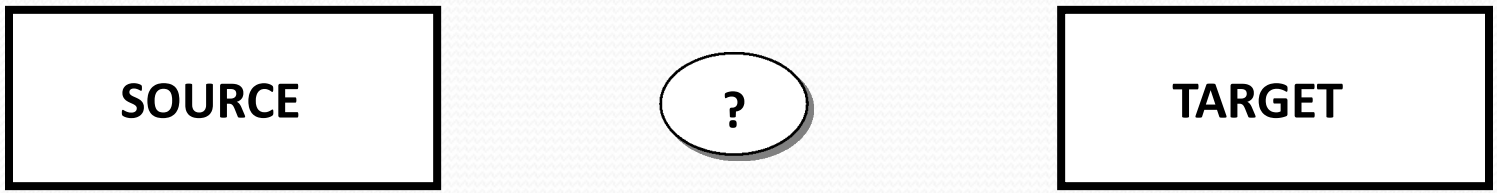
CONT.

Can we say ? :

- After filling the most suitable word ,A meanigful sentences can be formed.

- A wrong selection of word can differ the meaning of sentences.

OUR PROBLEM :



CONT.

Similarly we can say :

- Like FIB problem we have to fill these blanks/question mark with the articles.
- so that the story or the news article make sense when we will read.
- Here question mark can be filled with these given(below) articles.

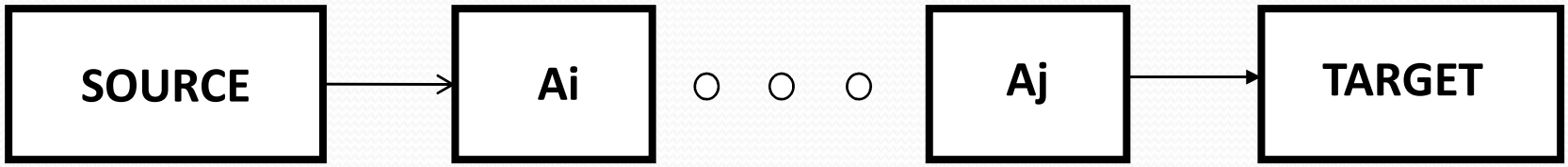
A1 ,A2 ,A3 ,…… ,An

Note :

- articles can be type of sports,politics, entertainment.

CONT.

So our aim is to develop a coherent chain linking the source and target documents into a meaningful story



Here Ai,Aj are intermediate articles.

CONT.

For example :

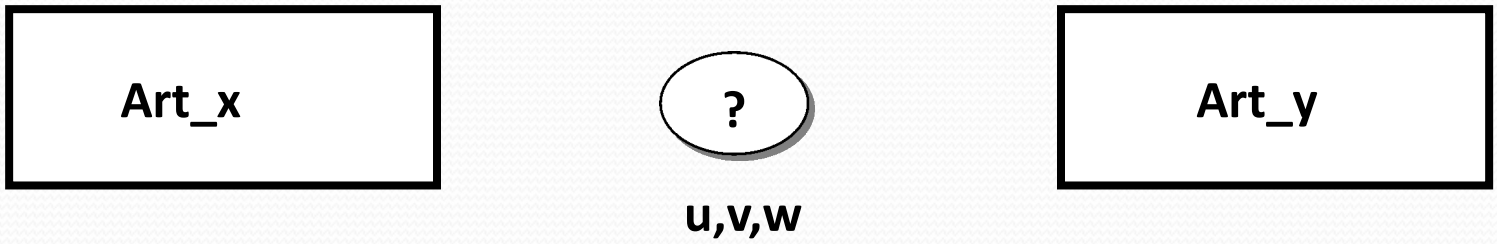
- 1.it can discover the hidden chain between JNU FEE HIKE PROTEST and CAA,NRC,NPR protest.

- 2. it can discover the chain between Jessica lal murder case and Nitish katara murder case.

Latest exapmle:

- 3.It can discover the chain between **COVID-19** pandemic and economic crisis(impact).

EXAMPLE :



Linking articles index :

x,u,v,w,y

Link Topics :

Art_x_ **COVID-19** pandemic

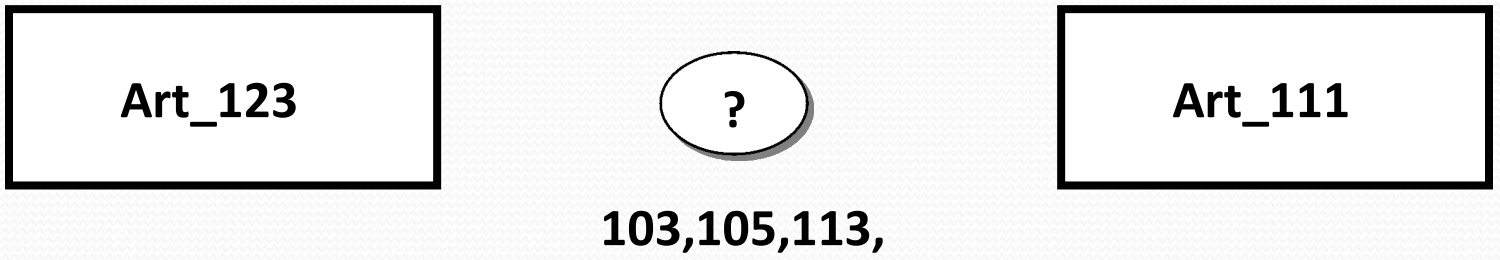
Art_u_Lockdown in india

Art_v_demand decreases

Art_w_manufacturing closed due to lockdown

Art_y_economy crisis(impact)

EXAMPLE :



Linking articles index :

123,103,105,113,111

Link Topics :

Art_123_68% turnout in Gujarat Phase-I

Art_103_Gujarat final phase poll tomorrow

Art_105_Development won in Gujarat

Art_113_BJP victory celebrated

Art_111_Modi elected leader of Gujarat BJP

Our approach :

INPUT :

PICK TWO ARTICLES
(START,TARGET)

OUTPUT :

BRIDGE THE GAP
With a smooth chain of article

- Finally our game plan is to find out a good chain/a relevant chain

Work flow :

STEP I : Chain Computation:

➢For a similarity measure between documents we use the Bhattacharya's distance.

Bhattacharyya's Distance

$DB = - \ln (BC(p,q))$

where $BC(p,q) = \sum_{x \in X} (p(x).q(x))^{1/2}$ is the Bhattacharyya coefficient

source:http://en.wikipedia.org/wiki/Bhattacharyya_distance

➢After generation of chain we will have to do its evaluation

CONT.

- We evaluate the characteristics of a good chain by its coherence
- We have different methods to find the coherence of a chain.

STEP 2 : Chain Evaluation :

Now that we have a chain, we evaluate our results using coherence of the output chain as a parameter.

- Coherence1(d1, …,dn) = i=1…n-1min Σw 1(w € di ∩ di+1)

- This is straight from the below mentioned paper of Dafna Shahaf.

Source(ref) : Dafna Shahaf and Prof. Carlos Guestrin : Connecting the dots between news articles. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2010

- Finally, we compare different combinations of weight-assignment methods and shortest path algorithms to see which combination gives us a better coherence.

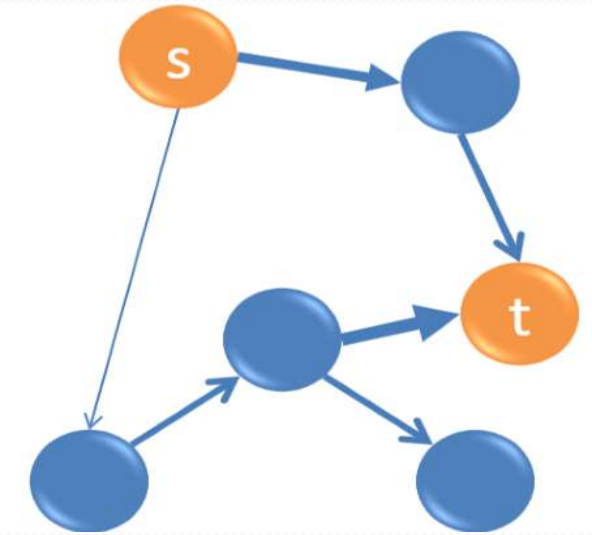
WHAT MAKES THE STORY GOOD?

➢A good chain makes the story good.

What is a good chain?

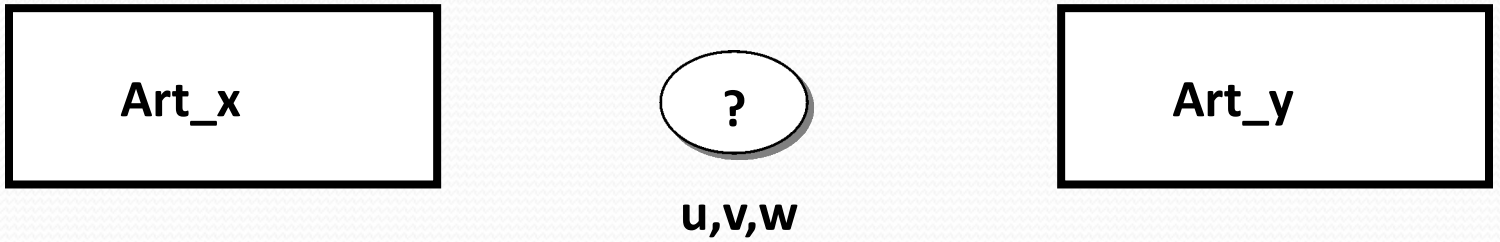
➢A good chain should have these attributes

•Every transition is strong
•No jitteriness(back - and - forth)
• Short(5-6 articles)



Source:google images

CONT.



Linking articles index :

x,u,v,w,y

Link Topics :

Art_x_ **COVID-19** pandemic

Art_u_Lockdown in india

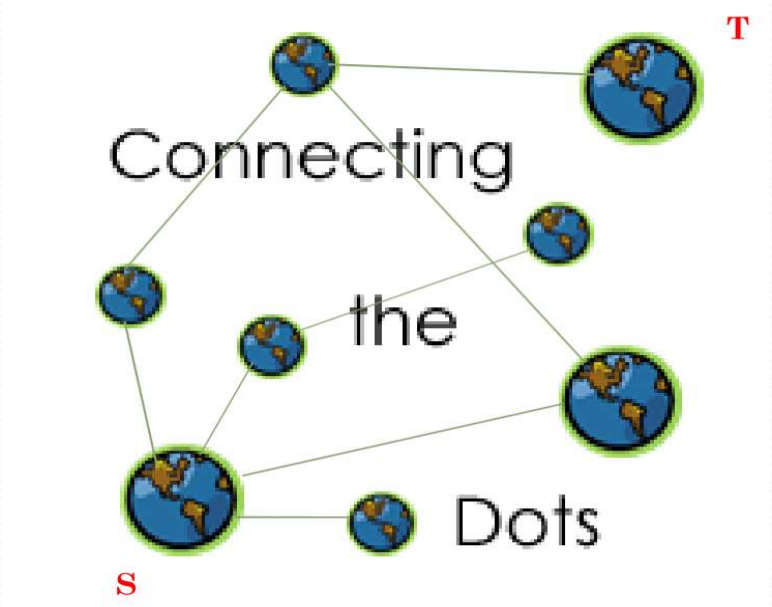
Art_v_demand decreases

Art_w_manufacturing closed due to lockdown

Art_y_economy crisis(impact)

Why did I select this topic? :

CONNECTING THE DOTS BETWEEN NEWS ARTICLES



Source:google images

CONT.

Why did I select this topic :

Getting information off the internet is like taking a drink from a fire hydrant.

-Mitchell Kapor

- News browsing is one of the primary uses of the internet
- Information Overload is everywhere so these days we are facing this problem.
- The problem is for entire sectors, from scientists to intelligence analysts and web use

CONT.

- all of whom are constantly struggling to keep up with the larger and larger amounts of content published everyday.

- With this much data, it is often easy to miss the big picture

We have problems of :

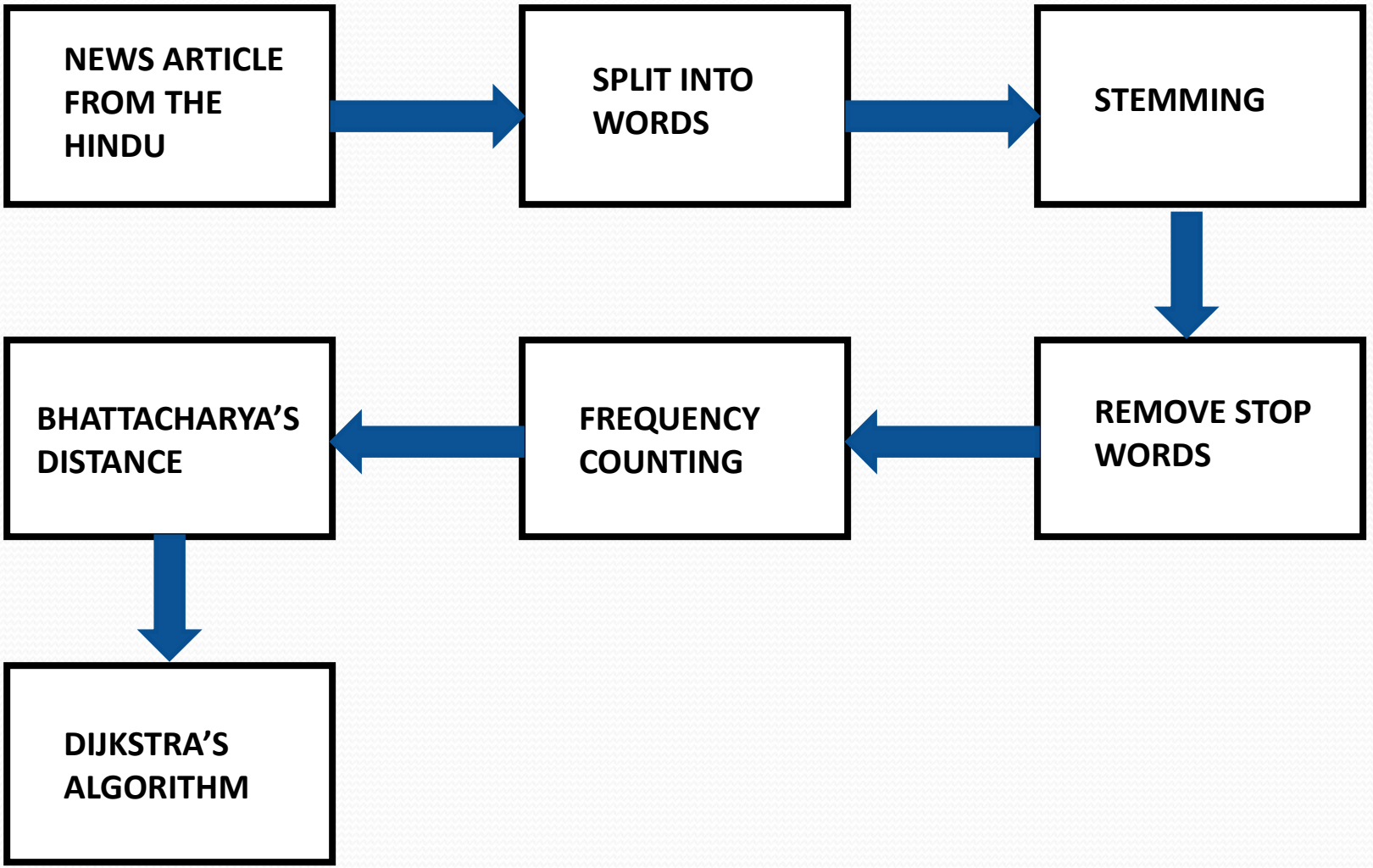
- Tackling information overload

- Navigate between topics

- Searching for relevant news is difficult

So This project will help to give you basic idea to tackle above problems.

WORK FLOW(CHAIN COMPUTATION) :



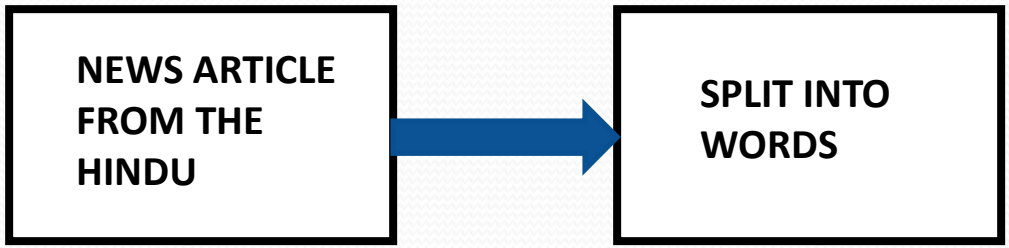
Example :

NEWS ARTICLE FROM THE HINDU

osama bin laden, leader of terror outfit al- qaeda and alleged mastermind of the 9/11 attacks in new york city, was killed in pakistan in a special forces operation by the united states, u.s. president barack obama announced on sunday night.

In a statement issued shortly after 11.30 p.m., mr. obama confirmed that osama, high on the list of u.s. authorities most wanted men, had been killed after a “fire-fight” in abbottabad, a military cantonment town not far from islamabad. **Len=100**

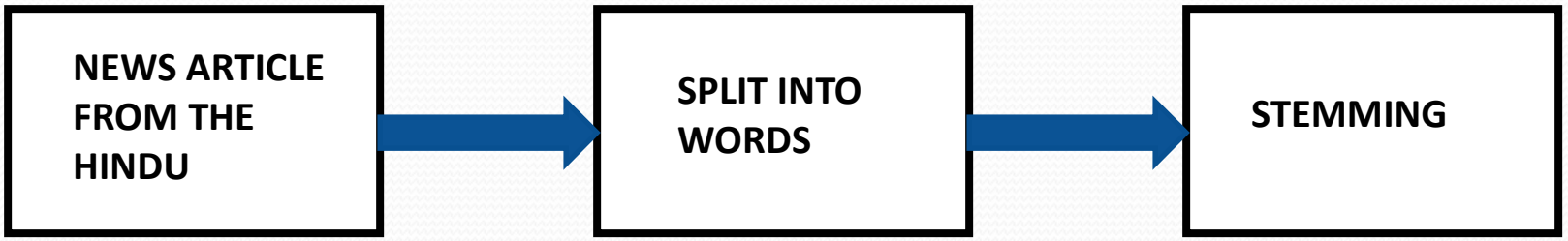
CONT.



['', 'osama', 'bin', 'laden', '', 'leader', 'of', 'terror', 'outfit', 'al-', 'qaeda', 'and', 'alleged', 'mastermind', 'of', 'the', '9/11', 'attacks', 'in', 'new', 'york', 'city', '', 'was', 'killed', 'in', 'pakistan', 'in', 'a', 'special', 'forces', 'operation', 'by', 'the', 'united', 'states', '', 'u', 's', '', 'president', 'barack', 'obama', 'announced', 'on', 'sunday', 'night', '', '', 'in', 'a', 'statement', 'issued', 'shortly', 'after', '11', '30', 'p', 'm', '', '', 'mr', '', 'obama', 'confirmed', 'that', 'osama', '', 'high', 'on', 'the', 'list', 'of', 'u', 's', '', 'authorities', '', 'most', 'wanted', 'men', '', 'had', 'been', 'killed', 'after', 'a', '‘“fire-fight”', 'in', 'abbottabad', '', 'a', 'military', 'cantonment', 'town', 'not', 'far', 'from', 'islamabad', '']

Len=100

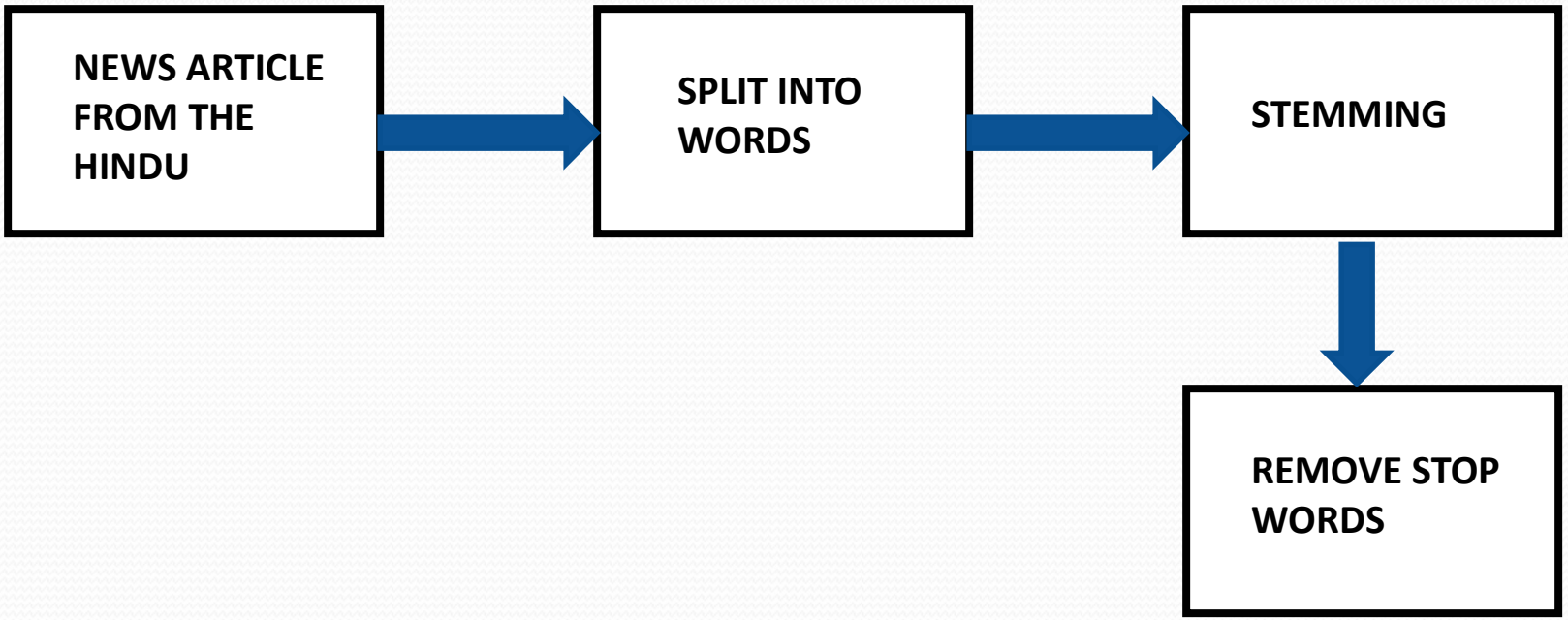
CONT.



['', 'osama', 'bin', 'laden', '', 'leader', 'of', 'terror', 'outfit', 'al-', 'qaeda', 'and', 'alleg', 'mastermind', 'of', 'the', '9/11', 'attack', 'in', 'new', 'york', 'citi', '', 'wa', 'kill', 'in', 'pakistan', 'in', 'a', 'special', 'forc', 'oper', 'by', 'the', 'unit', 'state', '', 'u', 's', '', 'presid', 'barack', 'obama', 'announc', 'on', 'sunday', 'night', '', '', 'in', 'a', 'statement', 'issu', 'shortli', 'after', '11', '30', 'p', 'm', '', '', 'mr', '', 'obama', 'confirm', 'that', 'osama', '', 'high', 'on', 'the', 'list', 'of', 'u', 's', '', 'author', '', 'most', 'want', 'men', '', 'had', 'been', 'kill', 'after', 'a', '‘“fire-fight”', 'in', 'abbottabad', '', 'a', 'militari', 'canton', 'town', 'not', 'far', 'from', 'islamabad', '']

Len=44

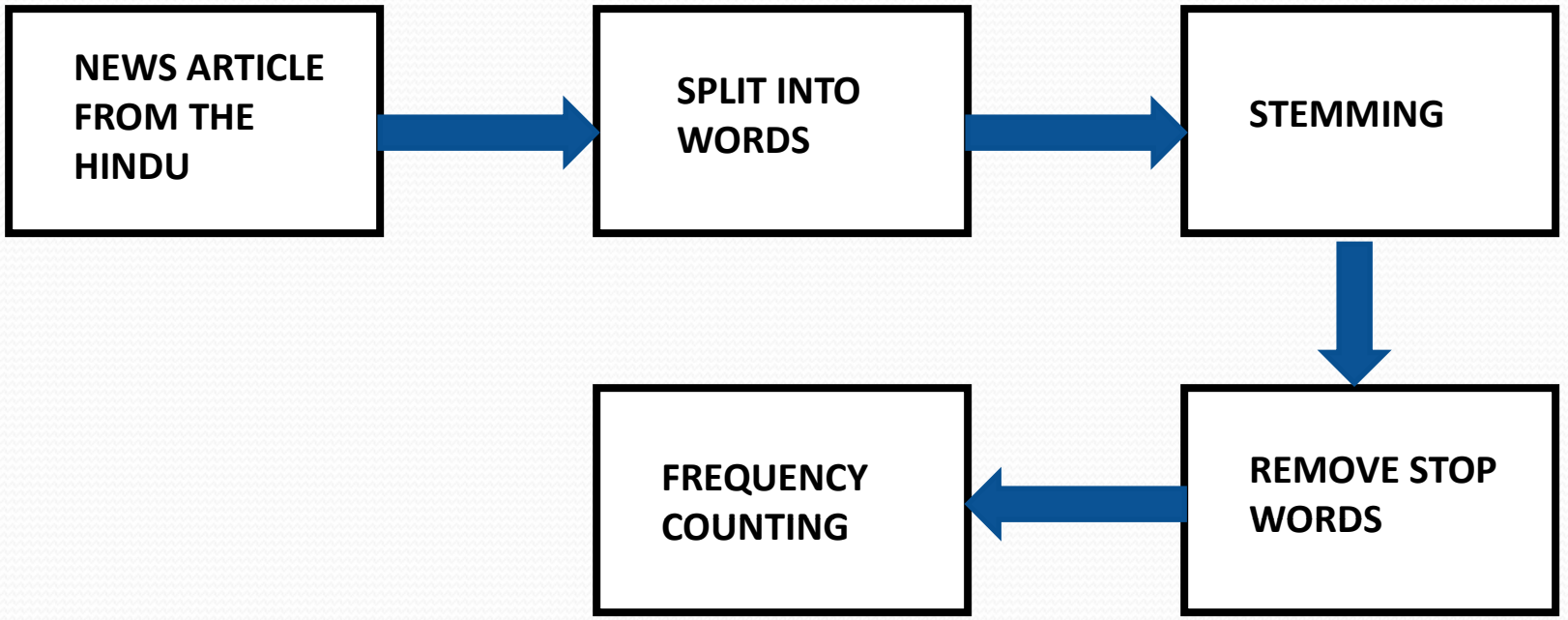
CONT.



['osama', 'bin', 'laden', 'leader', 'terror', 'outfit', 'al-', 'qaeda', 'alleg', 'mastermind', '9/11', 'attack', 'york', 'citi', 'wa', 'kill', 'pakistan', 'special', 'forc', 'oper', 'unit', 'presid', 'barack', 'obama', 'announc', 'sunday', 'night', 'statement', 'issu', 'shortli', '11', '30', 'obama', 'confirm', 'osama', 'list', 'author', 'kill', '"fire-fight"', 'abbottabad', 'militari', 'canton', 'town', 'islamabad']

Len=44

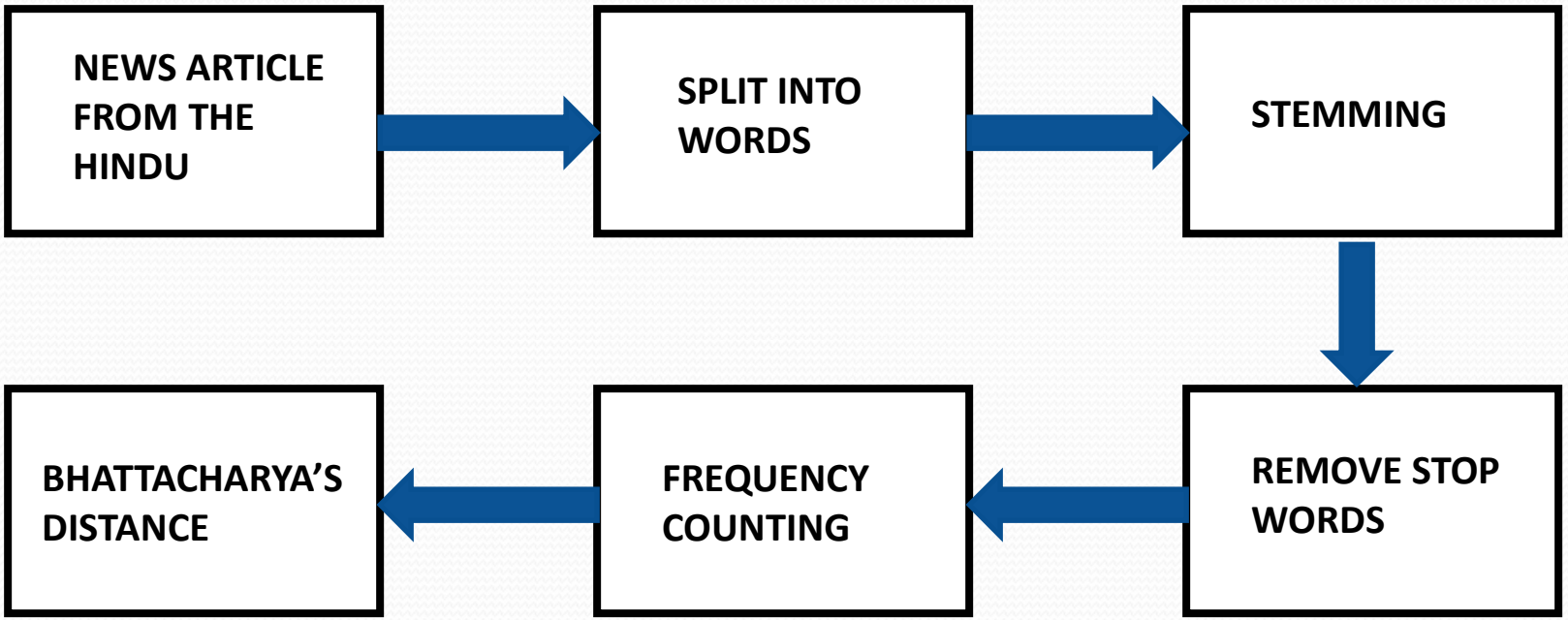
CONT.



[['bin', 1], ['laden', 1], ['leader', 1], ['terror', 1], ['outfit', 1], ['al-', 1], ['qaeda', 1], ['alleg', 1], ['mastermind', 1], ['9/11', 1], ['attack', 1], ['york', 1], ['citi', 1], ['wa', 1], ['pakistan', 1], ['special', 1], ['forc', 1], ['oper', 1], ['unit', 1], ['presid', 1], ['barack', 1], ['announc', 1], ['sunday', 1], ['night', 1], ['statement', 1], ['issu', 1], ['shortli', 1], ['11', 1], ['30', 1], ['confirm', 1], ['list', 1], ['author', 1], ['“fire-fight”', 1], ['abbottabad', 1], ['militari', 1], ['canton', 1], ['town', 1], ['islamabad', 1], ['osama', 2], ['kill', 2], ['obama', 2]]

Len=44

CONT.



Bhattacharyya's Distance

$DB = - \ln (BC(p,q))$

where $BC(p,q) = \sum_{x \in X} (p(x).q(x))^{1/2}$ is the Bhattacharyya coefficient

source:http://en.wikipedia.org/wiki/Bhattacharyya_distance

CONT.

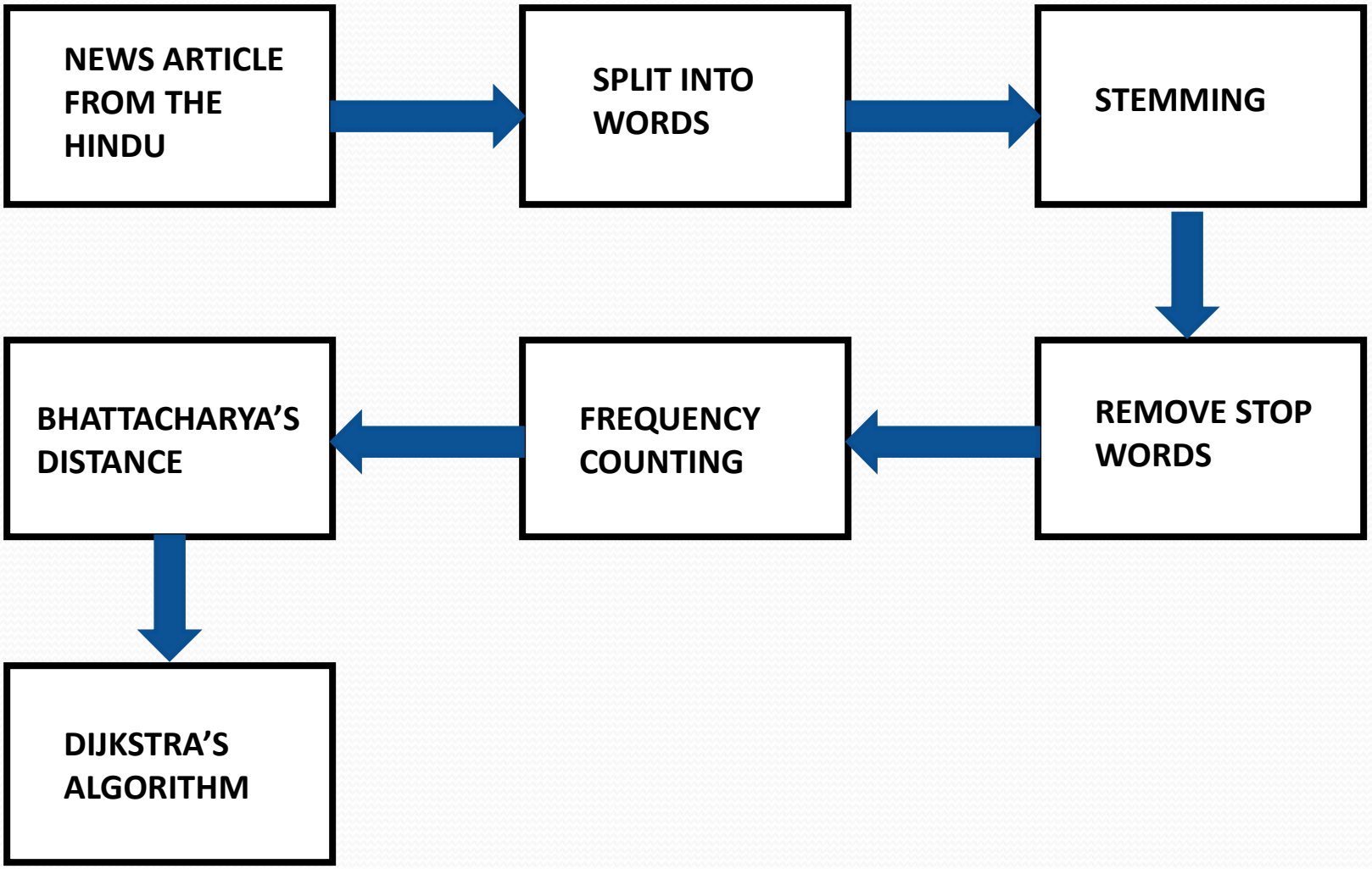
Bhattacharyya's Distance

DB = - ln (BC(p,q))

where BC(p,q) = $\sum_{x \in X} (p(x).q(x))^{1/2}$ is the Bhattacharyya coefficient

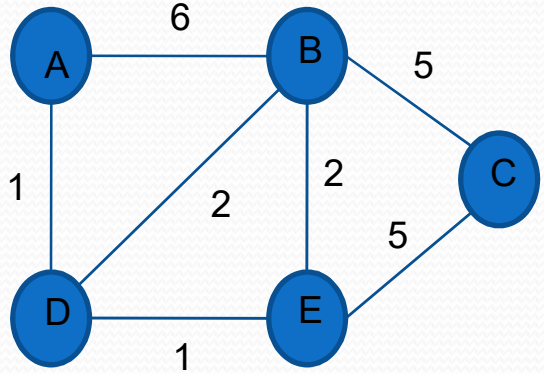
- We have used x as the common words between the two articles.
- p(x) is the number of occurrences of x in document p divided by total number of words in document p.
- q(x) is used similarly for document q.
- Bhattacharya's distance is closer to 0 for more similar articles
- farther than 0 for dissimilar articles due to the –ve logarithm taken of the Bhattacharya's coefficient.

CONT.


NEWS ARTICLE FROM THE HINDU
SPLIT INTO WORDS
STEMMING
REMOVE STOP WORDS
FREQUENCY COUNTING
BHATTACHARYA'S DISTANCE
DIJKSTRA'S ALGORITHM

	A	B	C	D	E
A	0	6	0	1	0
B	6	0	5	2	2
C	0	5	0	0	5
D	1	2	0	0	1
E	0	2	5	1	0

DIJKSTRA'S ALGORITHM



	A	B	C	D	E
A	0	6	0	1	0
B	6	0	5	2	2
C	0	5	0	0	5
D	1	2	0	0	1
E	0	2	5	1	0

VERTEX	SHORTEST DISTANCE FROM A	PREVIOUS VERTEX
A	0	
B	3	D
C	7	E
D	1	A
E	2	D

NOTE:

- Here we can use CosineSimilarity also as a weight instead of distance as a weight.

- With cosine similarity as weight we can use same algorithm Dijkstra,A* for chain computation

- Then we can compare the both results

Results :

➢Results on the same source and target article for different combination of methods.

Weight	Algorithm	Chain(linking articles index)
Bhatt. Dist	Dijkstra	345-343-330-349-342
Bhatt. Dist	A*	----------
CosineSimilarity	Dijkstra	----------
CosineSimilarity	A*	----------

Future Work

- We used here news articles .we can use different articles other than news articles may help in important scientific discoveries.

- With millions of articles are being produced daily worldwide, this process needs to be implemented onto a large scale.

References:

[1] Dafna Shahaf and Prof. Carlos Guestrin : Connecting the dots between news articles. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2010.

[2] http://en.wikipedia.org/wiki/Bhattacharyya_distance

[3] Dafna Shahaf , Prof. Carlos Guestrin and Eric Horvitz : Trains of thought-Generating information maps. International World Wide Web Conference (WWW), 2012.

- In this COVID-19 pandemic,
- STAY AT HOME,STAY SAFE,STAY HAPPY.

Thank You!