CONNECTING THE DOTS BETWEEN NEWS ARTICLES

CSE – S518 Artificial Intelligence(MINI PROJECT)

SAURABH SHUKLA(CSJMA16001390053)

Guide: Dr. Renu Jain

Abstract:

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. Thousands of news articles are published everyday, indicating an extensive media ranging from sports and politics to culture. 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 user. all of whom are constantly struggling to keep up with the larger and larger amounts of content published every day. With this much data, it is often easy to miss the big picture.

In other context Also, we are facing lots of difficulties in navigating between different topics and to find the hidden connections between them. Methods for automatically connecting the dots. Given two news articles, the system automatically finds a coherent story. Better understanding of the progression of the story.

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

THE IDEA:

Acknowledgement:

I would like to acknowledge the contributions of some of the individuals who went out of their way in this **COVID-19** pandemic to help me complete this project.

First of all , I am very grateful to Dr. Renu Jain for giving the deep subject knowledge throughout this course .this leads to me to select this Project. she always keeps us motivated for project work.specially thanks to madam for keep patience.

I would also like to thanks Dafna Shahaf and Prof. Carlos Guestrin for writing the mentioned paper .which academically encouraged me to take up this project.

Previous Work:

The inspiration of this project comes from the paper published in 2010, “Connecing the dots between news articles” by Dafna Shahaf and Prof. Carlos Guestrin.

Introduction :

Our approach :

INPUT :

PICK TWO ARTICLES

(START,GOAL)

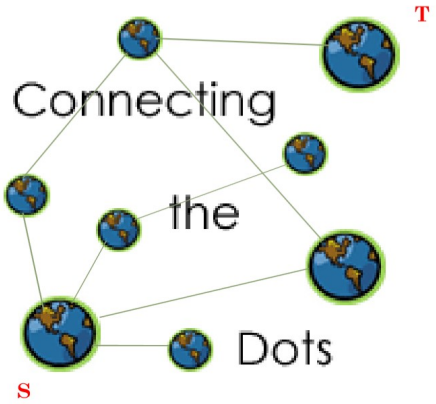
OUTPUT :

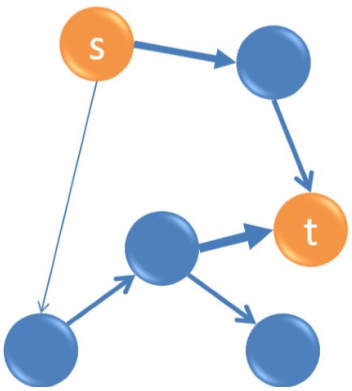
BRIDGE THE GAP

With a smooth chain of article

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

WHAT MAKES THE STORY GOOD?





Given a set of documents, a source and a target document, we aim at generating a coherent chain linking the source and target documents into a meaningful story.

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 example:

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

As we have selected the news domains, given two news articles, our aim is to generate a coherent chain linking them together.

WORK FLOW:

STEP I : Chain Computation:

For a similarity measure between documents we use the Bhattacharya's distance that is defined as:

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

Where

BC(p,q) = x € X Σ (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 thus calculated is closer to 0 for more similar articles and farther than 0 for dissimilar articles due to the –ve logarithm taken of the Bhattacharya's coefficient.

Using this Bhattacharya's distance as the edge weights between two documents, we model the entire dataset as a graph, and calculate the adjacency matrix for it.

On this adjacency matrix, when we get the source and target articles we compute the shortest path using one of the shortest path algorithms out of Dijkstra's algorithm and A* algorithm.

STEP 2 : Chain Evaluation :

Now that we have a chain, we evaluate our results using coherence of the output chain as a parameter. We also use the fact that a chain is only as strong as its weakest link. For this we use two evaluation measures:

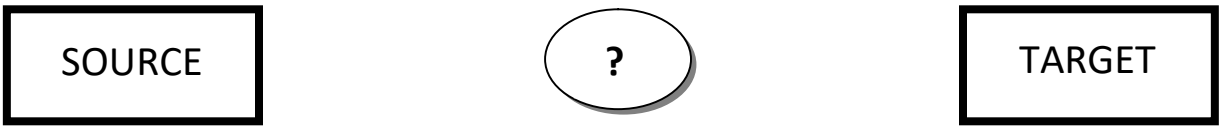
• **Coherence1**(d1, …,dn) = i=1…n-1min Σw 1(w € di ∩ di+1)

This is the minimal transition score .

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

For Example(worked example) :

Problem statement :

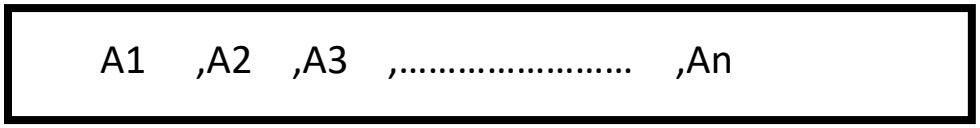


Here is given source and target article .we have to develop a chain of articles linking these two source and target article.

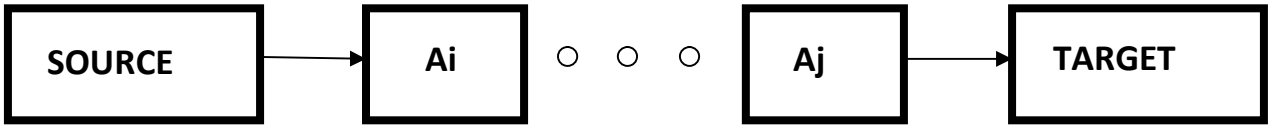
In short we have to develop a chain of article so that a relevant news or a meaningful story can be formed.

It is like FIB(fill in the blanks) problem.but here we have to fill these blanks with the news 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.



So the new picture will be



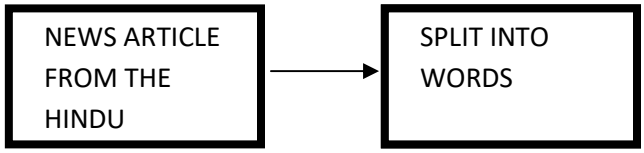
Here Ai,Aj are intermediate articles.

BLOCK DIAGRAM OF PROCESS:

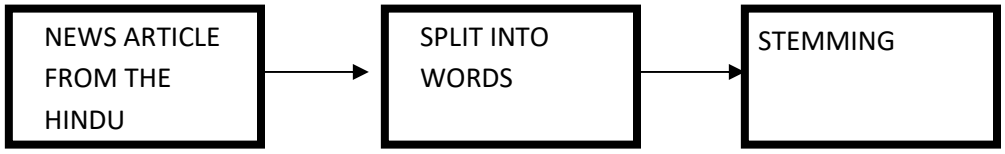
Osama bin Laden, leader of terror outfit al- Qaeda and alleged mastermind of the 9/11 attacks in New York City, was killed in Pakistan in a Special Forces operation by the United States, U.S. President Barack Obama announced on Sunday night.

NOTE:

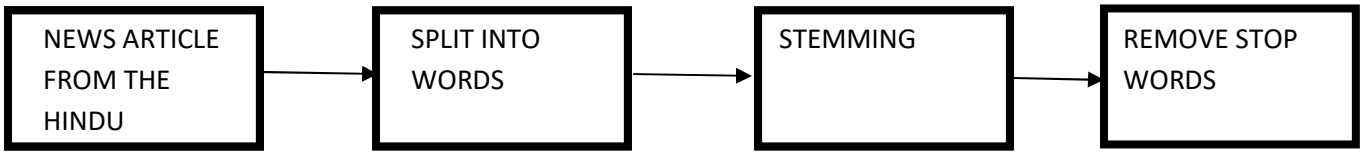
Here i have taken only one paragraph of article(345_US FORCES KILL OSAMA BIN LADEN) to show the process.



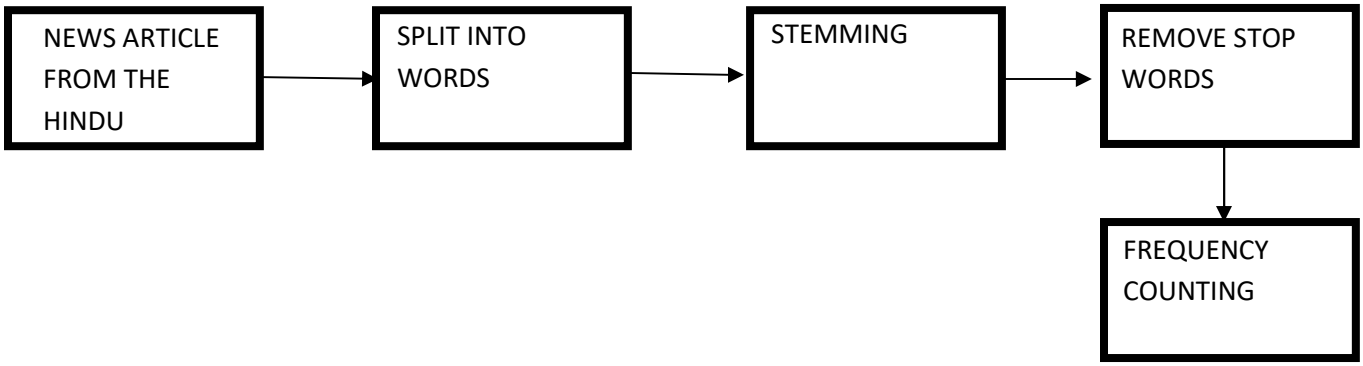
['', 'osama', 'bin', 'laden', '', 'leader', 'of', 'terror', 'outfit', 'al-', 'qaeda', 'and', 'alleged', 'mastermind', 'of', 'the', '9/11', 'attacks', 'in', 'new', 'york', 'city', '', 'was', 'killed', 'in', 'pakistan', 'in', 'a', 'special', 'forces', 'operation', 'by', 'the', 'united', 'states', '', 'u',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



['', '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', '']



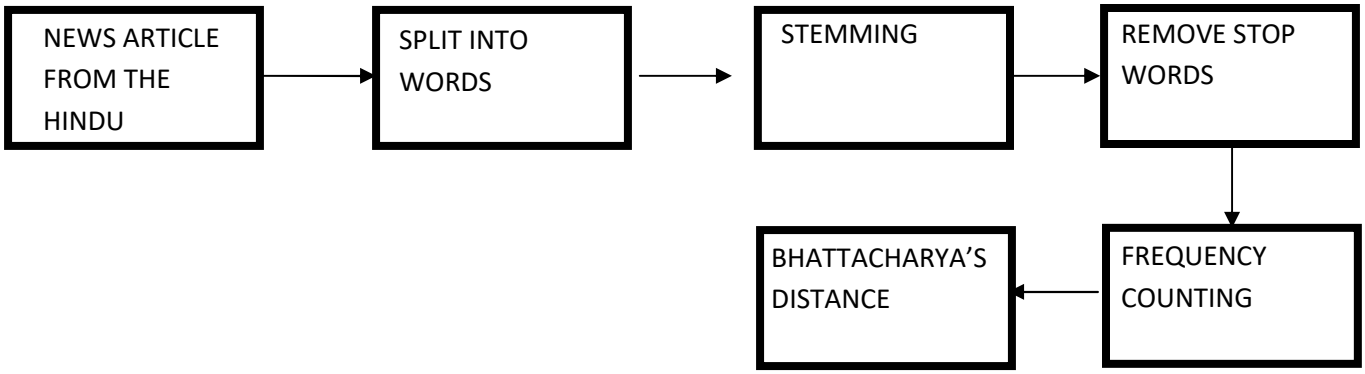
['osama', 'bin', 'laden', 'leader', 'terror', 'outfit', 'al-', 'qaeda', 'alleg', 'mastermind', '9/11', 'attack', 'york', 'citi', 'wa', 'kill', 'pakistan', 'special', 'forc', 'oper', 'unit', 'presid', 'barack', 'obama', 'announc', 'sunday', 'night']



[['osama',1], ['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],['kill',1],['pakistan',1],['special',1],['forc',1],['oper',1],['unit',1],['presid',1],['barack',1],['obama',1],['announc',1],['sunday',1],['night',1]]

NOTE:

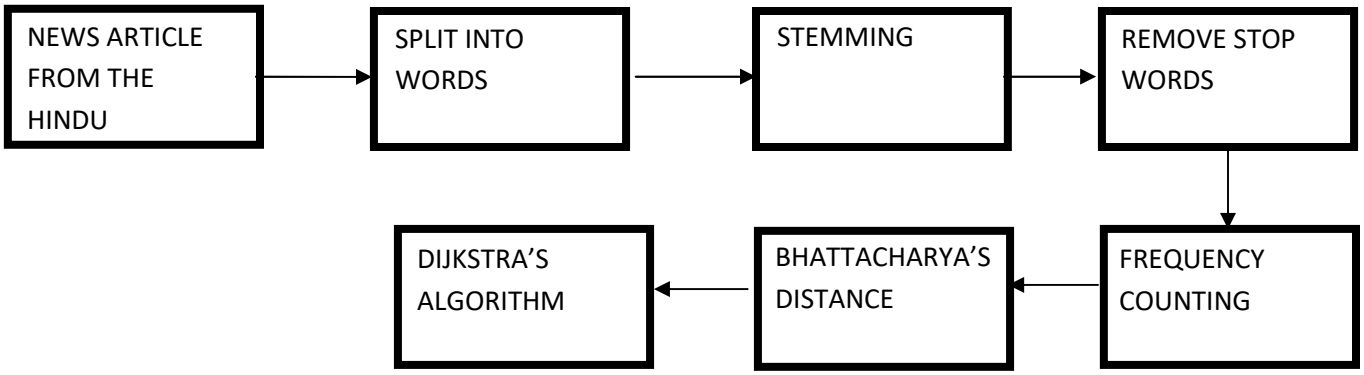
Here every word has only 1 frequency in the paragraph.
If any word appears two or more time in the paragraph.
For example if osama appears two time then
[‘Osama’,2]



Bhattacharyya's Distance

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

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



At last we have to make the good chain.

What is a good chain?

A good chain should have these attributes

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

Conclusions:

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*	345-331-344-346-342

Given a source article 'article345_US FORCES KILL OSAMA BIN LADEN' and a target article 'article342_HIS DEATH WILL BREAK THE IRON FIST OF AL_QAEDA IN IRAQ', the chain generated is

LINK TOPICS:

article345_US FORCES KILL OSAMA BIN LADEN

↓

article343_Inconceivable that Osama had no support system in Pakistan_ US

↓

article330_Osama bin Laden buried at sea

↓

article349_Osamas Pakistan home is no more

↓

article342_His death will break the iron fist of al_Qaeda in Iraq

After chain generation we will do its evaluation.

But due to time constraint we will not go into details of evaluation.

We can evaluate the characteristics of a good chain by its coherence. We can use two methods to find the coherence of a chain. One of them Is called coherence1.

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

This is straight from the above mentioned paper of Dafna Shahaf.

And the other method is called coherence2.

In this we use cosine similarity as below

- **Coherence2**(d1, …,dn) = i=1…n-1min { cosine-similarity (di , di+1) }

This measure similarity between the documents.

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

NOTE:

1.Here we can use CosineSimilarity also as a weight .

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

3.Then we can compare the both results.

4.We can also evaluate the chain which one is more good by comparing percentage of coherence but we will not go into the details of this due to time constraint.

Coherence2(d1, …,dn) = i=1…n-1min { cosine-similarity (di , di+1) }

This measure guarantees a value between 0 and 1, and hence, can be used as the percentage similarity between documents.

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