CONNECTING THE DOTS BETWEEN NEWS ARTICLES

GUIDE : Prof. Amitabha Mukerjee

Ankit Modi (10104)
Chirag Gupta (10212)

Problem Statement

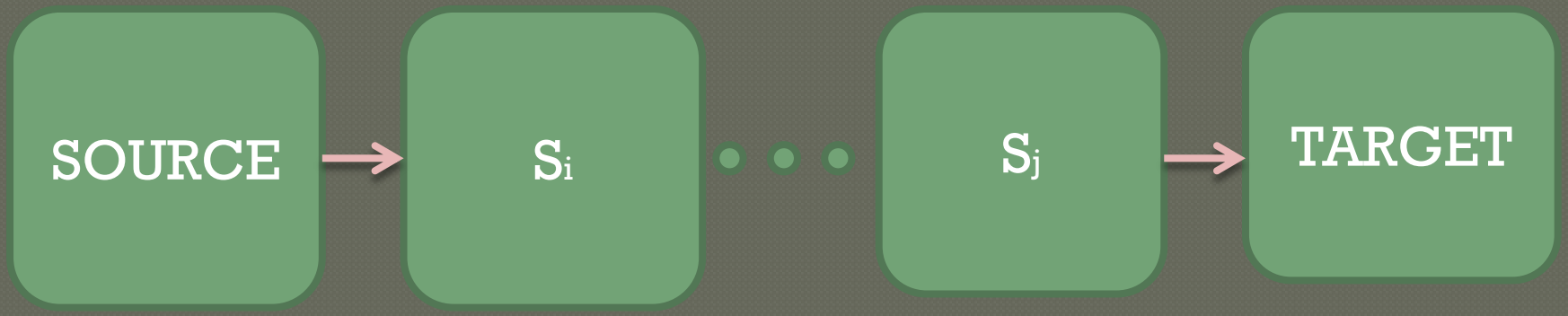
SOURCE

?

TARGET

$S_1, S_2, S_3, \ldots S_n$

Problem Statement



$S_1, S_2, S_3, \ldots S_n$

Motivation

- Problem ?
  Tackling *information overload*

Motivation

- Problem ?
  Tackling *information overload*

  Seeing *bigger picture*

Motivation

- Problem ?
  Tackling *information overload*

  Seeing *bigger picture*

  Navigate between topics

Motivation

- Domain ?

  News browsing : One of primary uses of Internet

  Politics, Sports, Entertainment etc

  Searching for relevant news is difficult

Framework

Corpus of news articles from The Hindu

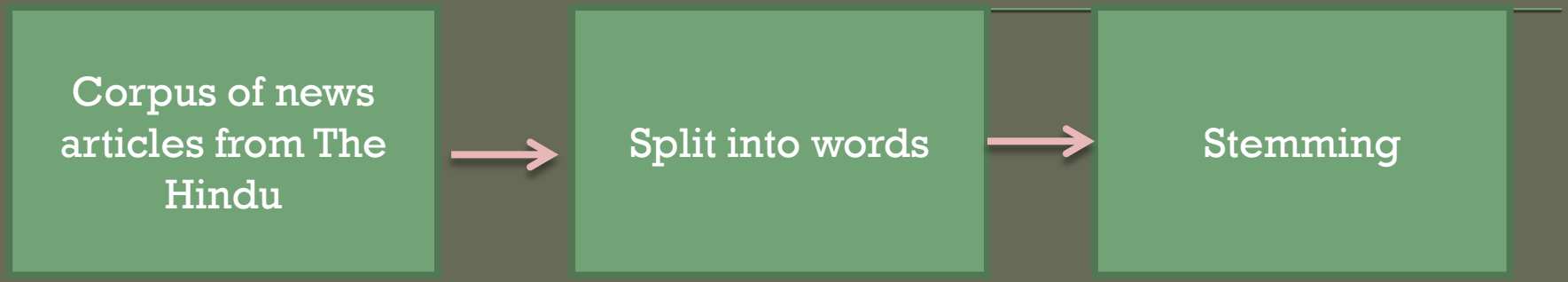
a delhi court on wednesday convicted sukhdev pehalwan, the third accused in the 2002 nitish katara murder case, saying that at the time of the incident he too was "present with convicts vikas yadav and vishal yadav," currently serving life term in tihar jail.

Framework

45

['a', 'delhi', 'court', 'on', 'wednesday', 'convicted', 'sukhdev', 'pehalwan,', 'the', 'third', 'accused', 'in', 'the', '2002', 'nitish', 'katara', 'murder', 'case,', 'saying', 'that', 'at', 'the', 'time', 'of', 'the', 'incident', 'he', 'too', 'was', 'present', 'with', 'convicts', 'vikas', 'yadav', 'and', 'vishal', 'yadav', 'currently', 'serving', 'life', 'term', 'in', 'tihar', 'jail', '']

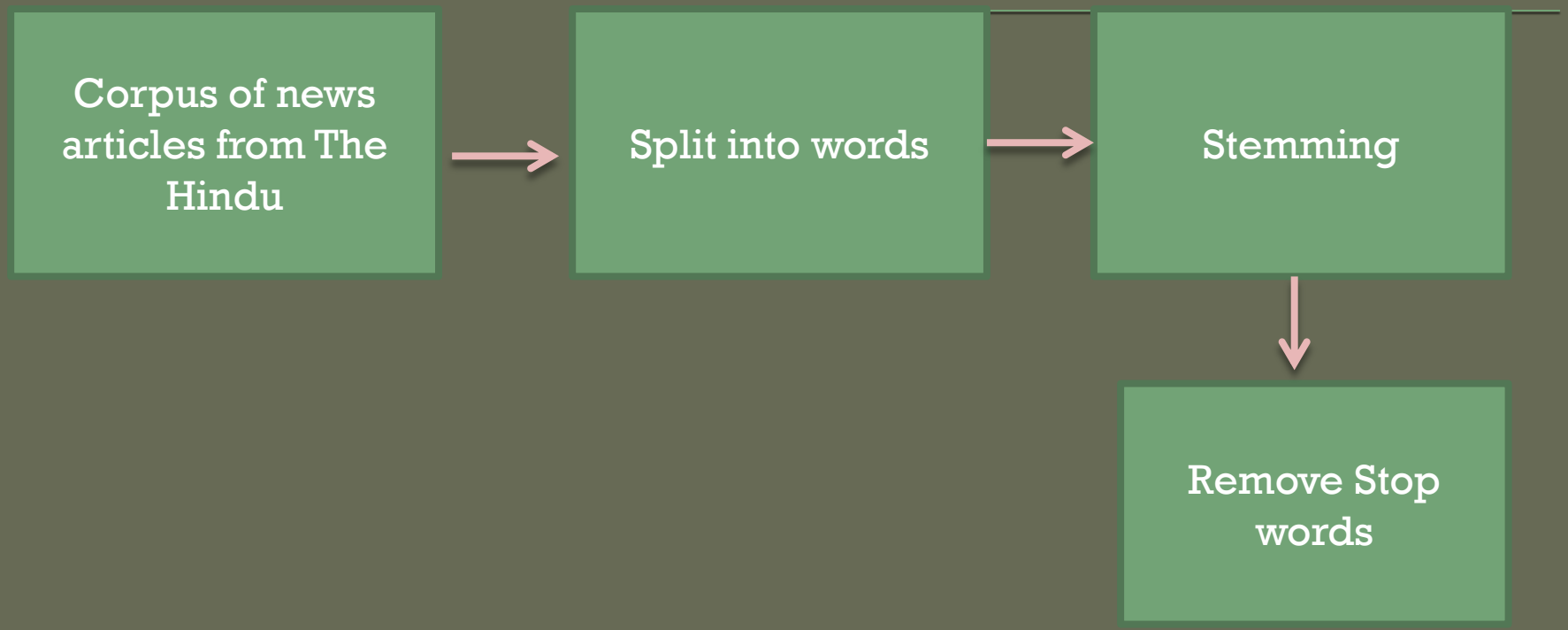
Framework



45

['a', 'delhi', 'court', 'on', 'wednesdai', 'convict', 'sukhdev', 'pehalwan,', 'the', 'third', 'accus', 'in', 'the', '2002', 'nitish', 'katara', 'murder', 'case,', 'sai', 'that', 'at', 'the', 'time', 'of', 'the', 'incid', 'he', 'too', 'wa', 'present', 'with', 'convict', 'vika', 'yadav', 'and', 'vishal', 'yadav', 'current', 'serv', 'life', 'term', 'in', 'tihar', 'jail']

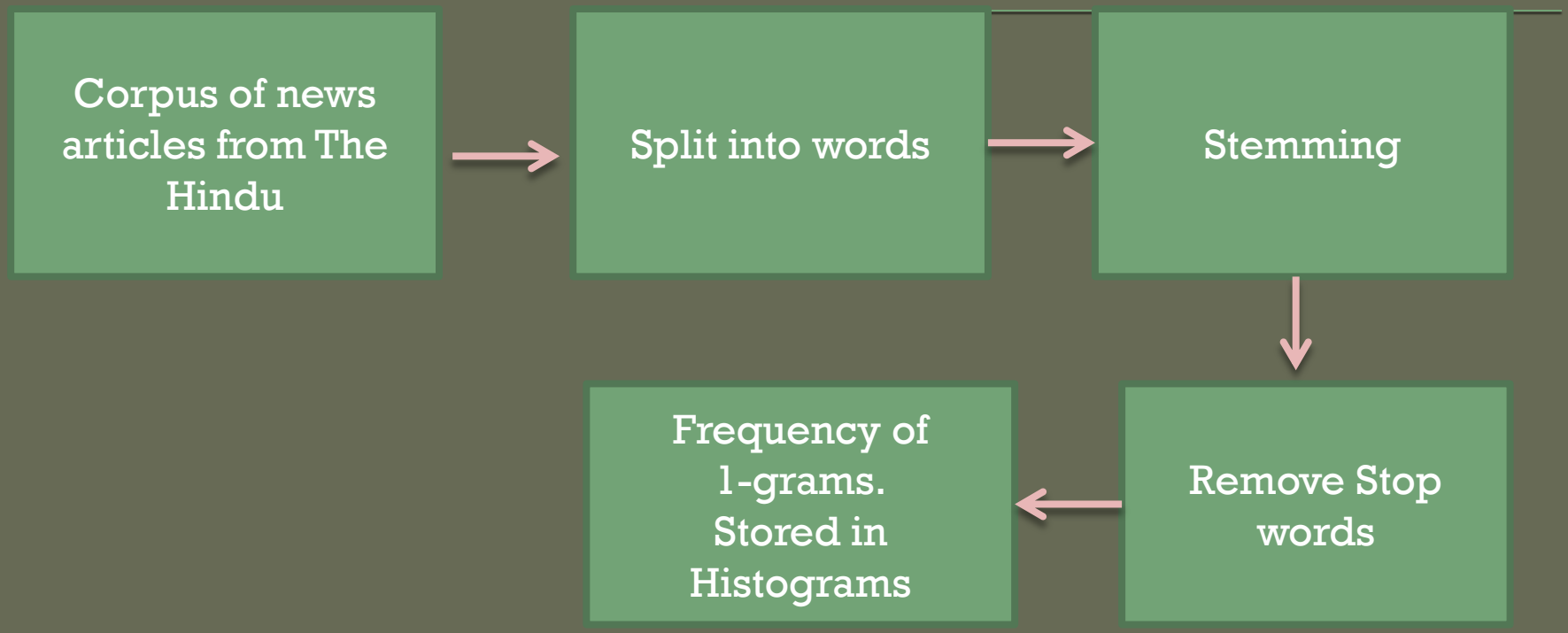
Framework



29

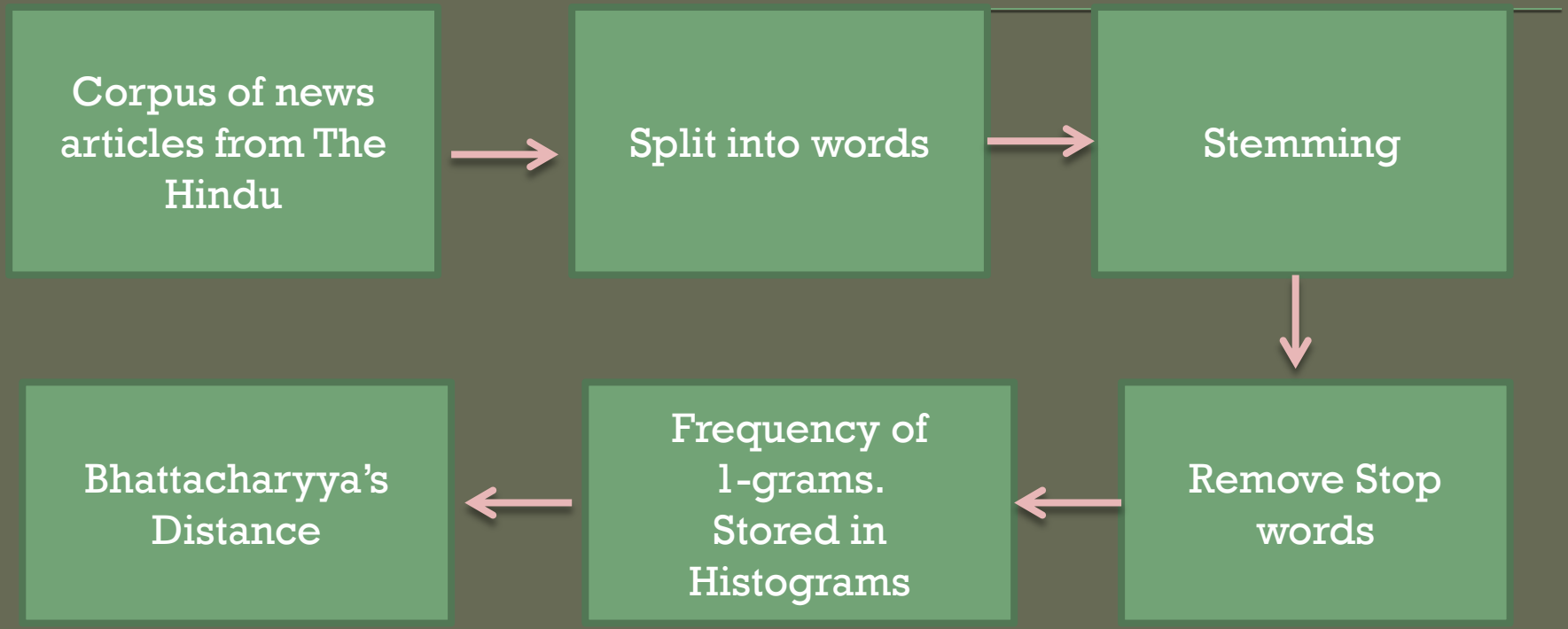
['delhi', 'court', 'wednesdai', 'convict', 'sukhdev', 'pehalwan,', 'third', 'accus', '2002', 'nitish', 'katara', 'murder', 'case,', 'sai', 'time', 'incid', 'wa', 'present', 'convict', 'vika', 'yadav', 'vishal', 'yadav', 'current', 'serv', 'life', 'term', 'tihar', 'jail']

Framework



[['delhi', 1], ['court', 1], ['wednesdai', 1], ['sukhdev', 1], ['pehalwan,', 1], ['third', 1], ['accus', 1], ['2002', 1], ['nitish', 1], ['katara', 1], ['murder', 1], ['case,', 1], ['sai', 1], ['time', 1], ['incid', 1], ['wa', 1], ['present', 1], ['vika', 1], ['vishal', 1], ['current', 1], ['serv', 1], ['life', 1], ['term', 1], ['tihar', 1], ['jail', 1], ['yadav', 2], ['convict', 2]]

Framework



Bhattacharyya's Distance

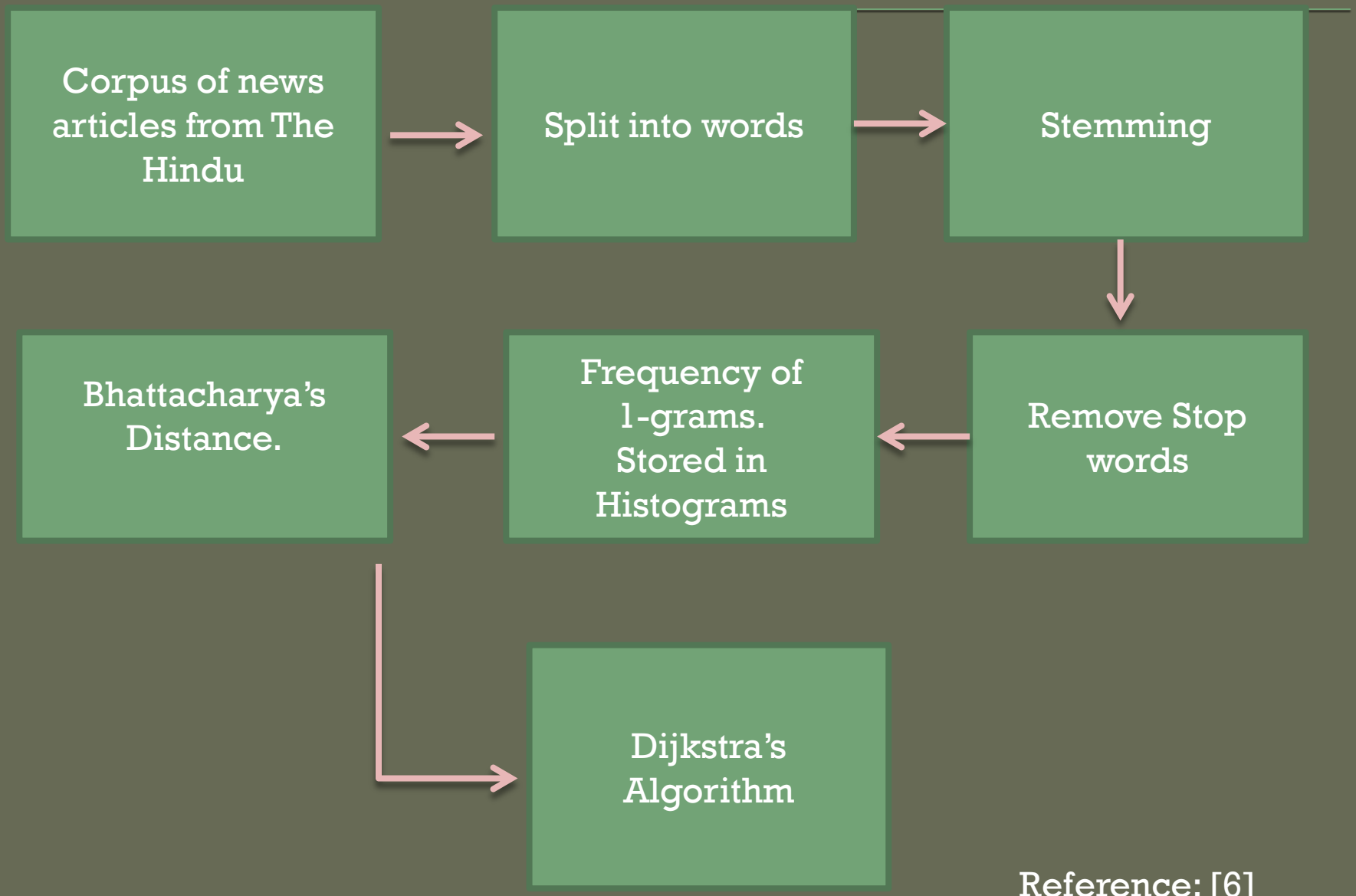
DB = - ln (BC(p,q)):

where

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

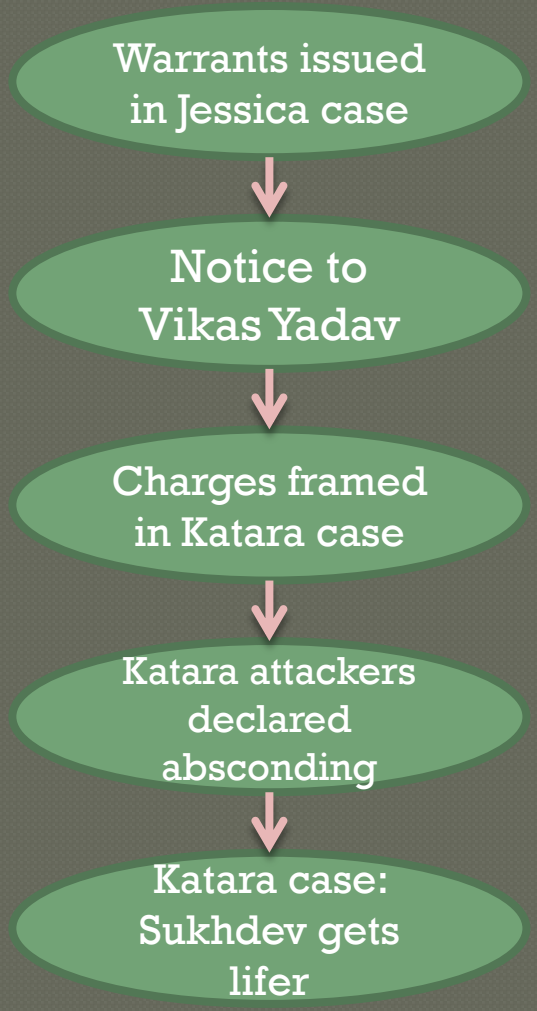
Reference: [7]

Framework



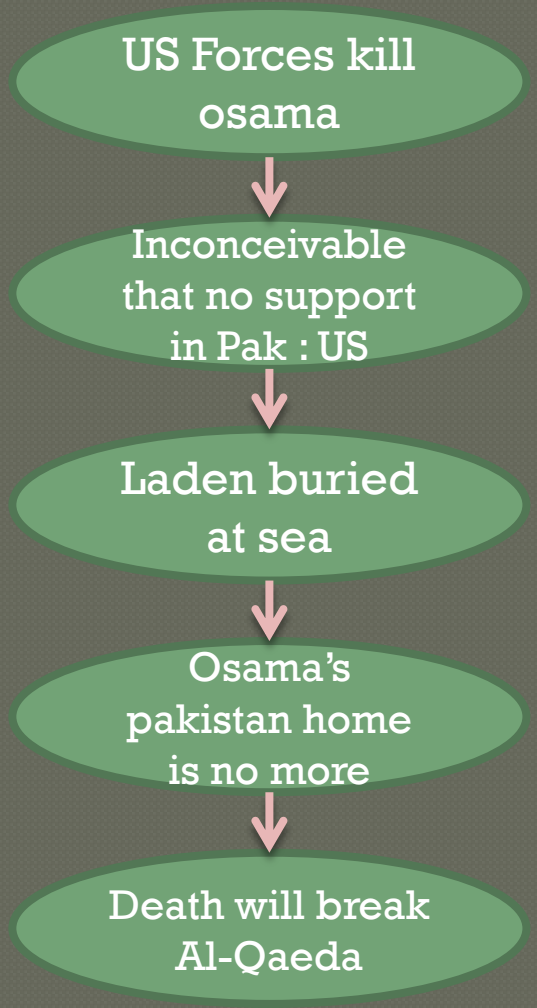
Reference: [6]

Sample Results




```
>>>

LINKING ARTICLES INDEX

221
233
232
246
247

LINK TOPICS

221_Warrants issued in Jessica case.txt
233_Notice to Vikas Yadav.txt
232_Charges framed against third accused in Katara case.txt
246_Katara attackers declared absconding.txt
247_Katara murder_Sukhdev Pehalwan gets lifer.txt
>>>
```

Sample Results




```
>>>

LINKING ARTICLES INDEX

345
343
330
349
342

LINK TOPICS

345_US FORCES KILL OSAMA BIN LADEN.txt
343_Inconceivable that Osama had no support system in Pakistan_ US.txt
330_Osama bin Laden buried at sea.txt
349_Osamas Pakistan home is no more.txt
342_His death will break the iron fist of al_Qaeda in Iraq.txt
```

Evaluation

- Coherence $(d_1, \ldots, d_n) = {}^{n-1}\Sigma_{i=1}\ \Sigma_w\ 1(w \in d_i \cap d_{i+1})$
  Every time a word appears in two consecutive articles, we score a point
  Drawback : Weak links

- Coherence $(d_1, \ldots, d_n) = {}_{i=1\ldots n-1}\min\ \Sigma_w\ 1(w \in d_i \cap d_{i+1})$
  Minimal transition score

Reference: [1]

```
=[]

 in range(len(sentences)):
istances.append([])
ength_tempi = sum(sentences[i][j][1] for j in range(len(sentences[i])))
ount=0

or j in range(len(sentences)):
    count=0
    length_tempj = sum(sentences[j][m][1] for m in range(len(sentences[j])))

    if len(sentences[i]) < len(sentences[j]):
        for k in range(len(sentences[i])):
                for l in range(len(sentences[j])):
                        if sentences[i][k][0] == sentences[j][l][0] and sentences[i][k][0]!='':
                            count = round(count + round(math.sqrt(float(sentences[i][k][1]*sentences[j][l][1])/(length_tempi * length_tempj)),5),
    else:
        for k in range(len(sentences[j])):
                for l in range(len(sentences[i])):
                        if sentences[j][k][0] == sentences[i][l][0] and sentences[j][k][0]!='':
                            count = round(count + round(math.sqrt(float(sentences[j][k][1]*sentences[i][l][1])/(length_tempi * length_tempj)),5),
    if count == 0.0:
        distances[i].append(0.0)
    else:
        distances[i].append(round(-math.log(count),2))
t distances
t target_distance

n("distances.txt","w")
 dump(distances,f)
se()


n("distances.txt","r")
nces=json.load(f)
se()
```

Code Snapshot

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] Dafna Shahaf , Prof. Carlos Guestrin and Eric Horvitz : Trains of thought-Generating information maps. *International World Wide Web Conference (WWW), 2012.*
- [3] Michael D. Lee, Brandon Pincombe and Matthew Welsh : An Empirical Evaluation of Models of Text Document Similarity. In Proceedings of the 27th Annual Conference of the Cognitive Science Society (2005).
- [4] Deept Kumar, Naren Ramakrishnan, Richard F. Helm, and Malcolm Potts : Algorithms for Storytelling. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 6, JUNE 2008
- [5] M. Shahriar Hossain, Joseph Gresock, Yvette Edmonds, Richard Helm, Malcolm Potts and Naren Ramakrishnan. Connecting the Dots between PubMed Abstracts. 2012
- [6]http://networkx.github.com/documentation/latest/reference/generated/networkx.algorithms.shortest_paths.weighted.dijkstra_path.html#networkx.algorithms.shortest_paths.weighted.dijkstra_path
- [7] http://en.wikipedia.org/wiki/Bhattacharyya_distance

Thank you

Questions ?

Other Approaches

- [5] used *Soergel distance* to calculate distance between documents and then *A* algorithm* to find the chain
- [1] used bipartite graph and the notion of *influence* to find the chain
- [2] used notion of *m-coherence* for evaluation of results