Connecting the Dots Between News Articles

Guide : Amitabha Mukerjee
Ankit Modi (10104, amodi@)
Chirag Gupta (10212, gchirag@)

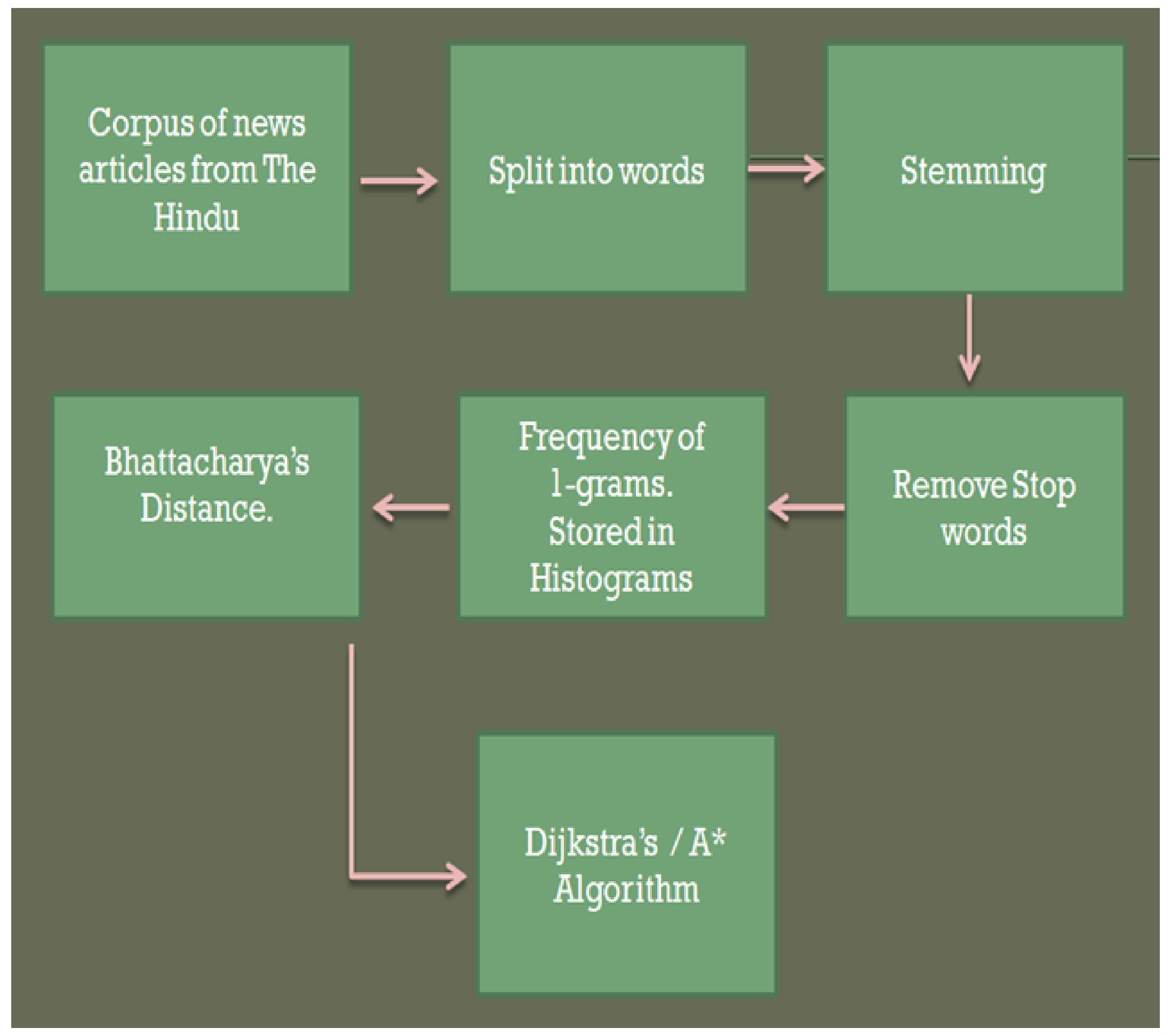
Motivation

- Tackling Information Overload
- Navigate Between Topics
- Seeing the bigger picture
- News Browsing : Primary use of Internet

Related work

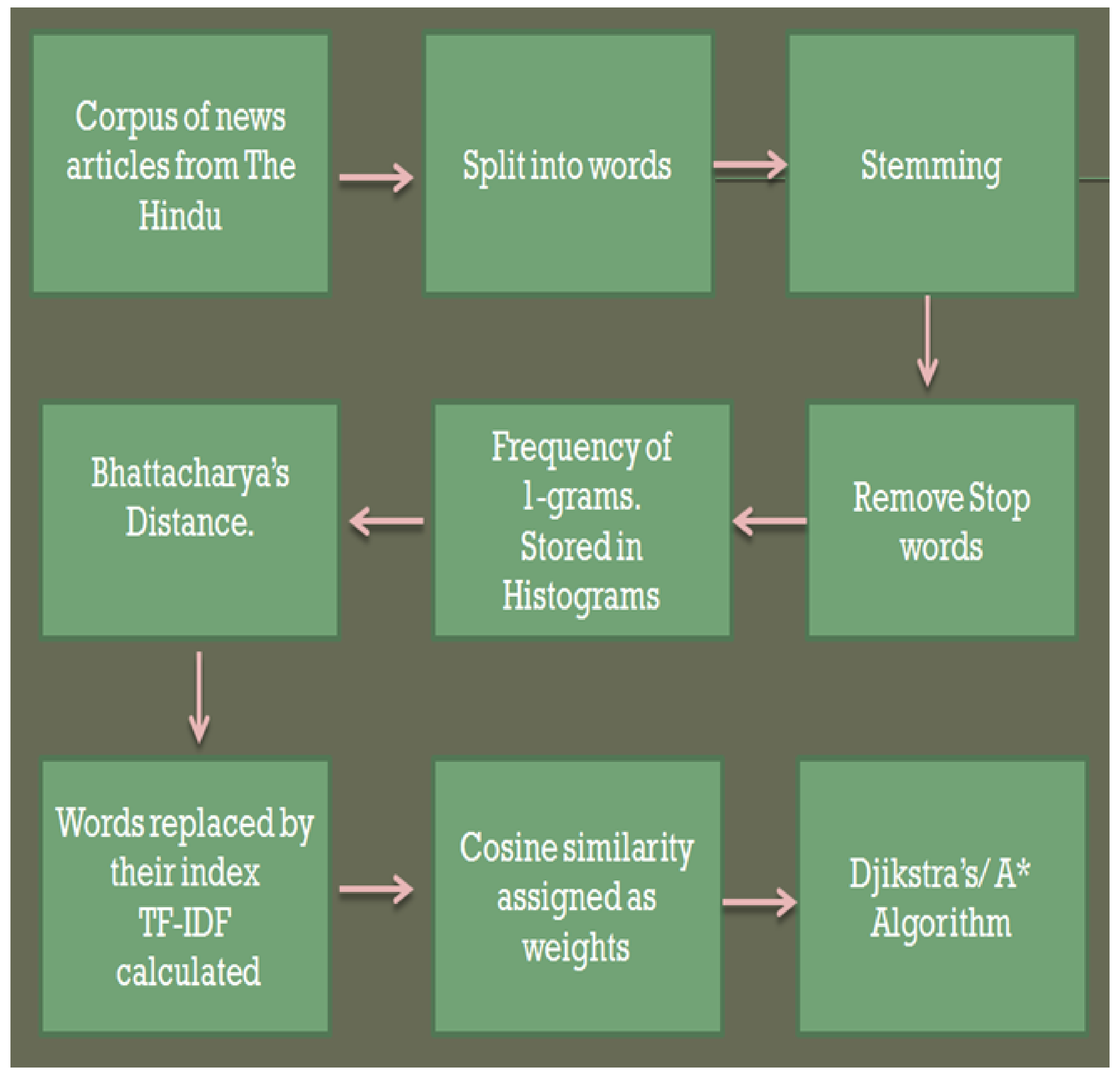
- Dafna Shahaf and Prof. Carlos Guestrin used Bipartite Graph and notion of influence (2010)
- Hossain, Gresock, Edmond, helm, potts and ramakrishnan used *Soergel Distance* and *A* algorithm* (2012)
- Shahaf, Guestrin and Horvitz used m-coherence to compare similarity between articles (2012)
- R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *CIKM '04, 2004*
- Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05, 2005*

Approach 1



Corpus of news articles from The Hindu → Split into words → Stemming → Remove Stop words → Frequency of 1-grams. Stored in Histograms → Bhattacharya's Distance. → Dijkstra's / A* Algorithm

Approach2



Corpus of news articles from The Hindu → Split into words → Stemming → Remove Stop words → Frequency of 1-grams. Stored in Histograms → Bhattacharya's Distance. → Words replaced by their index TF-IDF calculated → Cosine similarity assigned as weights → Dijkstra's/ A* Algorithm

Coherence

We evaluate our results using *coherence* as a parameter. We used two different methods to calculate *coherence*.

•**Coherence1**$(d_1, \dots, d_n) = {}_{i=1\dots n-1}\min \Sigma_w 1(w \in d_i \cap d_{i+1})$
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.*

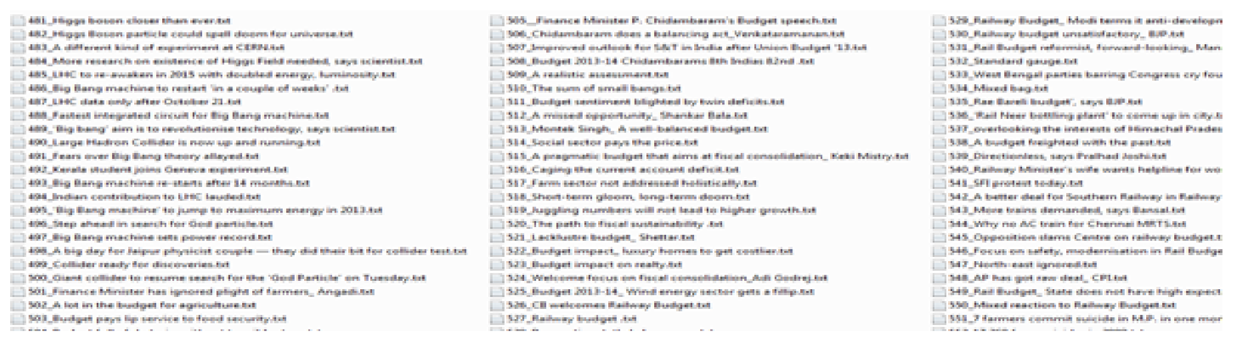
•**Coherence2**$(d_1, \dots, d_n) = {}_{i=1\dots n-1}\min \{ \text{cosine-similarity } (d_i , d_{i+1}) \}$
This measure guarantees a value between 0 and 1, and hence, gives the percentage similarity between documents.

Tools And Resources used

- **Gensim:** www.radimrehurek.com/gensim
- **NetworkX:** http://networkx.github.io/documentation/latest/index.html
- **matplotlib:** http://matplotlib.org/

Data Set

We have created our own corpus of 1000 articles downloaded from The Hindu spread over around 40 topics. Some of the topics are Nitish Katara murder case, Jessica Lal murder case, US presidential elections, India's world cup triumph etc.
We expect it to be very useful for future projects related to Data-Mining. Following is a snapshot of our data set

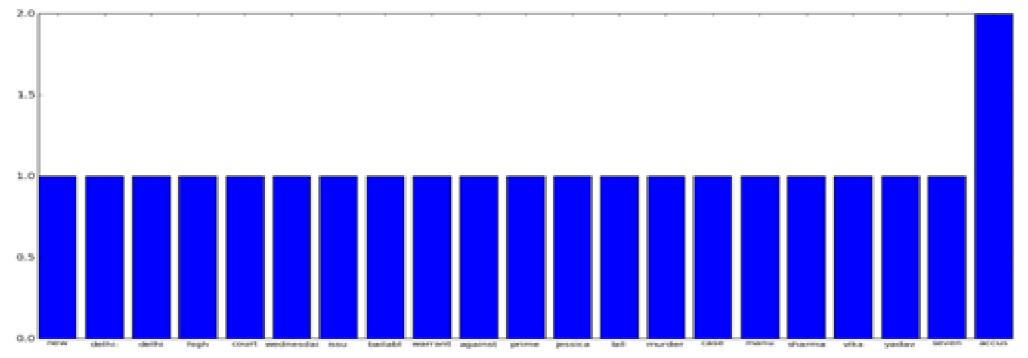


481_Higgs boson closer than ever.txt
482_Higgs Boson particle could spell doom for universe.txt
483_A different kind of experiment at CERN.txt
484_More research on existence of Higgs Field needed, says scientist.txt
485_LHC to re-awaken in 2015 with doubled energy, luminosity.txt
486_Big Bang machine to restart 'in a couple of weeks' .txt
487_LHC data only after October 21.txt
488_Fastest integrated circuit for Big Bang machine.txt
489_'Big bang' aim is to revolutionise technology, says scientist.txt
490_Large Hadron Collider is now up and running.txt
491_Fears over Big Bang theory allayed.txt
492_Kerala student joins Geneva experiment.txt
493_Big Bang machine re-starts after 14 months.txt
494_Indian contribution to LHC lauded.txt
495_'Big Bang machine' to jump to maximum energy in 2013.txt
496_Step ahead in search for God particle.txt
497_Big Bang machine sets power record.txt
498_A big day for Jaipur physicist couple — they did their bit for collider test.txt
499_Collider ready for discoveries.txt
500_Giant collider to resume search for the 'God Particle' on Tuesday.txt
501_Finance Minister has ignored plight of farmers_ Angadi.txt
502_A lot in the budget for agriculture.txt
503_Budget pays lip service to food security.txt
505__Finance Minister P. Chidambaram's Budget speech.txt
506_Chidambaram does a balancing act_Venkataramanan.txt
507_Improved outlook for S&T in India after Union Budget '13.txt
508_Budget 2013-14 Chidambarams 8th Indian 82nd .txt
509_A realistic assessment.txt
510_The sum of small bangs.txt
511_Budget sentiment blighted by twin deficits.txt
512_A missed opportunity_ Shankar Bala.txt
513_Montek Singh_ A well-balanced budget.txt
514_Social sector pays the price.txt
515_A pragmatic budget that aims at fiscal consolidation_ Keki Mistry.txt
516_Caging the current account deficit.txt
517_Farm sector not addressed holistically.txt
518_Short-term gloom, long-term doom.txt
519_Juggling numbers will not lead to higher growth.txt
520_The path to fiscal sustainability .txt
521_Lacklustre budget_ Shettar.txt
522_Budget impact_ luxury homes to get costlier.txt
523_Budget impact on realty.txt
524_Welcome focus on fiscal consolidation_Adi Godrej.txt
525_Budget 2013-14_ Wind energy sector gets a fillip.txt
526_CB welcomes Railway Budget.txt
527_Railway budget .txt
529_Railway Budget_ Modi terms it anti-developn
530_Railway budget unsatisfactory_ BJP.txt
531_Rail Budget reformist, forward-looking_ Man
532_Standard gauge.txt
533_West Bengal parties barring Congress cry fou
534_Mixed bag.txt
535_Rae Bareli budget', says BJP.txt
536_'Rail Neer bottling plant' to come up in city.t
537_overlooking the interests of Himachal Prades
538_A budget freighted with the past.txt
539_Directionless, says Pralhad Joshi.txt
540_Railway Minister's wife wants helpline for wo
541_SFI protest today.txt
542_A better deal for Southern Railway in Railway
543_More trains demanded, says Bansal.txt
544_Why no AC train for Chennai MRTS.txt
545_Opposition slams Centre on railway budget.t
546_Focus on safety, modernisation in Rail Budge
547_North-east ignored.txt
548_AP has got raw deal_ CPI.txt
549_Rail Budget_ State does not have high expect
550_Mixed reaction to Railway Budget.txt
551_7 farmers commit suicide in M.P. in one mor

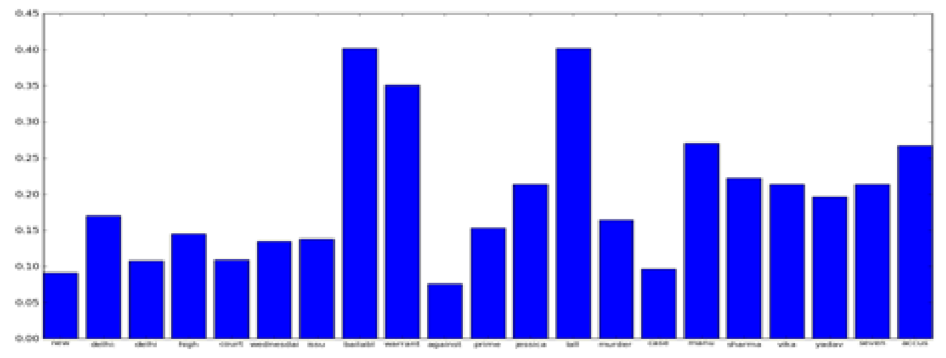
Results

Weight	Algorithm	Chain	Coherence1	Coherence2
Bhatt. Dist	Dijkstra	123-117-105-113-111	11 (105-113)	16.05% (105-113)
Cosine-Similarity	Dijkstra	123-103-102-110-111	16 (110-111)	22.88% (110-111)
Bhatt. Dist	A*	123-118-103-117-111	18 (117-111)	19.24% (118-103)
Cosine-similarity	A*	123-103-117-110-111	16 (110-111)	20.45% (117-110)

Comparison between Word-count and TF-IDF

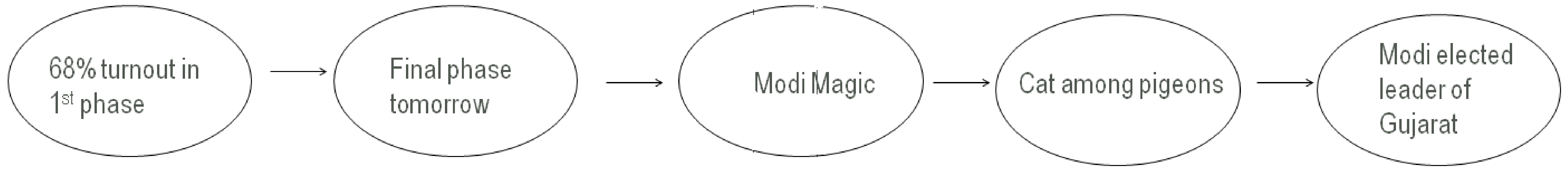


unigram

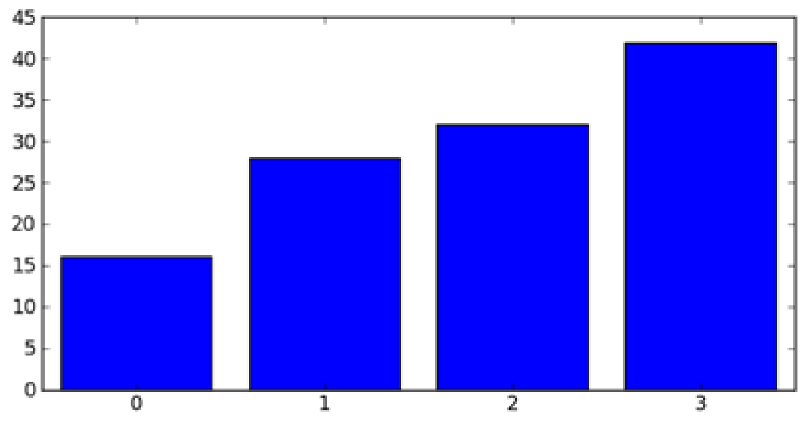


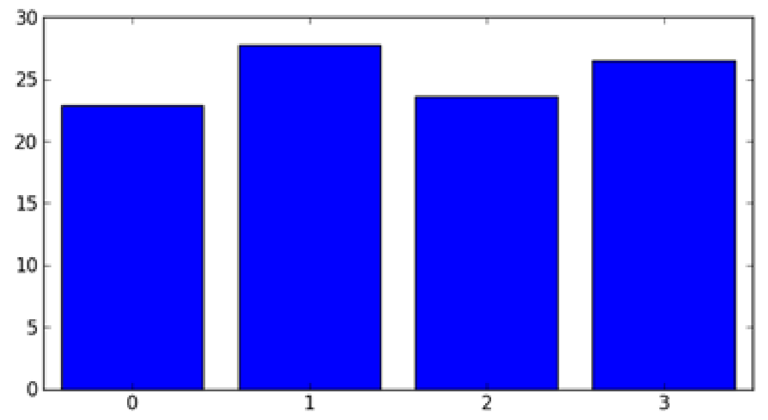
TF-IDF

Comparison between *Coherence1* and *Coherence2*

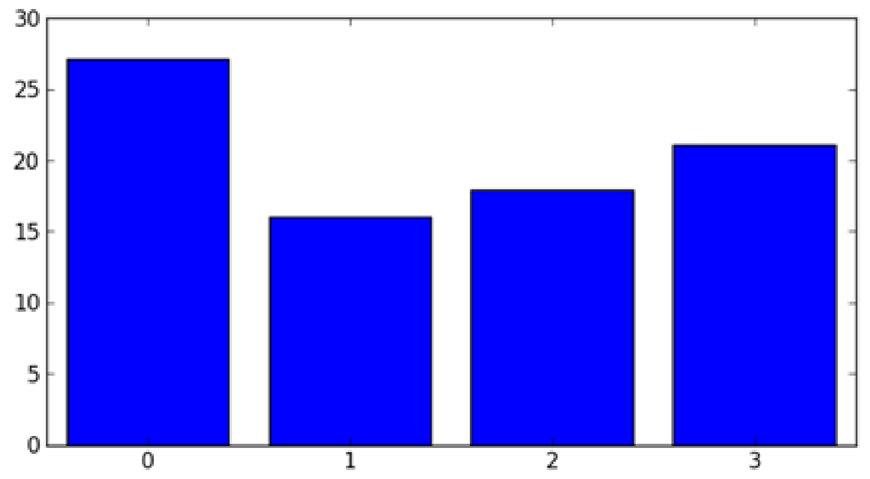


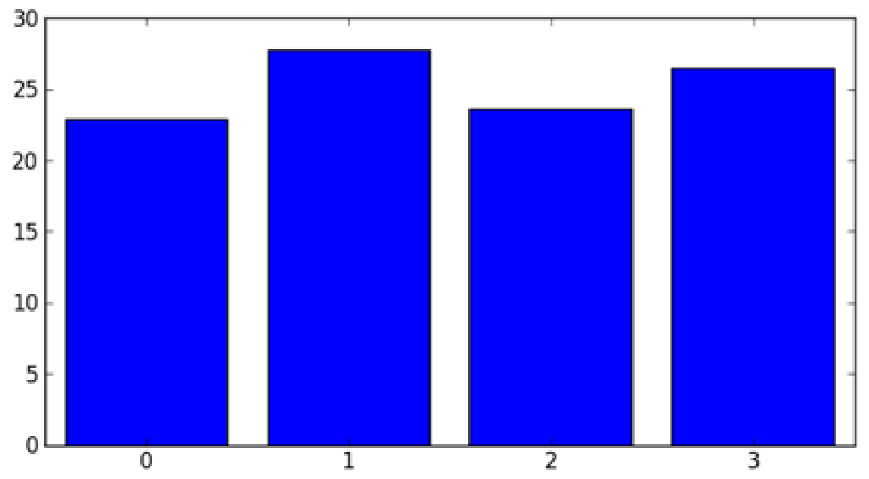
68% turnout in 1st phase → Final phase tomorrow → Modi Magic → Cat among pigeons → Modi elected leader of Gujarat





Comparison using *coherence2* between *Bhattacharya's Distance* and *Cosine Similarity*





Conclusion

- TF-IDF is better than uniram
- *Coherence2* is a better evaluating parameter than *cohorence1*
- Cosine similarity using TF-IDF gives more cohorent chain than Bhattacharaya's distance.

Future Work

- Notion of *m-coherence* can be used instead of *coherence* for better evaluation of results.
  Ref: Dafna Shahaf , Prof. Carlos Guestrin and Eric Horvitz : Trains of thought-Generating information maps. *International World Wide Web Conference (WWW), 2012*
- Using different corpuses (other than News articles) may help in important scientific discoveries.

www.PosterPresentations.com