Topic Transition Modeling

Computing for Data Science

PGDBA Group 11

14/11/2015

"All models are false, some are useful"

-George E. P. Box

How it started …

- The initial idea of the Topic for the project came from a very random conversation on a bus and we decided to model how these conversations flowed.
- A lose analogy(term) used was to model it on the lines of Random Walk“ .. Something like a “Random Talk” ☺

Content

- Objective
- Model assumptions
- Steps
  1. Data Collection
  2. Data Cleaning
  3. Topic Modeling
  4. Topic Transition Modeling
  5. Visualization
- Conclusion
- Challenges
- References

Objective

1. To identify the latent topics that underlie a conversation and
2. study their distribution across the entire timestamp to arrive at stationary probabilities

The probability distributions obtained can then be modeled to understand the transitions of the conversation from one topic to another. To achieve the objective, certain assumptions were made.

Model assumptions

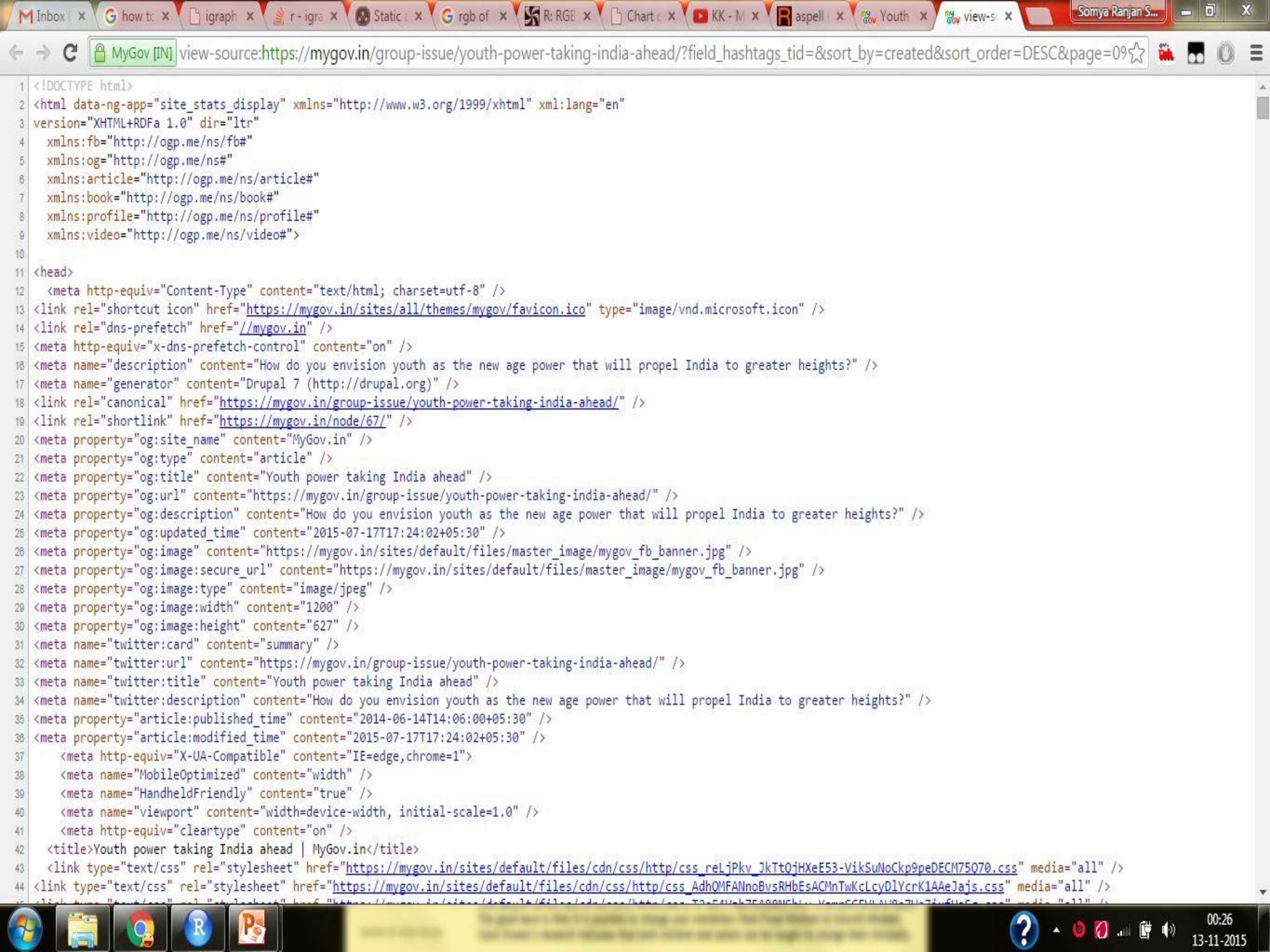
- Topic transition is a homogeneous Markov chain process
- A set of consecutive comments is a document
- For topic modeling
    1. Order of words in a document is neglected (i.e., exchangeable)
    2. Order of documents in a corpus is also neglected

"In God we trust. All others must bring data."

-W. Edwards Deming

Data Collection

- Data was scraped from a national discussion forum https://mygov.in
- Crawlers were build using Java program
- Data consists of comments of people from across the country
- Comments were made in a time stamp of one year

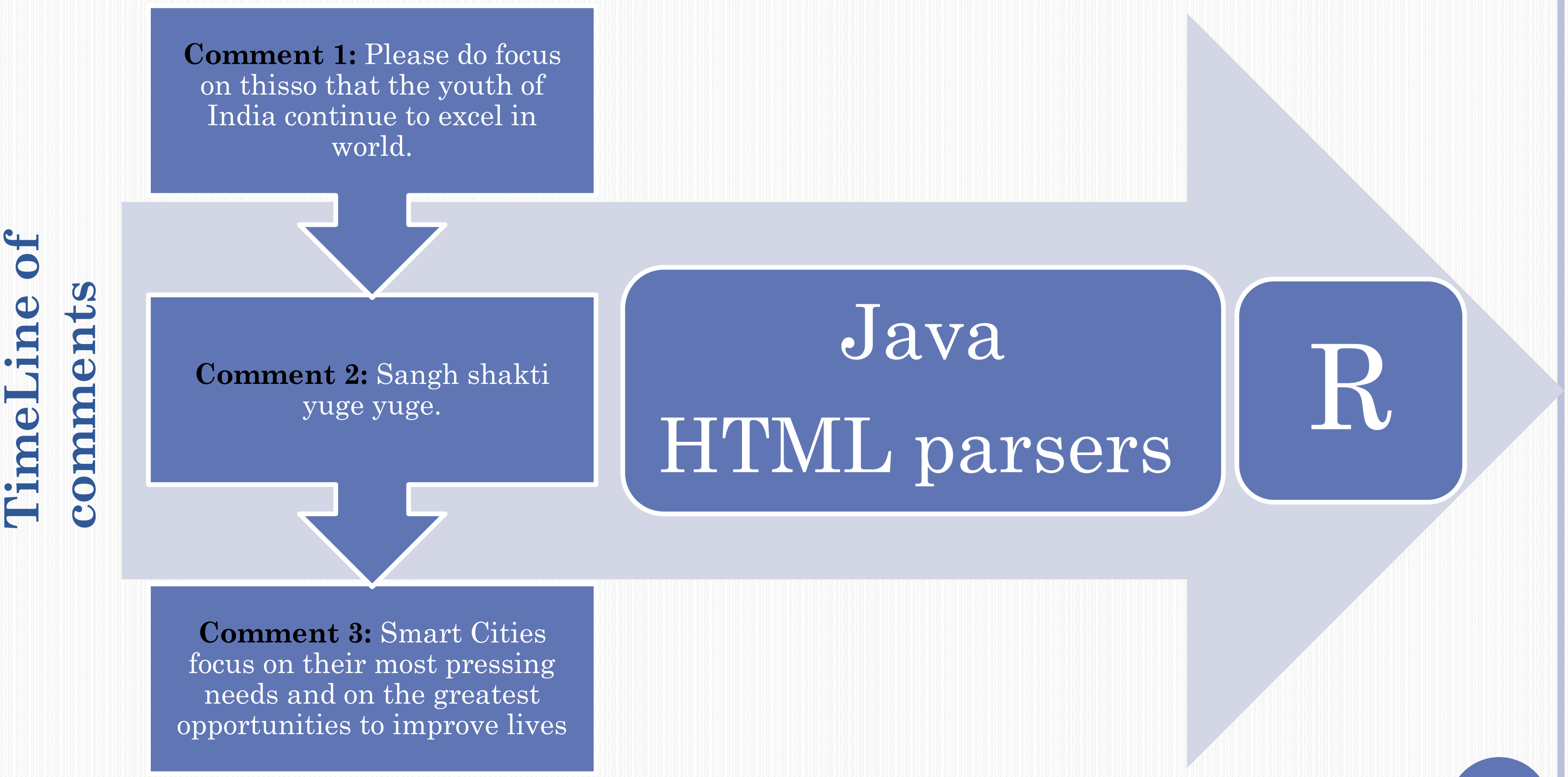


Inbox
how to
igraph
r - igra
Static
rgb of
R: RGB
Chart
KK - M
aspell
Youth
view-s
Somya Ranjan S...
MyGov [IN]
view-source:https://mygov.in/group-issue/youth-power-taking-india-ahead/?field_hashtags_tid=&sort_by=created&sort_order=DESC&page=0%
00:26
13-11-2015

```
<!DOCTYPE html>
<html data-ng-app="site_stats_display" xmlns="http://www.w3.org/1999/xhtml" xml:lang="en"
version="XHTML+RDFa 1.0" dir="ltr"
  xmlns:fb="http://ogp.me/ns/fb#"
  xmlns:og="http://ogp.me/ns#"
  xmlns:article="http://ogp.me/ns/article#"
  xmlns:book="http://ogp.me/ns/book#"
  xmlns:profile="http://ogp.me/ns/profile#"
  xmlns:video="http://ogp.me/ns/video#">

<head>
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<link rel="shortcut icon" href="https://mygov.in/sites/all/themes/mygov/favicon.ico" type="image/vnd.microsoft.icon" />
<link rel="dns-prefetch" href="//mygov.in" />
<meta http-equiv="x-dns-prefetch-control" content="on" />
<meta name="description" content="How do you envision youth as the new age power that will propel India to greater heights?" />
<meta name="generator" content="Drupal 7 (http://drupal.org)" />
<link rel="canonical" href="https://mygov.in/group-issue/youth-power-taking-india-ahead/" />
<link rel="shortlink" href="https://mygov.in/node/67/" />
<meta property="og:site_name" content="MyGov.in" />
<meta property="og:type" content="article" />
<meta property="og:title" content="Youth power taking India ahead" />
<meta property="og:url" content="https://mygov.in/group-issue/youth-power-taking-india-ahead/" />
<meta property="og:description" content="How do you envision youth as the new age power that will propel India to greater heights?" />
<meta property="og:updated_time" content="2015-07-17T17:24:02+05:30" />
<meta property="og:image" content="https://mygov.in/sites/default/files/master_image/mygov_fb_banner.jpg" />
<meta property="og:image:secure_url" content="https://mygov.in/sites/default/files/master_image/mygov_fb_banner.jpg" />
<meta property="og:image:type" content="image/jpeg" />
<meta property="og:image:width" content="1200" />
<meta property="og:image:height" content="627" />
<meta name="twitter:card" content="summary" />
<meta name="twitter:url" content="https://mygov.in/group-issue/youth-power-taking-india-ahead/" />
<meta name="twitter:title" content="Youth power taking India ahead" />
<meta name="twitter:description" content="How do you envision youth as the new age power that will propel India to greater heights?" />
<meta property="article:published_time" content="2014-06-14T14:06:00+05:30" />
<meta property="article:modified_time" content="2015-07-17T17:24:02+05:30" />
    <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
    <meta name="MobileOptimized" content="width" />
    <meta name="HandheldFriendly" content="true" />
    <meta name="viewport" content="width=device-width, initial-scale=1.0" />
    <meta http-equiv="cleartype" content="on" />
  <title>Youth power taking India ahead | MyGov.in</title>
  <link type="text/css" rel="stylesheet" href="https://mygov.in/sites/default/files/cdn/css/http/css_reLjPkv_JkTtQjHXeE53-VikSuNoCkp9peDECM75Q70.css" media="all" />
<link type="text/css" rel="stylesheet" href="https://mygov.in/sites/default/files/cdn/css/http/css_AdhQMFANnoBvsRHbEsACMnTwKcLcyDlYcrK1AAeJajs.css" media="all" />
```

Flow Chart



TimeLine of comments

Comment 1: Please do focus on thisso that the youth of India continue to excel in world.

Comment 2: Sangh shakti yuge yuge.

Comment 3: Smart Cities focus on their most pressing needs and on the greatest opportunities to improve lives

Java

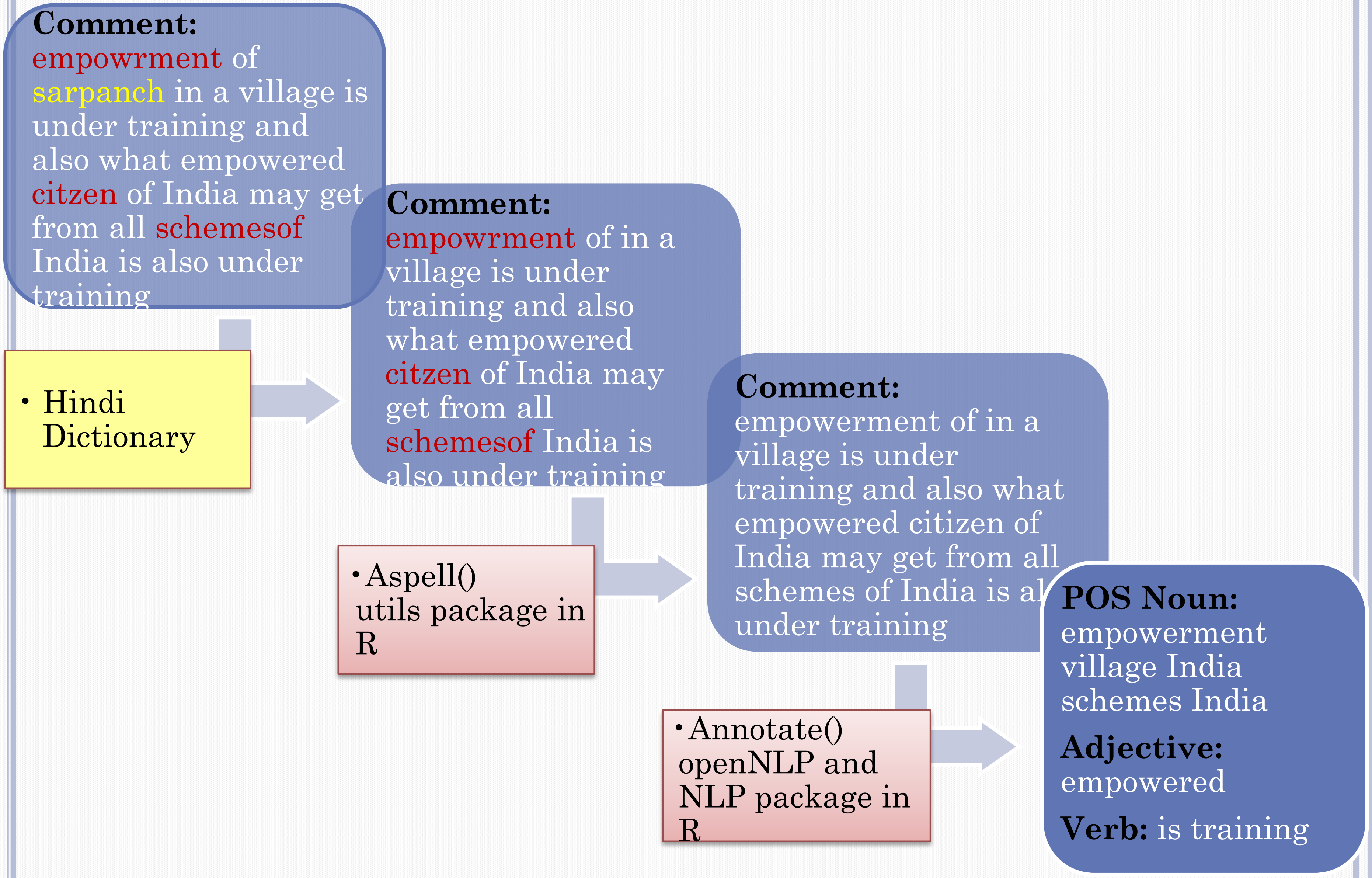
HTML parsers

R

Data Cleaning

- Removal of Hindi words using a self prepared dictionary
- Correction of misspelt English words using aspell function provided by utils package in R
- Extraction of different parts of speech (POS) using annotate function provided by NLP and open NLP package in R

Flow Chart



Comment: empowrment of sarpanch in a village is under training and also what empowered citzen of India may get from all schemesof India is also under training

- Hindi Dictionary

Comment: empowrment of in a village is under training and also what empowered citzen of India may get from all schemesof India is also under training

- Aspell() utils package in R

Comment: empowerment of in a village is under training and also what empowered citizen of India may get from all schemes of India is al under training

- Annotate() openNLP and NLP package in R

POS Noun: empowerment village India schemes India

Adjective: empowered

Verb: is training

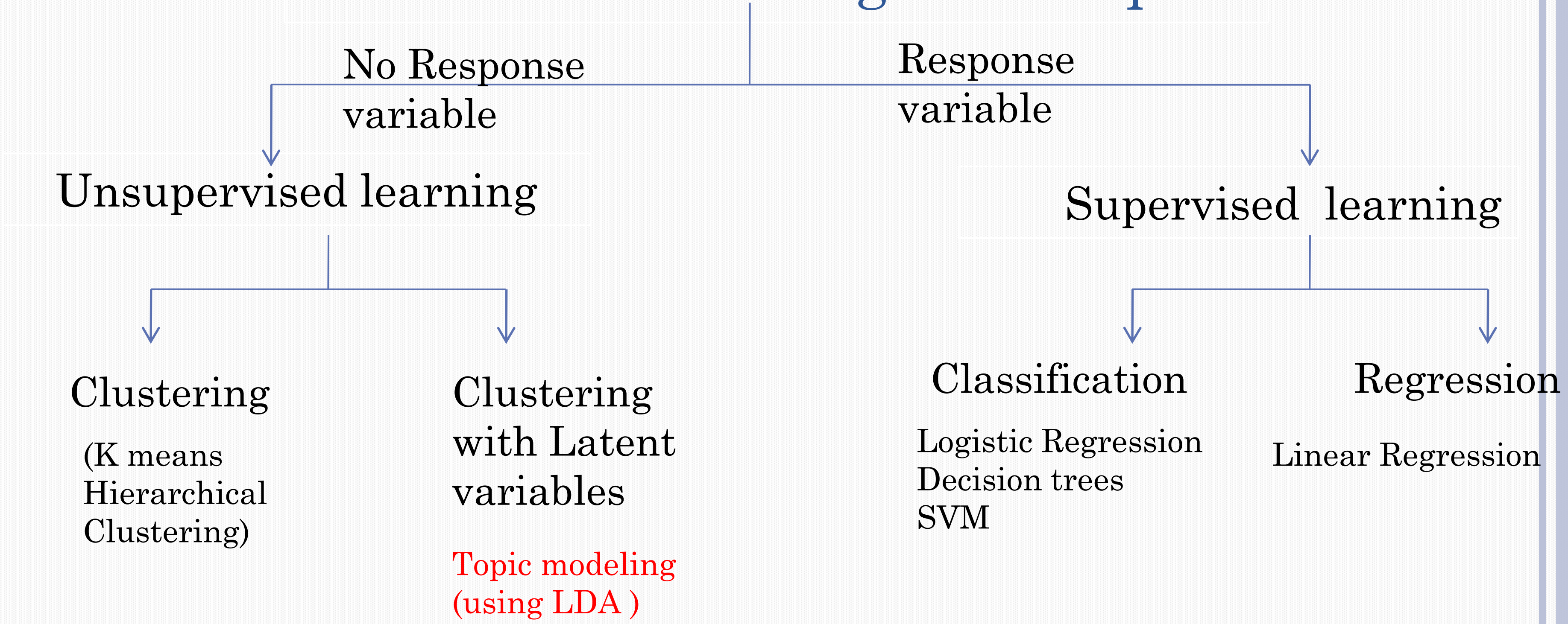
OUTPUT AFTER DATA CLEANING

Document	Nouns + verbs + Adjectives
1	representation Parliament youth bulk population way voice .schools colleges youth participation country decision
2	heritage city developments heritage contrast Smart city project people nature zones Govt
3	scheme Mahatma Gandhi guarantee scheme India employee contract basis contract employee basis trailer future
4	passport office court bench vigilance court city Bhopal city power system city offices institutions

"All generalizations are false, including this one."

-Mark Twain

Machine Learning Techniques



No Response variable

Response variable

Unsupervised learning

Supervised learning

Clustering

(K means
Hierarchical
Clustering)

Clustering with Latent variables

Topic modeling
(using LDA)

Classification

Logistic Regression
Decision trees
SVM

Regression

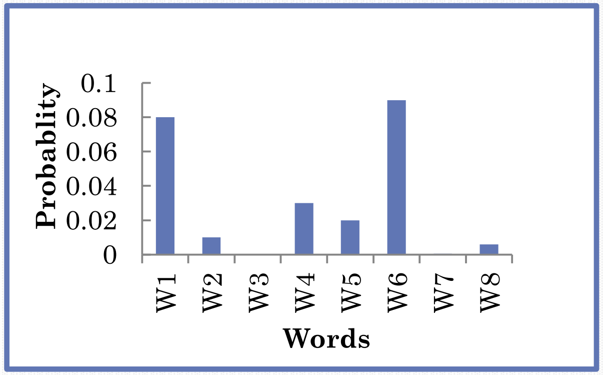
Linear Regression

Latent Dirichlet allocation (LDA) is a generative model that aim to explain observed variables in terms of latent variables (topics). Using latent variables, it describes why some of the data is similar.

http://chdoig.github.io/pygotham-topic-modeling/#/3/1

LATENT DIRICHLET ALLOCATION(LDA)

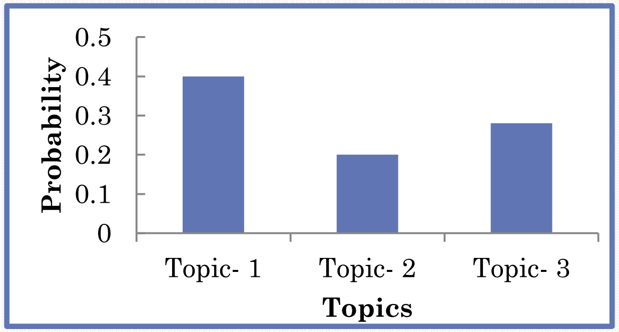
Topic →a probability distribution over all words in the corpus.



Similar distributions over words will be there for all the topics

Discover patterns of words-use…. …. …..topic

Document → a probability distribution over all topics



Similar distributions over topics will be there for all the documents

Connect documents that exhibit similar patterns.. ..Using proportion of topics (as a mixture)

What is known? The prior probabilities of words in a document

OBJECTIVE – To find the above distributions by providing a prior number of topics

LDA CONTD…

To describe corpus through LDA, following information required:

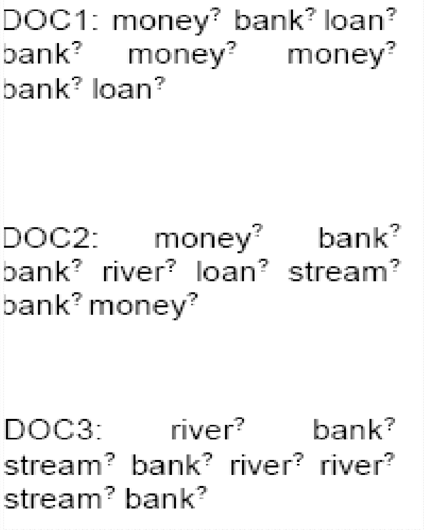
OBJECTIVE

- Distribution of words in a topic
- Distribution of topics in a document
- **z(i,j)** Topic assignment for w(i,j)

- Other information
  - **α – Dirichlet prior parameter** (sampled once for given data).
  - **β - Dirichlet prior parameter** (sampled once for given data)
  - No.of topics, documents and words …

HOW LDA WORKS?

Documents



DOC1: money? bank? loan?
bank? money? money?
bank? loan?

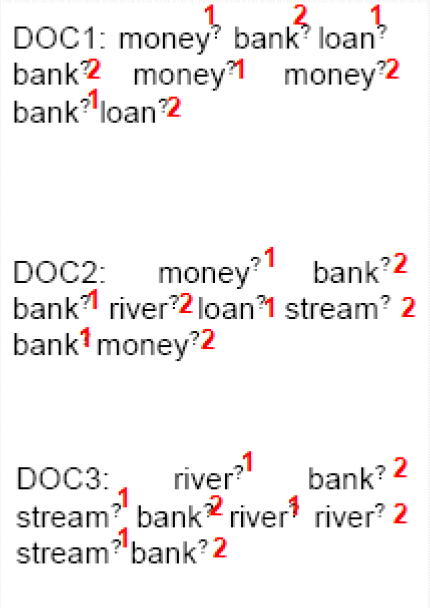
DOC2: money? bank?
bank? river? loan? stream?
bank? money?

DOC3: river? bank?
stream? bank? river? river?
stream? bank?

Topics

Topic -1
Topic -2

Randomly assign topics to each word



DOC1: money?1 bank?2 loan?1
bank?2 money?1 money?2
bank?1 loan?2

DOC2: money?1 bank?2
bank?1 river?2 loan?1 stream? 2
bank1 money?2

DOC3: river?1 bank? 2
stream?1 bank?2 river1 river? 2
stream?1 bank? 2

Count

C^{DT} → Topics per document

C^{WT} → Words per topic

	topic1	topic2
money	3	2
bank	3	6
Loan	2	1
River	2	2
Stream	2	1

	doc1	doc2	doc3
topic1	4	4	4
topic2	4	4	4

Factors affecting topic assignments

How likely is a word w for a topic j?
Probability of word w under topic *j*

How dominante is a topic j in a doc d?
Probability that topic *j* document *d*

HOW LDA WORKS? CONTD…

P(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w

Ex: P("money"/topic 1) = 3/12 =0.25

P("money"/topic 2) = 2/12 = 0.16

P(topic t | document d) = the proportion of words in document d that are currently assigned to topic t.

Ex: P(topic1/ doc1) = 4/8=0.5

P(topic2/doc1) = 4/8= 0.5

Prob **("word i" has index topic zi** /Given that "word i" is in doc d1, and all other assignments are known)

Proportional to

P(word wi /topic zi) * P(topic zi /doc1)

For first word "money" in doc1,

Prob that "money" belongs to topic 1 = 0.25*0.5 =0.125

Prob that "money" belongs to topic 2 = 0.5*0.5 =0.08

So, Index is changed for this word, and tables are updated with new indexes.

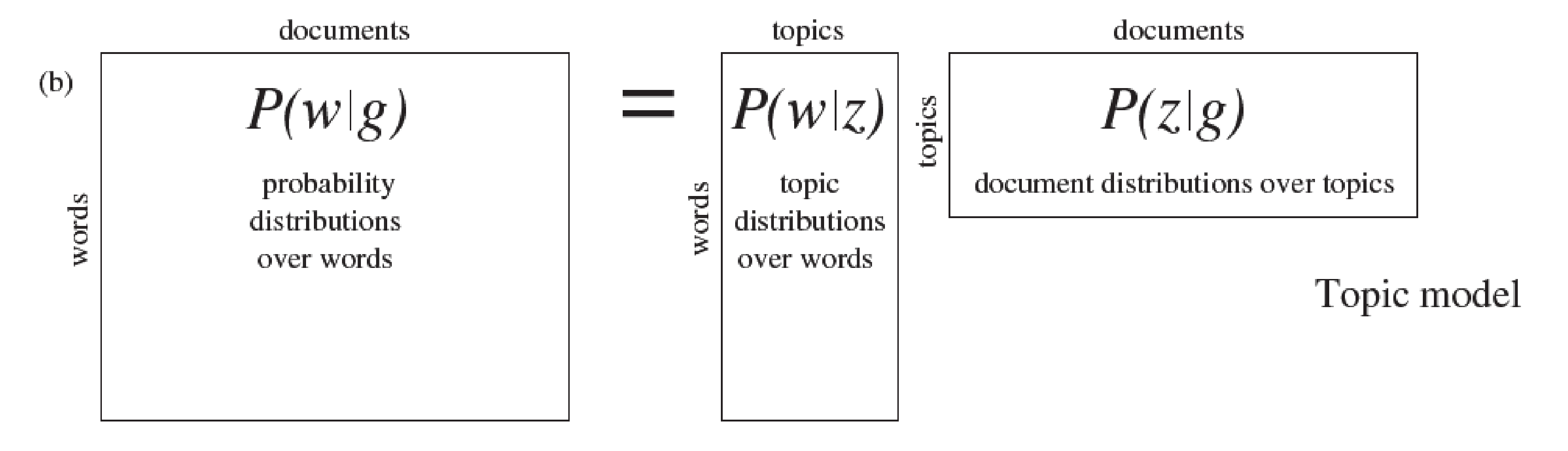
Gibbs Sampling for LDA

Probability that topic ***j*** is chosen for word ***w***

Count number of times a word token $\boldsymbol{w_i}$ was assigned to a topic ***j*** across all docs

$$P(z_i = j \mid \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{\boxed{C^{WT}_{w_i j}} + \beta}{\sum_{w=1}^{W} \boxed{C^{WT}_{wj}} + W\beta} \frac{\boxed{C^{DT}_{d_i j}} + \alpha}{\sum_{t=1}^{T} \boxed{C^{DT}_{d_i t}} + T\alpha}$$

Count number of times a topic ***j*** was already assigned to some word token in doc $\boldsymbol{d_i}$


documents
(b)
words
$P(w|g)$
probability distributions over words
=
topics
words
$P(w|z)$
topic distributions over words
topics
documents
$P(z|g)$
document distributions over topics
Topic model

Latent variable model:

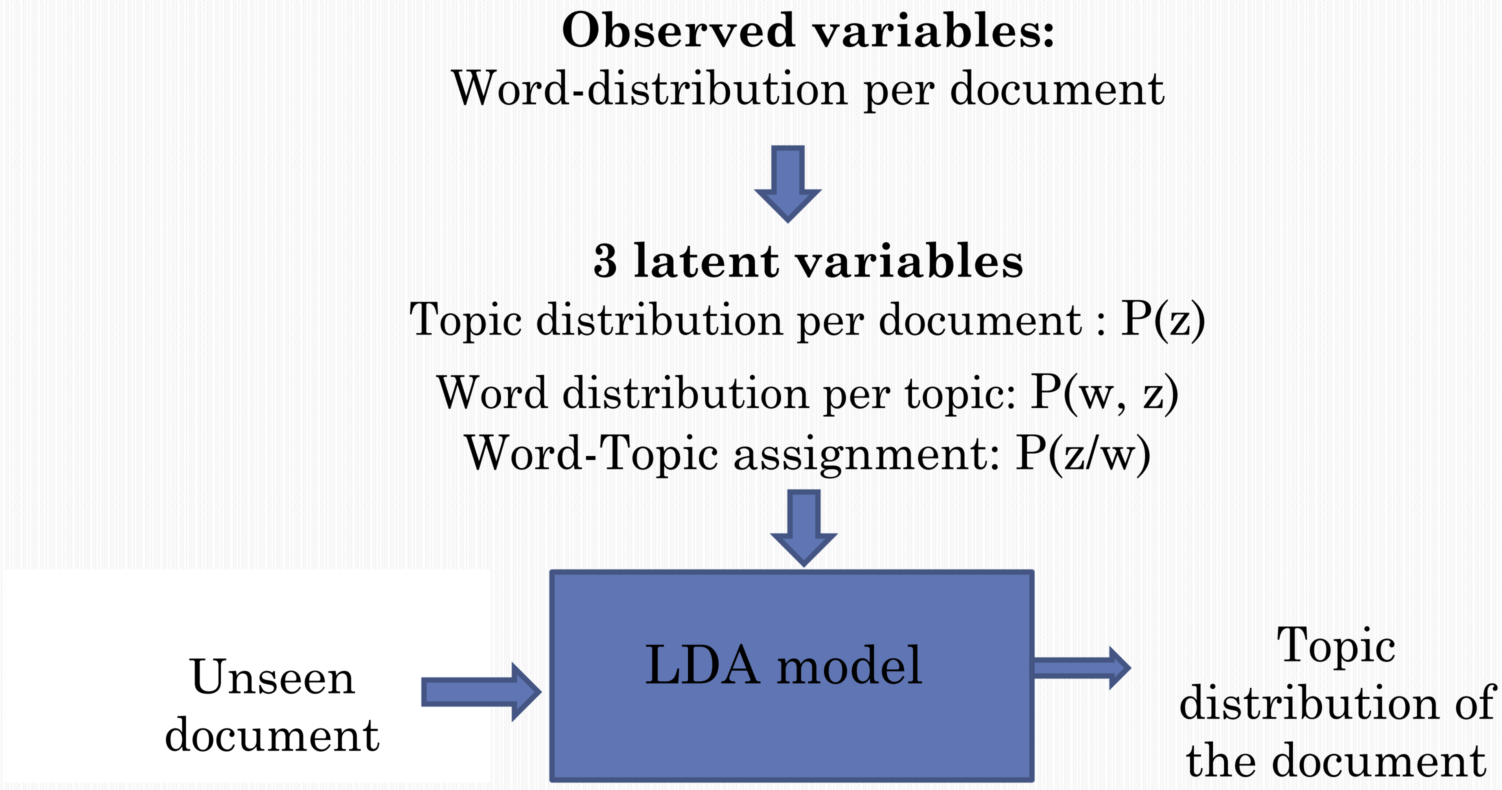
P(W|Z) is the probability of sampling word 'B' from the topic's word distribution.

P(Z/g) is Probability of the topic in the given document's topic distribution

Observed Varaibles:

P(W/g) is probability of the word in the given document

In Nut shell, LDA…..



Observed variables:

Word-distribution per document

3 latent variables

Topic distribution per document : P(z)

Word distribution per topic: P(w, z)

Word-Topic assignment: P(z/w)

Unseen document

LDA model

Topic distribution of the document

Forming a Dictionary : Need

- The Topics generated by LDA are a collection of words. For the topics to make sense to a human, they have to be classified into one of topics which is relatable in terms of real world.
- This exercise can also be thought of as recognizing the direction in which the collection of words in a topic inherently point to.

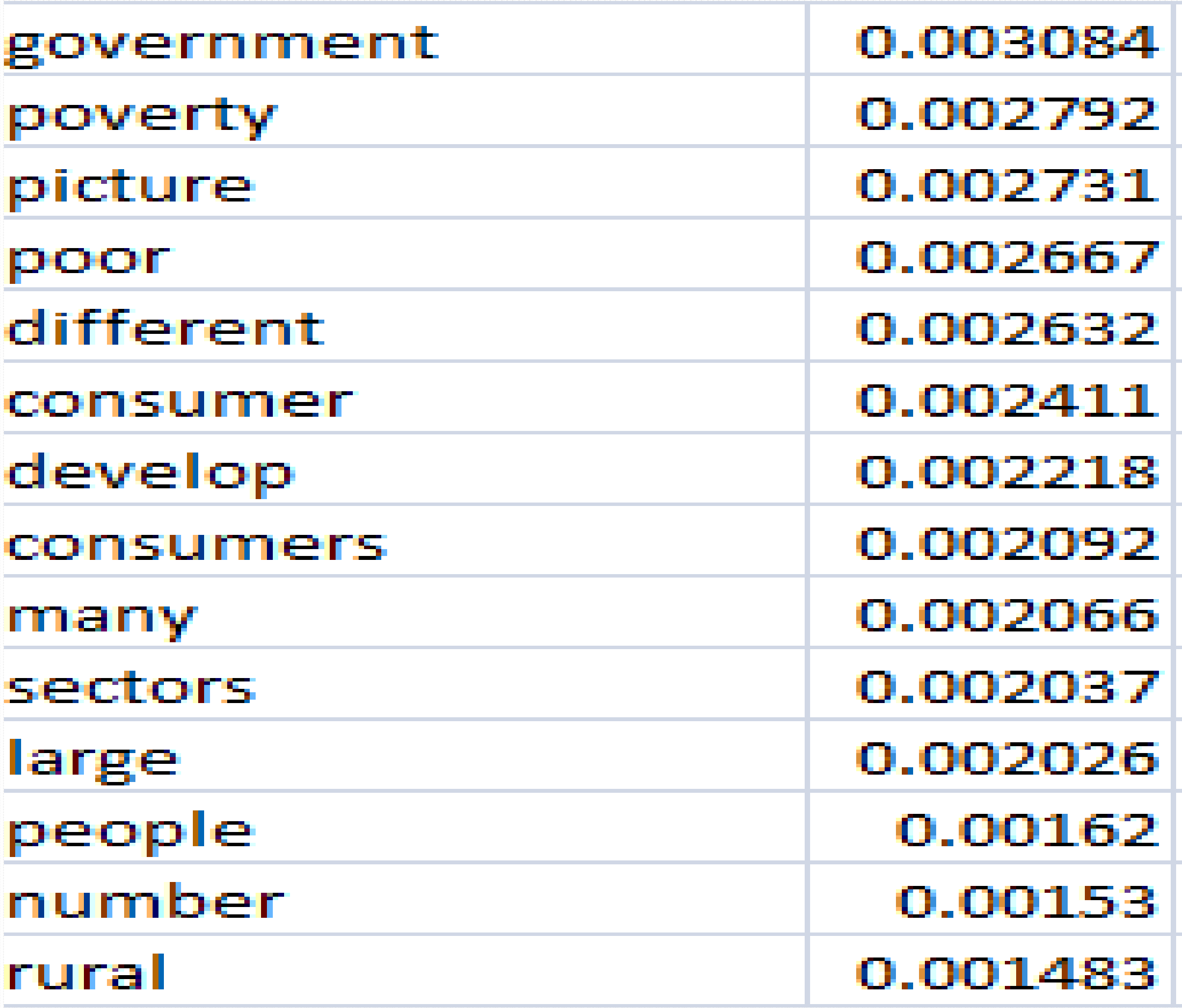
Forming a Dictionary : Methodology

- This necessitated the forming of a dictionary which would serve as a measure to categorize the topics in terms of real world topics.
- To form a dictionary, we first parsed the PDF of the major topics to form a collectable set of Documents upon which LDA was run to produce 3 Topics. Then top 1000 words according to the probability were sorted and chosen.
- Doing this over a variety of subjects gave us a comparative frame of reference.

Forming a Dictionary : Pitfalls to avoid

- The LDA gives probability outputs in from of Log transforms which have to be converted to Probabilities before sorting them in descending order and using them. Some distributions of topics might follow a uniform distribution thereby making the cut-off threshold difficult to decide. (General use here, 1000 words).
- It would be better to first visually look at a distribution and then select a threshold for the number of representative words in a Topic. As a general principle, use of thick-tailed distributions should be avoided.
- Selections of texts and number of topics should be done in such a way so as to get a 'sharp' word probability distribution.
- The process of running LDA over a number of texts and consequent 'cleaning' should be implemented on a system with a system with competitive H/W so as to meet the computation intensive demands of the process.

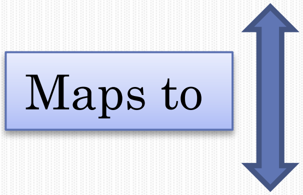
The dictionary for a topic : Economics



government	0.003084
poverty	0.002792
picture	0.002731
poor	0.002667
different	0.002632
consumer	0.002411
develop	0.002218
consumers	0.002092
many	0.002066
sectors	0.002037
large	0.002026
people	0.00162
number	0.00153
rural	0.001483

HOW TO LABEL TOPICS

Topic 1	Topic 2	Topic 3	Topic 4
education	city	neutrality	polybag
youth	people	internet	farmers
skill	gt	access	youth
skills	nbsp	telecom	skill
development	youth	companies	india



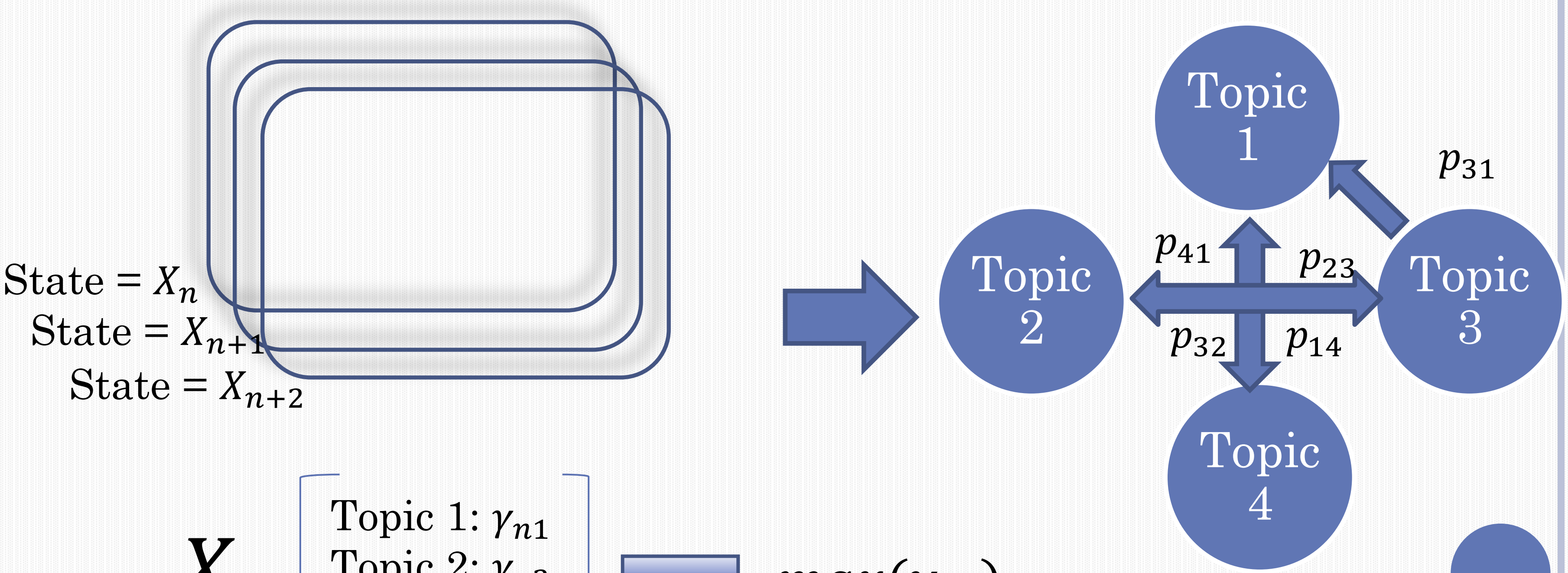
Maps to

Pol. Sci.	Geography	Economics	Environment

Topic Transition Modeling

Problem

Under the information of topic distribution for time sequenced state of conversation learning the transitioning tendencies of topics



$$X_n \left[\begin{array}{l} \text{Topic 1: } \gamma_{n1} \\ \text{Topic 2: } \gamma_{n2} \\ \text{Topic 3: } \gamma_{n3} \\ \text{Topic 4: } \gamma_{n4} \\ \quad \dots \end{array}\right] = max(\gamma_{ni})$$

Model

Conversation ~ Discrete Time Markov Chain

Sequence of random variables where given the entire past history of states, present state depends only on the most recent past state.

$\Pr(X_{n+1} = x \mid X_1 = x_1, X_2 = x_{2,\ldots,}X_n = x_n) =$

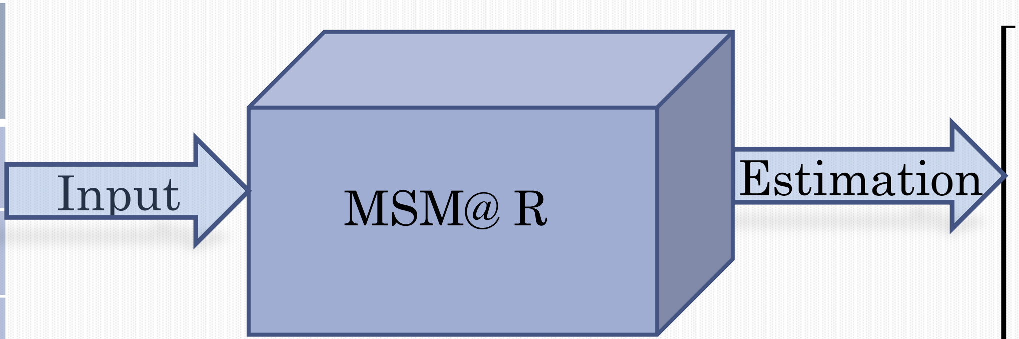
$\Pr(X_{n+1} = x \mid X_n = x_n)$,if $\Pr\left(X_1 = x_1, X_2 = x_{2,\ldots,}X_n = x_n\right) > 0$

ASSUMPTIONS

- Conversation can be classified into distinct non-overlapping states.
- Present state of conversation depends only on the most recent past.
- Time-Homogeneous with finite state space.
- Sampling times are non-informative and exact w.r.t to transitions.

Model Estimation

ID	Time	State
1	1	4
1	2	4
1	3	3
…	…	…



$$\begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & & \vdots \\ \vdots & \cdots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix}$$

MLE Estimate:

$$\hat{M} = \{\hat{\theta}\}\ where$$

$$\{\hat{\theta}\} = \frac{n_{rc}}{\sum_{j=1}^{h} n_{rj}}$$

$$L(\theta) = \prod_{r=1}^{h} \prod_{c=1}^{h} \theta^{n_{rc}}$$

Markov Chain transition modeling:

On applying LDA to 30000 reviews/comments,

LDA model provides topic distribution for each document:

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
Doc1	Z11	Z12	Z13	Z14
Doc2	Z21	Z22	Z23	Z24
Doc3	Z31	Z32	Z33	Z34
		……………………		
		……………………		
DocN	**Zn1**	**Zn2**	**Zn3**	**Zn4**

$$Transition\,Matrix = \begin{bmatrix} p_{11} & \cdots & p_{14} \\ \vdots & \ddots & \vdots \\ p_{41} & \cdots & p_{44} \end{bmatrix}_{4\,x\,4}$$

$$[z_{11}\ \ z_{12}\ z_{13}\ z_{14}] \begin{bmatrix} p_{11} & \cdots & p_{14} \\ \vdots & \ddots & \vdots \\ p_{41} & \cdots & p_{44} \end{bmatrix}_{4\,x\,4} = [z_{21}\ \ z_{22}\ z_{23}\ z_{24}]$$

Assuming conversation is follows Markov chain and it is homogenous,

Doc1 Topic X Transition matrix
= Doc2 Topic probability distribution

So , from above equation, We would get four linear equations:

$$z_{11}p_{11} + z_{12}p_{21} + z_{13}p_{31} + z_{14}p_{41} + 0.p_{21} + 0.p_{22} + 0.p_{23} \ldots\ldots. + 0.p_{44} = z_{21}$$

$$0.p_{11} + 0.p_{21} + 0.p_{31} + 0.p_{41} + z_{11}p_{12} + z_{12}p_{22} + z_{13}p_{32} + z_{14}p_{42} \ldots\ldots. + 0.p_{44} = z_{22}$$

$$0.p_{11} + 0.p_{21} + 0.p_{31} + 0.p_{41} + \cdots \ldots\ldots z_{11}p_{13} + z_{12}p_{23} + z_{13}p_{33} + z_{14}p_{43} \ldots + 0.p_{44} = z_{32}$$

$$0.p_{11} + 0.p_{21} + 0.p_{31} + 0.p_{41} + \cdots \ldots\ldots z_{11}p_{14} + z_{12}p_{24} + z_{13}p_{34} + z_{14}p_{44} = z_{42}$$

Doc2 Topic X Transition matrix
= Doc3 Topic probability distribution
Above equation, would provide 4 such linear equations

-16 variables to be estimated using 120 equations using conversation transition information from doc1 to doc 30.

- Using linear program modelling with inequality constraints, we can solve for transition parameters.

Dynamic Transition modeling:

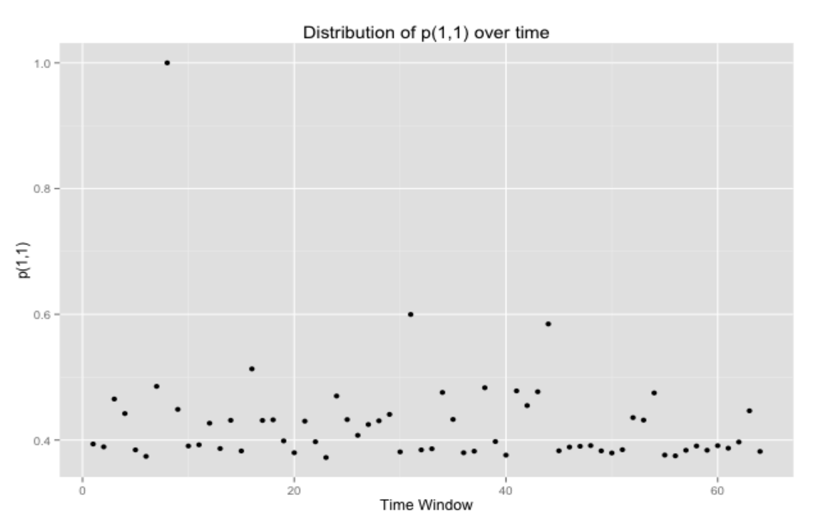
-Transition matrix is estimated using window of 30 conversations.

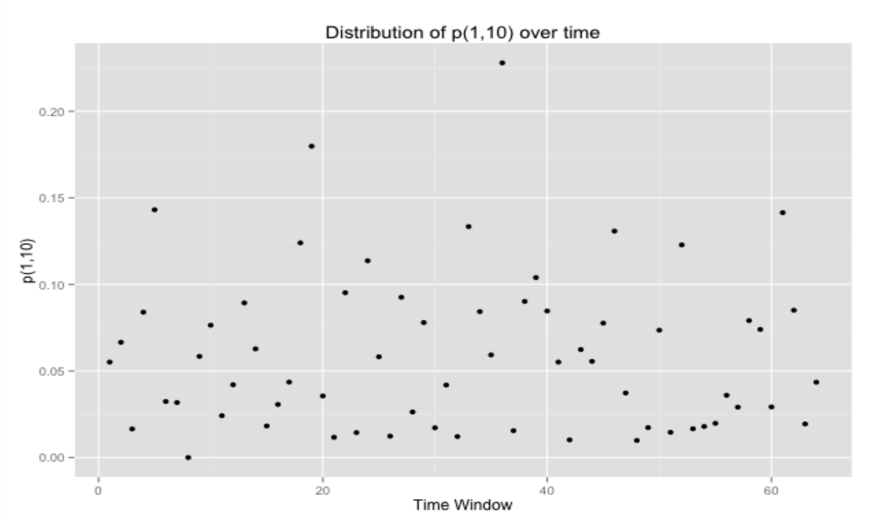
- Window slides down by 10 conversations, and transition matrix is estimated using window of 30 conversations in the current window.

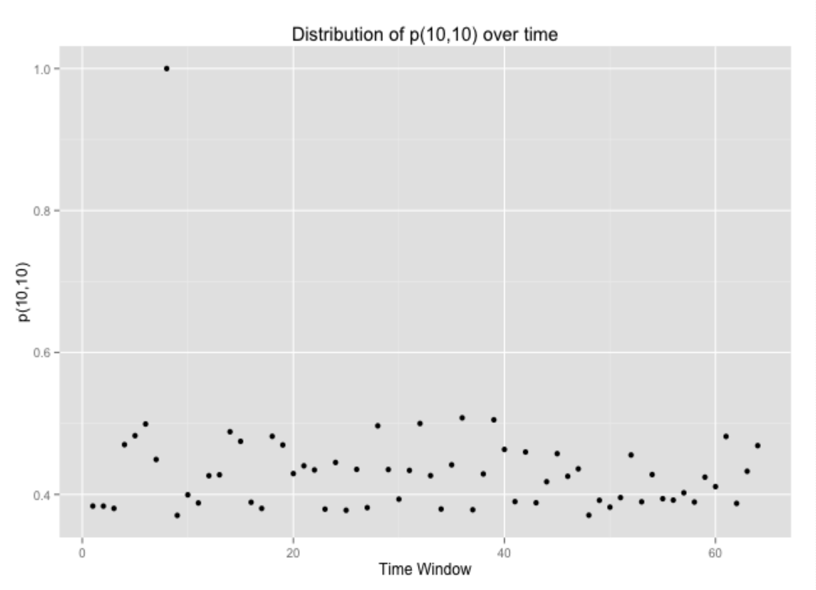
Model/Assumption Assessment

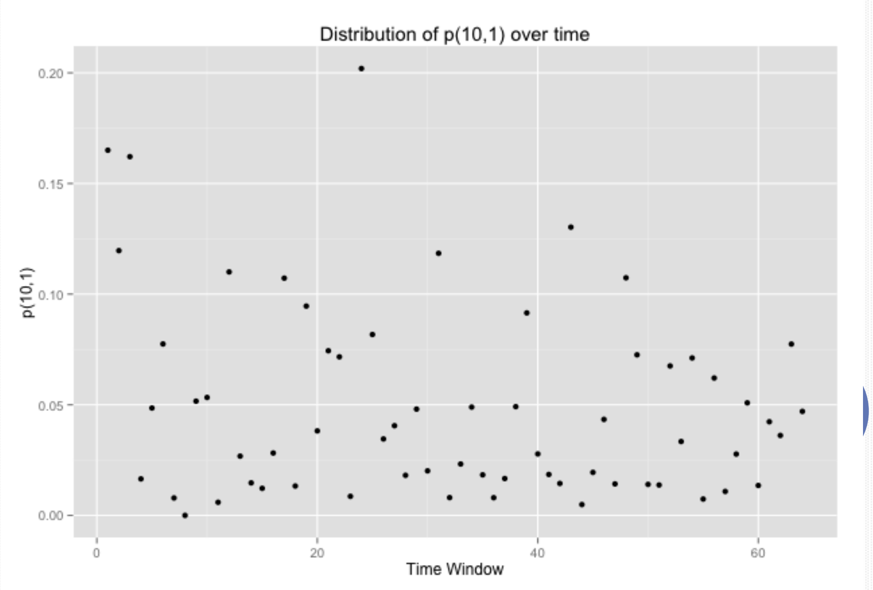
- Likelihood Measure of the Fit
- Observed v/s Expected Prevalence
- Piecewise Homogeneous Markov Chain Modeling
- Nature of Stationarity
- Continuous Time Markov Process Modeling
- Constrained Linear Programming Modeling

Time Series of Transition probabilities

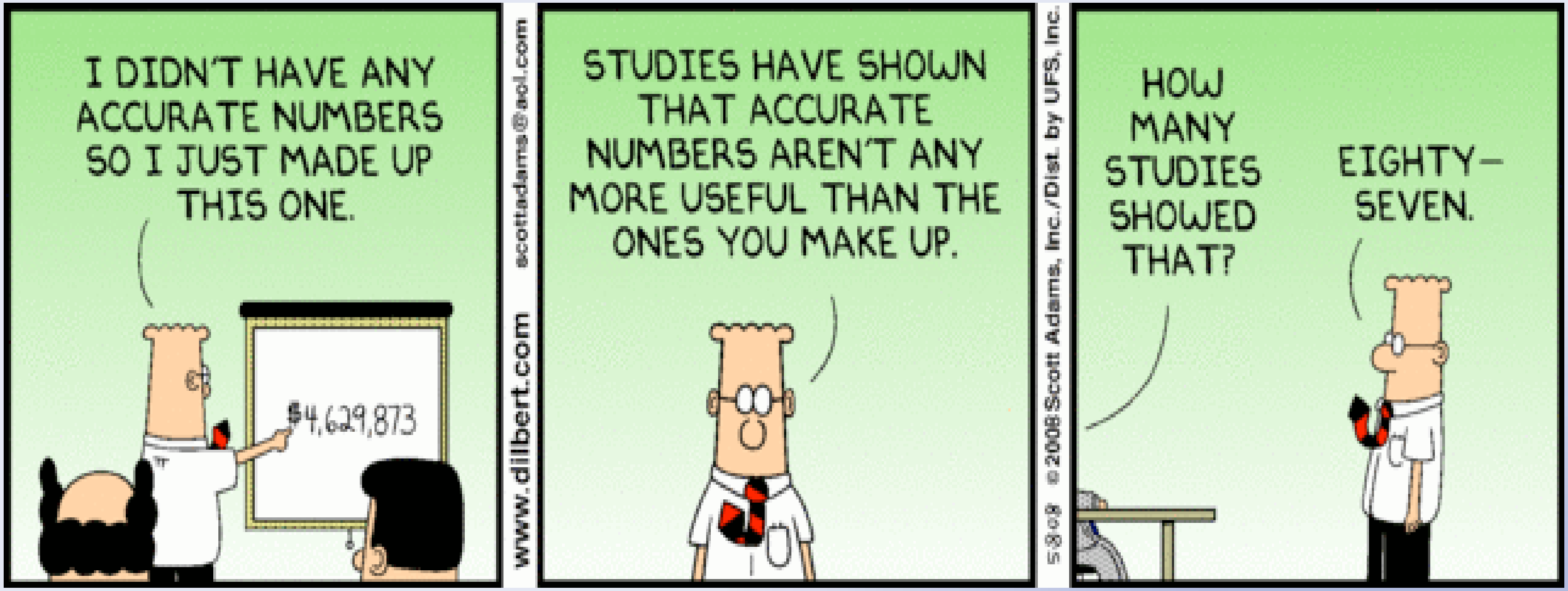

Distribution of p(1,1) over time


Distribution of p(1,10) over time


Distribution of p(10,10) over time


Distribution of p(10,1) over time

"87% OF STATISTICS ARE MADE UP ON THE SPOT"


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

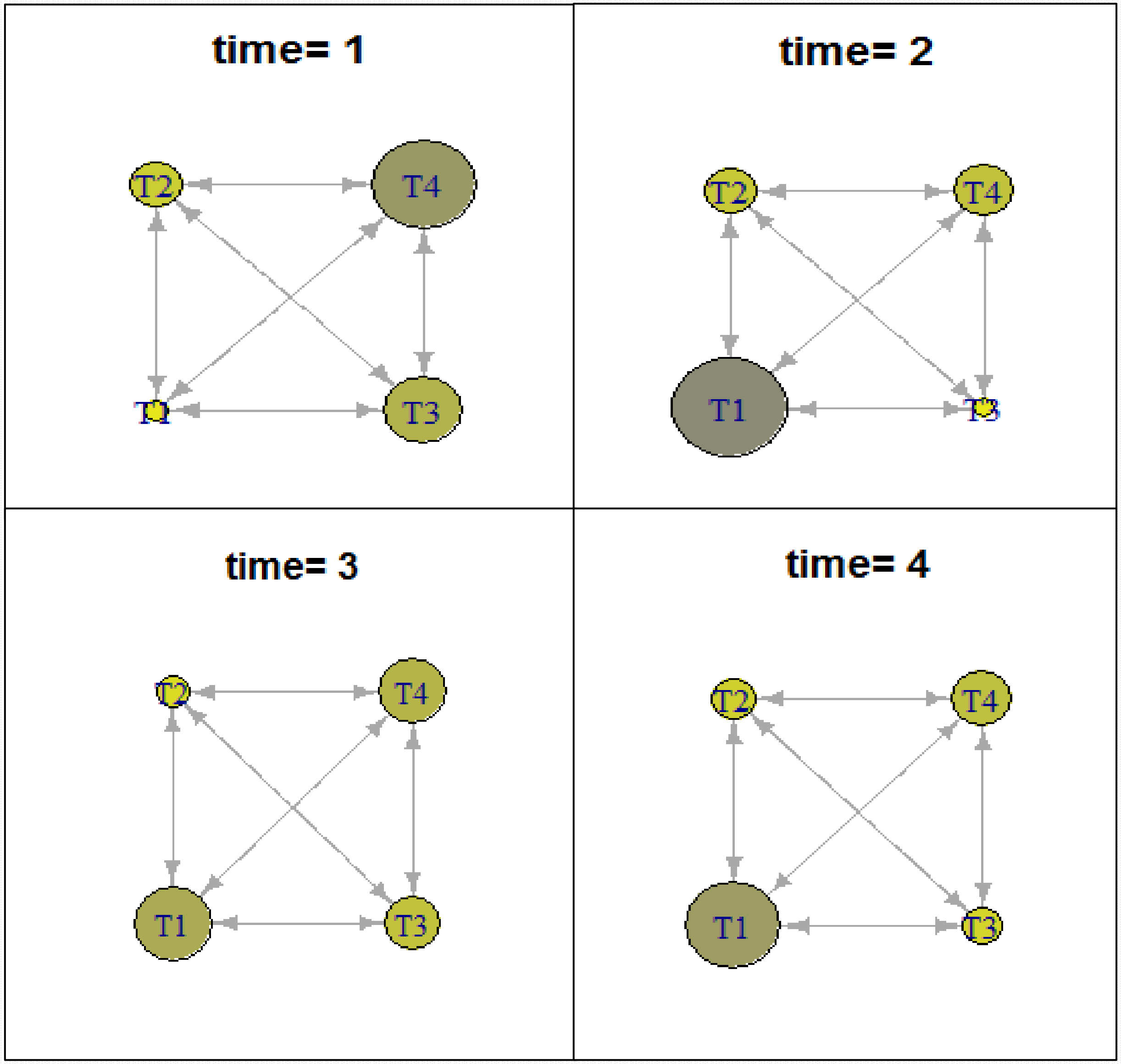
www.dilbert.com scottadams@aol.com

5-8-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.

Visualization

- Visualization was carried out using igraph package in R
- Intuitive
  1. Temporal variations can be easily showcased
  2. Nodes represent the topics
  3. Node size represents the corresponding topic probabilities
  4. Changing node sizes represent the transitions

Visualization (Contd..)



Challenges

- Difficult to get enough volume of conversational data
- Though twitter data can be extracted easily and covers a large variety of topics, the data is very unclean and mining information is very cumbersome
- No translators available to convert Hindi words to English and hence there might be loss of useful information
- Building a dictionary of Hindi words
- Need for Human Evaluation of Topics modeled by LDA
- Getting the correct number of Topics
- Time In-homogeneous nature of Markov Chain

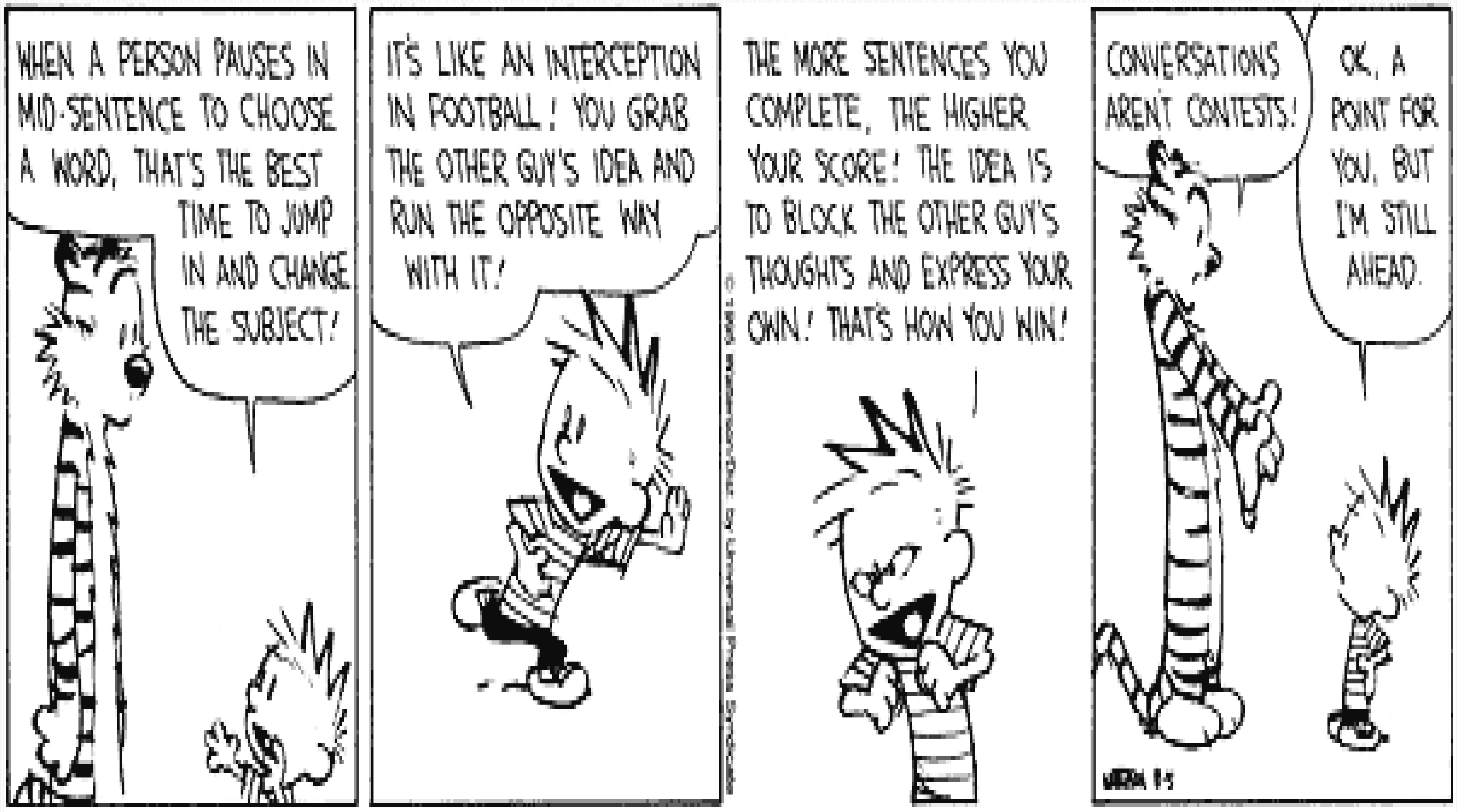
FUTURE WORK

- Modeling time in-homogeneous Markov chain
- Experimenting with the weighting in the DTM
- Building better dictionaries for labeling the topics
- Incorporating shingles for building the DTM

References :-

- http://kateto.net/network-visualization
- http://www.r-bloggers.com/temporal-networks-with-igraph-and-r-with-20-lines-of-code/
- Latent Dirichlet Allocation, David M. Blei et.al, Journal of Machine Learning Research 3 (2003) 993-1022, Submitted 2/02; Published 1/03

Conversations !!
… And our aspirations to model it ☺


WHEN A PERSON PAUSES IN MID-SENTENCE TO CHOOSE A WORD, THAT'S THE BEST TIME TO JUMP IN AND CHANGE THE SUBJECT!
IT'S LIKE AN INTERCEPTION IN FOOTBALL! YOU GRAB THE OTHER GUY'S IDEA AND RUN THE OPPOSITE WAY WITH IT!
THE MORE SENTENCES YOU COMPLETE, THE HIGHER YOUR SCORE! THE IDEA IS TO BLOCK THE OTHER GUY'S THOUGHTS AND EXPRESS YOUR OWN! THAT'S HOW YOU WIN!
CONVERSATIONS AREN'T CONTESTS!
OK, A POINT FOR YOU, BUT I'M STILL AHEAD.
WATTERSON 8-5

The End.