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](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
Comment 1: Please do focus on this so 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.
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
Comment: empowerment of sarpanch in a village is under training and also what empowered citizen of India may get from all schemes of India is also under training.

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

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

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

• Hindi Dictionary

• Aspell() utils package in R

• Annotate() openNLP and NLP package in R

POS Noun: empowerment village India schemes India

Adjective: empowered

Verb: is training
## OUTPUT AFTER DATA CLEANING

<table>
<thead>
<tr>
<th>Document</th>
<th>Nouns + verbs + Adjectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>representation Parliament youth bulk population way voice schools colleges youth participation country decision</td>
</tr>
<tr>
<td>2</td>
<td>heritage city developments heritage contrast Smart city project people nature zones Govt</td>
</tr>
<tr>
<td>3</td>
<td>scheme Mahatma Gandhi guarantee scheme India employee contract basis contract employee basis trailer future</td>
</tr>
<tr>
<td>4</td>
<td>passport office court bench vigilance court city Bhopal city power system city offices institutions</td>
</tr>
</tbody>
</table>
“All generalizations are false, including this one.”

-Mark Twain
Latent Dirichlet allocation (LDA) is a generative model that aims to explain observed variables in terms of latent variables (topics). Using latent variables, it describes why some of the data is similar.
LATENT DIRICHLET ALLOCATION (LDA)

**Topic** \( \rightarrow \) 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** \( \rightarrow \) 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
To describe corpus through LDA, following information required:

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

Other information

- $\alpha$ - Dirichlet prior parameter (sampled once for given data).
- $\beta$ - 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**

Randomly assign
topics to each word

- Topic -1
- Topic -2

**Factors affecting topic assignments**

*$\text{How likely is a word } w \text{ for a topic } j?*$
Probability of word $w$ under topic $j$

*$\text{How dominante is a topic } j \text{ in a doc } d?*$
Probability that topic $j$ document $d$

<table>
<thead>
<tr>
<th></th>
<th>topic1</th>
<th>topic2</th>
</tr>
</thead>
<tbody>
<tr>
<td>money</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>bank</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Loan</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>River</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Stream</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
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
   \[ P(\text{word wi } / \text{topic zi}) \times P(\text{topic zi } / \text{doc1}) \]

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 $w_i$, conditioned on all other assigned topics of words in this doc and all other observed vars.

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^{W} C_{wj}^{WT} + W \beta} \frac{C_{d_{i j}}^{DT} + \alpha}{\sum_{t=1}^{T} C_{di t}^{DT} + T \alpha}$$

Count number of times a topic $j$ was already assigned to some word token in doc $d_i$.

Count number of times a word token $w_i$ was assigned to a topic $j$ across all docs.
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 Variables:**

\( P(W|g) \) 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 $\rightarrow$ LDA model $\rightarrow$ 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 competitive H/W so as to meet the computation intensive demands of the process.
The dictionary for a topic: Economics

<table>
<thead>
<tr>
<th>Term</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>government</td>
<td>0.003084</td>
</tr>
<tr>
<td>poverty</td>
<td>0.002792</td>
</tr>
<tr>
<td>picture</td>
<td>0.002731</td>
</tr>
<tr>
<td>poor</td>
<td>0.002667</td>
</tr>
<tr>
<td>different</td>
<td>0.002632</td>
</tr>
<tr>
<td>consumer</td>
<td>0.002411</td>
</tr>
<tr>
<td>develop</td>
<td>0.002218</td>
</tr>
<tr>
<td>consumers</td>
<td>0.002092</td>
</tr>
<tr>
<td>many</td>
<td>0.002066</td>
</tr>
<tr>
<td>sectors</td>
<td>0.002037</td>
</tr>
<tr>
<td>large</td>
<td>0.002026</td>
</tr>
<tr>
<td>people</td>
<td>0.00162</td>
</tr>
<tr>
<td>number</td>
<td>0.00153</td>
</tr>
<tr>
<td>rural</td>
<td>0.001483</td>
</tr>
</tbody>
</table>
## HOW TO LABEL TOPICS

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>education</td>
<td>city</td>
<td>neutrality</td>
<td>polybag</td>
</tr>
<tr>
<td>youth</td>
<td>people</td>
<td>internet</td>
<td>farmers</td>
</tr>
<tr>
<td>skill</td>
<td>gt</td>
<td>access</td>
<td>youth</td>
</tr>
<tr>
<td>skills</td>
<td>nbsp</td>
<td>telecom</td>
<td>skill</td>
</tr>
<tr>
<td>development</td>
<td>youth</td>
<td>companies</td>
<td>india</td>
</tr>
</tbody>
</table>

**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

\[ \text{State} = X_n \]
\[ \text{State} = X_{n+1} \]
\[ \text{State} = X_{n+2} \]

\[
\begin{bmatrix}
\gamma_{n1} \\
\gamma_{n2} \\
\gamma_{n3} \\
\gamma_{n4} \\
\vdots
\end{bmatrix}
\]

\[
max(\gamma_{ni})
\]
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 \(Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) > 0\)

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**

<table>
<thead>
<tr>
<th>ID</th>
<th>Time</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Input**

**Estimation**

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

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

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

**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:

<table>
<thead>
<tr>
<th></th>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>Z11</td>
<td>Z12</td>
<td>Z13</td>
<td>Z14</td>
</tr>
<tr>
<td>Doc2</td>
<td>Z21</td>
<td>Z22</td>
<td>Z23</td>
<td>Z24</td>
</tr>
<tr>
<td>Doc3</td>
<td>Z31</td>
<td>Z32</td>
<td>Z33</td>
<td>Z34</td>
</tr>
<tr>
<td>DocN</td>
<td>Zn1</td>
<td>Zn2</td>
<td>Zn3</td>
<td>Zn4</td>
</tr>
</tbody>
</table>
Assuming conversation is follows Markov chain and it is homogenous,

\[
\text{Doc1 Topic X Transition matrix} = \text{Doc2 Topic probability distribution}
\]

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

\[
\begin{align*}
z_{11}p_{11} + z_{12}p_{21} + z_{13}p_{31} + z_{14}p_{41} + 0.p_{21} + 0.p_{22} + 0.p_{23} + \cdots + 0.p_{44} &= z_{21} \\
0.p_{11} + 0.p_{21} + 0.p_{31} + 0.p_{41} + z_{11}p_{12} + z_{12}p_{22} + z_{13}p_{32} + z_{14}p_{42} + \cdots &+ 0.p_{44} = z_{22} \\
0.p_{11} + 0.p_{21} + 0.p_{31} + 0.p_{41} + \cdots + z_{11}p_{13} + z_{12}p_{23} + z_{13}p_{33} + z_{14}p_{43} + \cdots + 0.p_{44} &= z_{32} \\
0.p_{11} + 0.p_{21} + 0.p_{31} + 0.p_{41} + \cdots + z_{11}p_{14} + z_{12}p_{24} + z_{13}p_{34} + z_{14}p_{44} &= z_{42}
\end{align*}
\]

\[
\text{Doc2 Topic X Transition matrix} = \text{Doc3 Topic probability distribution}
\]

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 vs 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”
Visualization

- Visualization was carried out using the igraph package in R
- Intuitive
  - Temporal variations can be easily showcased
  - Nodes represent the topics
  - Node size represents the corresponding topic probabilities
  - 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
CONVERSATIONS !
... AND OUR ASPIRATIONS TO MODEL IT 😊

The End.