Search Project

Identifying Transition Points Within Browsing History

Group:
22111013 - Atul Kumar
22111054 - Shubham Rathore
22111069 - Vivek Kumar Gautam

Problem statement

- Option 2: Identify transition points within browsing history. Dig into the LDA code, or your favorite clustering algorithm's code, to probabilistically characterize the location of the user within the topic at the time of a transition.
- Bonus points for using dwell time as a marker

Methodology

- Extract relevant data from Google Chrome's history
- Perform topic modeling
- Create a Transition Probability Matrix
- Model the topic transitions as a Discrete Time Markov Chain

Exploring  history database

- (Linux) Located at: `~/.config/google-chrome/Default/History`
- Run
  - `sqlite3 ~/.config/google-chrome/Default/History`
- Save to urls.csv:
  - `SELECT datetime(last_visit_time/1000000-11644473600, 'unixepoch'), id, url, title FROM urls ORDER BY last_visit_time DESC;`
- Save to visits.csv:
  - `SELECT id, url, datetime(visit_time/1000000-11644473600, 'unixepoch'), from_visit, (visit_duration/1000000) FROM visits ORDER BY visit_time DESC;`

Topic modeling with BERTopic

- What is BERTopic?
  - Topic modeling technique
  - Leverages BERT embeddings and c-TF-IDF
  - Creates dense clusters allowing for easily interpretable topics
- Clustering
  - Train BERTopic on the stored URL database
- Usefulness
  - Get title of a URL through database query or Beautiful Soup
  - Pass the string to BERTopic model
  - Get the topic index

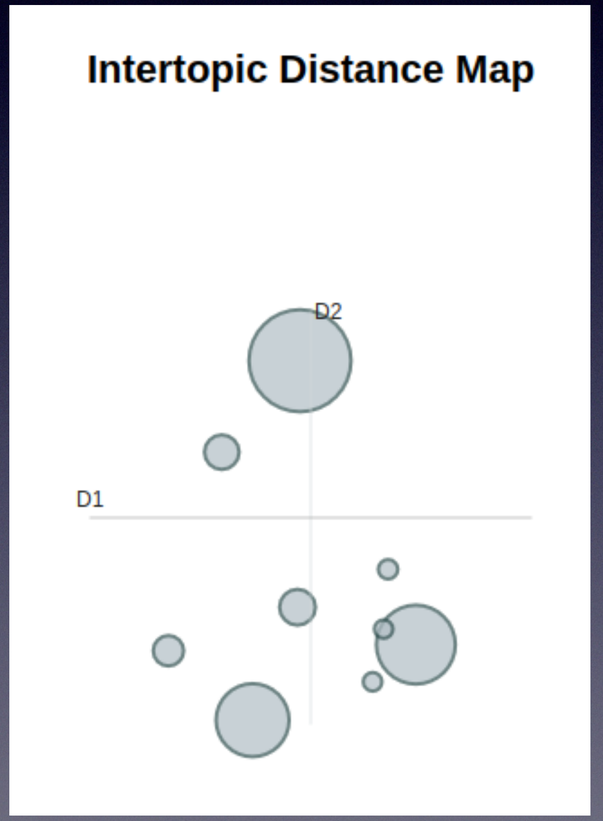


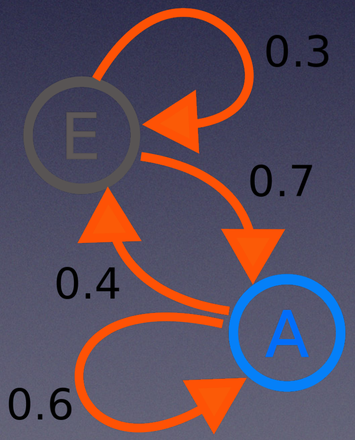
Fig 1 - *Topic clusters made from BERTopic*

Creating a Transition Probability Matrix

- Interpreting the transition probability matrix, `P`
  - `n x n` matrix, where `n` = number of topics
  - `P[i][j]` : Probability of transitioning from topic `i` to topic `j`
  - `P[i][i]` : Probability of staying at the same topic, interpretable as the 'location'
- Enumerate all transitions
  - `from_visit` in `visits.csv` tells us the parent link
  - Query the database to get topics of parent and child links
  - Increment `P[from][to]`
  - Repeat till done
- Calculating probabilities
  - For each row in `P`, divide each element of that row by its row sum

Discrete Time Markov Chain (DTMC)

- `Prob(next state | prev state) = Prob(next state | all prev states)`
- Each transition is a discrete time step
- Has a set of states within which transitions are happening
- Can be described by a stochastic matrix
- Probabilities after an arbitrary number of steps in future can be calculated
- How can a DTMC describe topic transitions?
  - Discrete time step -> link click
  - Set of states -> topics
  - Stochastic matrix -> Transition probability matrix
  - `t` steps in future -> Location of user within topic after `t` link clicks



A Markov chain with 2 states

Conclusion

- Chrome's history database was queried for useful URL data
- URL titles and BERTopic were used to form topic clusters
- Topic transitions were obtained using the URL data and the topic model
- DTMC probabilistically characterizes user's location within a topic

Thank you