

Analyzing Web History by Using Topic Modelling

Ritesh Kantule

Yogesh



Indian Institute of Technology Kanpur

CS616A: Human Centered Computing

Dr. Nisheeth Srivastava

21 August 2023

Abstract

The "Exploring Web History Analysis with Python" project presents a comprehensive approach to gaining insights into user behavior and interests by analyzing web history data. In the modern digital landscape, understanding user preferences and trends on the web is essential for informed decision-making and user-centric content optimization. This project employs a Python script that leverages various libraries, including pandas, numpy, matplotlib, requests, BeautifulSoup, nltk, gensim, and wordcloud, to extract, preprocess, and analyze web history records. The script covers data retrieval, cleaning, web scraping for titles, text preprocessing, LDA topic modeling, and visualization techniques. Latent Dirichlet Allocation (LDA) is employed for topic modeling, enabling the identification of hidden themes within the textual data. The generated visualizations, such as word clouds, topic trend plots, and cumulative hourly topic probabilities, provide intuitive representations of user interests and behavior patterns. The project's insights are valuable for organizations seeking to enhance user experience, optimize content strategies, and tailor offerings to match user preferences. By delving into the intricacies of web history data, this project contributes to the field of data-driven decision-making and user-centric design.

Method

1. Data Acquisition:

Loaded web history data from a CSV file using the pandas library.

We have converted the 'Time' column to a datetime format to enable accurate temporal analysis.

We Handled invalid URLs and data anomalies using error-handling mechanisms.

2. Text Preprocessing:

Tokenized text using a regular expression tokenizer to split into discrete words.

Removed stop words to filter out common and insignificant words.

Applied stemming using the PorterStemmer to reduce words to their base forms.

3. Latent Dirichlet Allocation (LDA) Topic Modeling:

Prepared the 'texts' list containing tokenized and preprocessed text from documents.

Created a dictionary using the gensim library to map words to numerical identifiers.

Generated a document-term matrix representing word frequencies for each document.

We utilized the LDA model to extract latent topics within the text data.

4. Visualization:

We utilized word clouds to represent the top words in each LDA-generated topic visually.

We plotted cumulative hourly topic probabilities to display topic distributions over time.

We created line plots to showcase topic trends and evolution over the analyzed period.

5. Topic Assignment and Analysis:

Assigned topics to web history records by calculating the probability distribution using

The LDA model. We selected the topic with the highest probability for each URL.

Grouped data by hours and days of the week to analyze temporal patterns.

6. Interpretation and Insights:

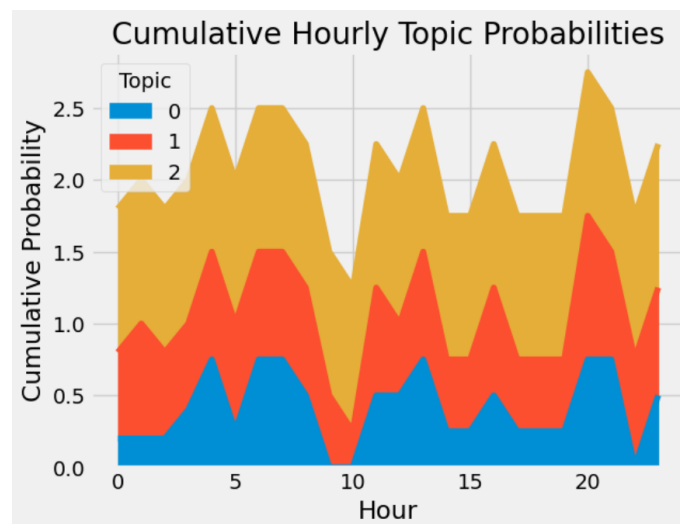
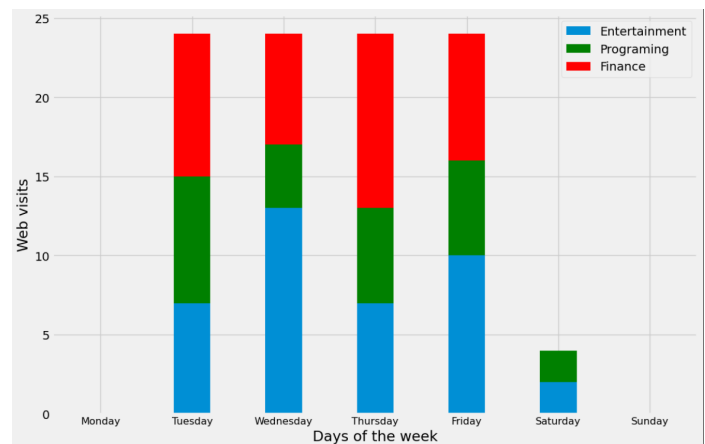
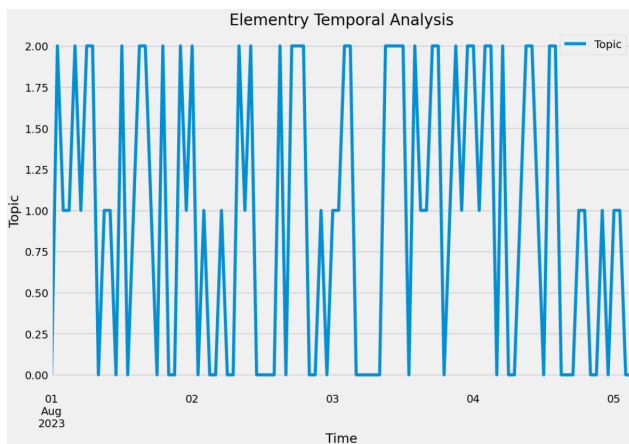
Analyzed generated visualizations and topic assignments to uncover user interests and behavior patterns. Extracted insights into preferred topics, temporal trends, and user engagement.



2. Temporal Trends and Patterns:

Cumulative hourly topic probabilities revealed fluctuations in user interests throughout the day.

Day-wise topic occurrences unveiled variations in preferences based on weekdays.



3. Content Optimization Potential:

The analysis offered valuable insights for content creators and providers to tailor offerings to user interests.

Understanding temporal patterns allowed for strategic scheduling of content delivery.

4. User Engagement Insights:

Topic trends over time highlighted changes in user engagement and evolving interests.

Topic assignments offered a means to track user interactions and preferences.

Conclusion

In conclusion, the "Exploring Web History Analysis with Python" project showcased the power of data analysis in uncovering valuable insights from web history data by applying a holistic approach that encompasses data retrieval, cleaning, web scraping, text preprocessing, topic modeling, and visualization techniques.

Applications

1. Enhanced Understanding of User Behavior:

The analysis revealed hidden topics and trends that provide a comprehensive view of user interests and preferences.

2. Data-Driven Decision-Making:

The insights derived from the project have practical applications for organizations aiming to optimize content strategies and improve user experience.

3. Future Possibilities:



The project lays the foundation for future explorations, including refining text preprocessing techniques, experimenting with different topic modeling algorithms, and integrating more advanced visualization methods.

Ultimately, the "Exploring Web History Analysis with Python" project underscores the importance of data-driven approaches in understanding user behavior, enriching content strategies, and fostering informed decision-making.

Discussion

The exploration of web history analysis using Python has illuminated crucial aspects of user behavior, content consumption, and the effectiveness of data-driven strategies. The project's methodologies and findings underscore the significance of accurate data preprocessing in handling challenges such as invalid URLs and unavailable titles. While the Latent Dirichlet Allocation (LDA) model effectively identifies latent topics, further experimentation with model parameters offers avenues for refinement. Ethical considerations regarding user privacy and consent highlight the importance of responsible data analysis. The project's applicability extends beyond web history analysis, offering insights into social media trends and customer sentiment. Integrating data retrieval, preprocessing, modeling, and visualization techniques presents a comprehensive approach to extracting valuable insights from web history data. This enables organizations to optimize content strategies and make informed decisions in a data-centric landscape.

References

-  Latent Dirichlet Allocation (Part 1 of 2)
-  Training Latent Dirichlet Allocation: Gibbs Sampling (Part 2 of 2)
- <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0>
- <https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13>