

# Integrating OCR and NER for the Systematic Analysis of Puranic Texts

Ishan Sharma Subeer Rai

September 25, 2024

---

**Supervised by:**

Prof. [Arnab Bhattacharya Dept. of Computer Science and Engineering IIT Kanpur]

## Abstract

The analysis of ancient Puranic texts is a challenging task due to their complex linguistic structures, multiple scripts, and historical variations in formatting. To address these challenges, a comprehensive methodology was designed and executed, combining Optical Character Recognition (OCR) for digitization, extensive text cleaning operations, and advanced Named Entity Recognition (NER) techniques. This multi-stage approach, detailed step-by-step, was critical in extracting meaningful entities from the Puranic texts, ensuring accuracy, and enabling advanced computational analysis.

The process of text digitization began with Optical Character Recognition (OCR), where EasyOCR was employed to convert scanned images of the Puranic manuscripts into machine-readable text. OCR is a key technology that transforms printed or handwritten texts into editable formats. However, OCR is particularly challenging when applied to ancient texts like the Puranas, which contain multiple scripts (e.g., Sanskrit, Devanagari), regional fonts, and diacritics. Many of these texts are also affected by poor print quality, faded ink, and irregular layouts, all of which make accurate recognition difficult. To overcome these hurdles, EasyOCR was configured with customized settings to handle these complexities. For instance, we adjusted parameters to improve accuracy when dealing with ornate fonts and diacritics common in Sanskrit and regional languages. Despite these configurations, the OCR output was not flawless; it still contained errors such as incorrect character recognition and misaligned text blocks. To address this, manual validation and corrections were undertaken to fix common OCR mistakes, including misidentified characters, missing diacritics, and improper word breaks. This combination of automated OCR with human oversight ensured the highest possible accuracy in converting the scanned texts into digital formats. OCR, therefore, played a foundational role in this methodology, enabling the transition from image-based text to a machine-readable format that could be further processed for analysis. Without this digitization, the next stages of text cleaning and NER would not have been feasible.

Once the texts were digitized through OCR, the next step involved extensive text cleaning and preprocessing. This step is crucial because the output from OCR, even after manual correction, still contains noise such as non-relevant symbols, page numbers, and footnotes, which must be removed to ensure clean data for Named Entity Recognition (NER). The first aspect of text cleaning involved removing extraneous elements. Using regex-based techniques, non-relevant characters like special symbols, page numbers, and footnotes were systematically eliminated from the text. These artifacts, if left unremoved, could interfere with the NER model's ability to correctly identify and classify entities. For instance, page numbers and footnotes could be mistakenly identified as named entities,

---

leading to erroneous results. Next, the process addressed the issue of spelling errors and language inconsistencies that arose from both OCR inaccuracies and the historical nature of the texts. This involved automatic spell-checking as well as manual reviews to correct misidentified words, especially those critical to Hindu mythology. For example, names of deities or places that were misspelled due to OCR errors were corrected during this stage. Furthermore, normalization of language variants was applied. Since the Puranas have been transcribed and translated into various languages over time, this step was essential for harmonizing terms and standardizing spellings. Synonyms were aligned with their most commonly accepted forms in Hindu mythology, ensuring consistency across the dataset. Once the text was cleaned of artifacts and errors, tokenization and sentence segmentation were applied. Tokenization breaks down the text into smaller units, such as words or phrases (tokens), while sentence segmentation splits the text into individual sentences. These processes are critical for preparing the data for NER, as they allow the model to process the text in manageable units and recognize entities with greater precision. By cleaning and structuring the data, this stage ensured that the text was ready for the next phase: named entity extraction.

The heart of the methodology lies in the application of Named Entity Recognition (NER) for identifying and classifying key entities from the Puranic texts. NER is a process within natural language processing (NLP) that involves identifying and categorizing named entities such as people, places, and organizations within a text. In this study, NER was used to extract specific entities relevant to Hindu mythology, such as deities, mythological events, sacred places, and characters. To execute this task, the Stanza NLP pipeline, developed by Stanford, was employed due to its advanced tokenization and entity recognition capabilities. The model was initially pre-trained on modern datasets, but given the unique language and context of the Puranas, it required significant customization and fine-tuning. This involved training the model on a custom dataset curated from Hindu mythology, which included annotated examples of deity names, mythological locations, and specific mythological events. By leveraging transfer learning techniques, pre-existing NER models were adapted to the specific linguistic and contextual requirements of the Puranas, significantly improving the model's performance in recognizing ancient and context-specific entities. However, entity extraction in historical texts like the Puranas presents unique challenges, especially with polysemous terms—words that have multiple meanings. For instance, the term "Kali" could refer to the goddess, the concept of time, or the personification of darkness depending on the context. To tackle this, rule-based post-processing techniques were developed to refine the initial entity classifications. Semantic analysis of the surrounding context was applied to ensure that each entity was classified correctly. Additionally, cross-referencing with established mythological databases was used to validate the recognized entities and correct any misclassifications. Moreover, batch processing was employed to handle the

---

voluminous nature of the Puranic texts. Given the extensive length of these scriptures, NER was implemented in a batch processing framework that allowed large sections of text to be processed in parallel. This made the process scalable and more efficient, enabling the model to process vast amounts of data without compromising accuracy. The output from NER included structured data with identified entities classified into various categories, such as PERSON, LOCATION, ORGANIZATION, EVENT, and more.

An essential part of the NER process was the contextual disambiguation of entities, particularly those that appeared with multiple meanings across different contexts. To achieve accurate classification, a combination of rule-based post-processing algorithms and semantic analysis techniques was applied. For example, the term "Kali" could represent the goddess in one context but refer to the concept of time in another. By analyzing the surrounding text, these ambiguities were resolved, and the correct classification was assigned. Cross-referencing with external mythological databases further enhanced the accuracy of entity recognition. Lesser-known figures and rare names from the Puranas were compared against established mythological resources to ensure that even obscure entities were correctly identified and classified. This cross-referencing also provided a verification layer, adding confidence to the extracted entities.

This methodology represents a significant leap in applying computational techniques to the study of ancient texts. By combining OCR, extensive text cleaning, and advanced NER, the study showcases how digital tools can enhance traditional humanities research. The structured data generated through this process makes the Puranas more accessible to scholars and researchers, enabling deeper exploration of these rich narratives. Furthermore, the scalability and adaptability of this approach mean that it can be applied to other historical and cultural texts, paving the way for future studies in digital humanities and computational linguistics.

# Contents

<b>Contents</b>	<b>1</b>
<b>1 Introduction</b>	<b>3</b>
<b>2 Literature Review</b>	<b>6</b>
2.1 Relevance to the Puranas . . . . .	6
2.2 Computational Approaches to Studying Sanskrit and Indic Texts . . . . .	6
2.3 Relevance to the Puranas . . . . .	7
2.4 Digital Humanities and Indian Religious Texts . . . . .	7
2.5 Relevance to the Puranas . . . . .	7
2.6 Puranic Studies and Thematic Analysis . . . . .	8
<b>3 Methodology</b>	<b>9</b>
3.1 Text Digitization: Converting Scanned Images to Machine-Readable Text . . . . .	9
3.2 Text Cleaning and Preprocessing . . . . .	10
3.3 Named Entity Recognition (NER) Application . . . . .	10
3.4 Contextual Disambiguation and Entity Classification . . . . .	11
3.5 Analysis and Visualization of Extracted Entities . . . . .	11
3.6 Validation, Feedback Loop, and Iterative Refinement . . . . .	12
<b>4 Data Analysis</b>	<b>13</b>
4.1 Most Frequently Mentioned Entities . . . . .	13
<b>5 Results</b>	<b>15</b>
5.1 Key Findings from the NER Process . . . . .	15
5.2 Cultural and Religious Insights from Extracted Entities . . . . .	16
5.3 Thematic Patterns and Recurrences Across Texts . . . . .	16
5.4 Challenges and Complexities in Entity Recognition . . . . .	16
5.5 Broader Implications of the NER Results on Puranic Studies . . . . .	17
5.6 Impact on Comparative Mythological and Religious Studies . . . . .	17
5.7 Technological Advancements in Textual Analysis . . . . .	17
5.8 Conclusion . . . . .	18

<b>6 Discussion</b>	<b>19</b>
6.1 SpaCy . . . . .	19
6.2 Stanford NER . . . . .	19
6.3 Hugging Face Transformers (BERT-based models) . . . . .	20
6.4 Flair NLP . . . . .	20
6.5 Polyglot NER . . . . .	20
6.6 Why These Models Are Not Sufficient for Our Use Case . . . . .	21
6.7 Why Our Method Stands Out . . . . .	21
<b>7 Conclusion</b>	<b>23</b>
7.1 Methodological Insights . . . . .	23
7.2 Dominance of Theological Entities . . . . .	24
7.3 Geographical Significance . . . . .	24
7.4 Recurrent Mythological Events . . . . .	24
7.5 Implications for Puranic Studies . . . . .	24

# 1. Introduction

The Puranas are a collection of ancient Hindu texts, dating from approximately 300 to 1500 CE, that provide an extensive narrative of Indian mythology, religious teachings, cosmology, and history. These texts, including prominent works like the Shiva Purana, Vishnu Purana, and Bhagavata Purana, form the bedrock of Hindu religious literature, offering insights into the lives of deities, sages, kings, and the cosmic order. However, their complex narrative style, voluminous content, and the intertwining of mythological and theological themes present significant challenges for analysis, particularly when employing traditional methods.

Named Entity Recognition (NER), a critical tool in Natural Language Processing (NLP), facilitates the extraction and classification of named entities—such as people, locations, and events—from unstructured text. By applying NER to the Puranas, this research aims to systematically identify and categorize key entities, making the texts more accessible and analyzable for scholars. This approach not only aids in understanding the roles and relationships of various entities but also assists in uncovering thematic patterns and cultural insights embedded within these ancient scriptures. Moreover, it embodies a technological approach that merges computational analysis with the rich narratives of the past, demonstrating how modern methods can enhance our understanding of historical texts.

**Importance of NER in the Study of the Puranas** Systematic Entity Analysis: The Puranas are rich with characters, locations, and events central to Hindu mythology. NER helps organize this complex information, allowing researchers to map relationships, trace genealogies, and explore thematic recurrences. This systematic analysis can uncover connections between deities and their stories, providing a clearer understanding of their cultural significance.

**Enhancement of Traditional Scholarship through Digital Humanities:** Integrating NER into the study of the Puranas bridges traditional literary analysis with modern computational techniques. This fusion enhances our understanding of these texts and provides new insights that might be difficult to achieve manually. For instance, the use of NER can reveal patterns and correlations across different Puranas, contributing to a more nuanced interpretation of Hindu mythology.

**Cultural and Religious Insight:** By extracting entities like deities, sacred places, and



mythological events, NER illuminates the significance of these elements within the texts, offering a deeper appreciation of their roles in Hindu religious and cultural contexts. This capability allows scholars to draw connections between textual elements and contemporary cultural practices, enriching the dialogue between past and present.

**Challenges in Accessing and Analyzing the Texts Lack of Readily Available Digital Texts:** The Puranas were often not available in digital text format, necessitating the conversion of scanned pages into editable text. This required Optical Character Recognition (OCR) technology, specifically EasyOCR, to extract text from images of printed books and manuscripts. The process involves:

**Text Extraction from Scanned Images:** Converting images into machine-readable text entails handling diverse print qualities, fonts, and occasional OCR errors that could impact the accuracy of the NER process.

**Text Cleaning and Preprocessing:** Following OCR, extensive text cleaning is required to rectify recognition errors, eliminate unwanted characters, and standardize the data for NER analysis. This preprocessing step is crucial for ensuring that the NER models receive high-quality input, which directly affects the reliability of the extracted entities.

**Complexity of Language and Contextual Recognition:**

**Linguistic Variability and Multilingual Challenges:** The Puranas were originally written in Sanskrit and have been translated into various languages, leading to significant variability in spelling, grammar, and terminology. This linguistic diversity complicates NER, necessitating models to adapt to the nuances of ancient and religious language.

**Contextual Ambiguities in Entity Recognition:** A major challenge in analyzing the Puranas is the contextual ambiguity of certain entities. Many words in the texts serve multiple roles; for example, the term "Ganga" can refer both to the sacred river and to the goddess personified as the river. Similarly, "Kali" can refer to the fierce goddess and to concepts like "time" or "darkness," depending on the context.

**Mythological Epithets and Titles:** Deities often have multiple names and titles (e.g., "Shiva" is also known as "Mahadeva," "Rudra," and "Nataraja"), complicating the NER process. The same entity may be presented differently across various contexts within the texts, which necessitates sophisticated algorithms that can discern these variations.

**Overlap Between Entity Types:** Some words can refer simultaneously to persons, places, or events, such as "Meru," which is both a mythical mountain and a symbol of the universe's axis. Distinguishing these contextual meanings is critical for accurate entity classification.

**Broader Impact of NER on Religious and Historical Studies** **Comparative Religious Studies:** NER enables systematic comparisons between different Puranas and other Hindu scriptures, revealing thematic overlaps, theological developments, and the evolving portrayal of deities over time. Such comparative analysis can provide deeper insights into the intertextual relationships and the evolution of religious thought within Hinduism.

Digital Preservation and Educational Applications: The structured data derived from NER can be used to create digital archives, interactive educational tools, and searchable databases, making the Puranas more accessible and engaging for academic study and public exploration. These innovations not only enhance scholarship but also democratize access to ancient knowledge, fostering greater public interest in cultural heritage.

Significance of This Research This research demonstrates the application of NER and OCR technologies as powerful tools for extracting and analyzing named entities within the Puranas. By systematically identifying deities, locations, and events, it provides valuable insights into the mythological, cultural, and theological dimensions of these ancient texts. Furthermore, it addresses the challenges of contextual ambiguities and the complexities inherent in recognizing entities with multiple meanings, enhancing our understanding of the rich narratives that have shaped Hindu thought and tradition for centuries.

In conclusion, the integration of technological approaches, such as NER and OCR, into the analysis of ancient texts like the Puranas not only revolutionizes our engagement with these complex narratives but also underscores the potential for interdisciplinary collaboration between the humanities and computational sciences. By leveraging modern technologies, this research paves the way for a more profound appreciation of the cultural and religious legacies embedded in these timeless scriptures.

## 2. Literature Review

NER, as a subfield of Natural Language Processing (NLP), has seen increasing application in the analysis of historical, religious, and literary texts. Traditionally, NER has been employed in modern datasets like news articles, scientific papers, and legal documents. However, its application to historical and religious texts has opened new frontiers in the digital humanities. This shift is largely driven by advancements in machine learning and the increasing availability of digitized ancient manuscripts.

Several studies have demonstrated the effectiveness of NER in identifying key figures, locations, and events in historical texts. For instance, research on ancient Greek and Roman literature has employed NER to map out mythological figures and historical places, providing a new lens through which to study classical antiquity (Mambrini Passarotti, 2019). Similarly, scholars working on the Bible and other religious texts have applied NER to track references to biblical characters, places, and theological concepts, highlighting the recurring themes in sacred narratives (Bamman Smith, 2012).

### 2.1 Relevance to the Puranas

The Puranas, as a corpus of ancient Indian literature, are filled with references to deities, mythological figures, and sacred places. Applying NER to this body of work offers an innovative method for extracting meaningful information that can be used to better understand the text’s narrative structure and thematic elements. Given the complexity and scale of the Puranas, computational approaches like NER provide a practical way to organize and analyze this vast content.

### 2.2 Computational Approaches to Studying Sanskrit and Indic Texts

In recent years, the field of NLP has turned its attention to non-Latin scripts and classical languages such as Sanskrit, Tamil, and other Indic languages. Researchers have faced challenges due to the unique grammatical structures and rich morphology of these languages. However, efforts such as the development of language-specific models and tools

(Kumar et al., 2020) have laid the groundwork for deeper analysis of ancient Indian texts.

Studies that focus on applying NLP techniques to Sanskrit literature often emphasize the extraction of semantic meaning from texts like the Vedas, Upanishads, and Puranas. For example, Mittal et al. (2020) applied NLP tools to extract verb-centric relationships from Sanskrit texts, highlighting the possibilities for studying the complex syntax and semantic layers present in these works.

## 2.3 Relevance to the Puranas

Given the Puranas' linguistic richness and their role in disseminating religious and philosophical ideas, the application of computational methods to study these texts is crucial. NER, in this context, can help scholars navigate the dense web of mythological narratives, making it easier to track the evolution of stories, characters, and places throughout the texts. Furthermore, tools that are developed for classical Sanskrit can also be adapted to study other Puranic languages and dialects.

## 2.4 Digital Humanities and Indian Religious Texts

The digital humanities field has witnessed a growing interest in digitizing and analyzing religious texts from different traditions. Indian religious literature, particularly the Vedas, Upanishads, Ramayana, and Mahabharata, has been a focal point of this interest. Scholars such as Pollock (2006) have argued for the importance of computational methods in the study of ancient Indian texts, noting that these methods provide a scalable way to handle large volumes of text while preserving interpretive depth.

NER applications in Indian religious texts, while still in their nascent stages, are rapidly gaining traction. For example, the Mahabharata Digital Corpus (MDC) and similar projects are attempting to digitize and apply NLP techniques to large-scale Indian epics, allowing for a more accessible and analyzable format. These projects have shown that computational analysis can enhance traditional philological approaches, offering new insights into the texts' structure and thematic focus.

## 2.5 Relevance to the Puranas

As a key part of the Indian religious canon, the Puranas are ripe for similar digital humanities initiatives. The use of NER in this thesis is aligned with these efforts, providing a bridge between traditional text-based scholarship and modern computational approaches. By extracting and analyzing entities from the Puranas, this study contributes to the growing field of digital humanities research on Indian religious literature.

## 2.6 Puranic Studies and Thematic Analysis

The Puranas have long been studied by scholars of Indian religion and history for their rich mythological content and religious significance. Traditional scholarship on the Puranas, such as the works of Ludo Rocher (1986) and Wendy Doniger (1978), has focused on their narrative complexity, theological teachings, and historical evolution. Rocher’s work, in particular, provides an extensive overview of the Puranic corpus and highlights the interplay between mythological stories and historical traditions.

Recent studies have continued to explore the Puranas’ role in shaping Hindu cosmology, ethics, and philosophy. Thematic analysis of the Puranas often centers around key deities like Vishnu, Shiva, and Brahma, as well as the cosmological and genealogical frameworks that structure their narratives. Scholars have also explored how the Puranas function as both religious texts and historical documents, preserving India’s cultural and social history.

Relevance to NER Analysis: NER provides a novel tool for engaging with the thematic richness of the Puranas. By automatically identifying key entities like deities, sacred locations, and important events, this study offers a data-driven approach to understanding the Puranas’ narrative structure. This computational method complements traditional scholarship by offering new ways to analyze recurring motifs, the roles of deities, and the geographical scope of the narratives.

## 3. Methodology

The methodology for analyzing the Puranas involved a comprehensive and systematic approach that combined Optical Character Recognition (OCR) and Named Entity Recognition (NER) technologies to extract, classify, and analyze named entities from these ancient texts. Given the complexity and richness of the Puranic narratives, this research utilized a multi-step process that included text digitization, extensive text preprocessing, entity recognition, contextual analysis, and validation. This technological approach not only facilitates the extraction of structured information from unstructured text but also enables a deeper and more nuanced understanding of these ancient scriptures.

### 3.1 Text Digitization: Converting Scanned Images to Machine-Readable Text

The first stage in the methodology was the digitization of the Puranas, which are often available only in printed form or as scanned images of manuscripts. To convert these images into machine-readable text, Optical Character Recognition (OCR) technology was employed.

- **\*Use of EasyOCR\***: EasyOCR was chosen for its adaptability in handling diverse print qualities, varying fonts, and scripts found in ancient and regional manuscripts. This tool effectively reads images and converts them into editable text, providing the foundational input for further analysis.

- **\*Challenges in OCR\***:
  - **\*Print Quality and Script Variations\***: The Puranas, written in Sanskrit and various regional languages, often feature ornate scripts, faded ink, and uneven text alignment, posing significant challenges for accurate OCR. These issues required careful handling to minimize errors.
  - **\*Dealing with Historical Scripts\***: Historical texts often contain diacritics, ligatures, and archaic spellings that are not always recognized correctly by standard OCR models. Custom settings and manual interventions were necessary to adjust for these script-specific features.
  - **\*Validation and Correction\***: Following OCR, a manual validation process was undertaken to correct common recognition errors, such as misidentified characters or missing diacritics, ensuring that the resulting text was as accurate as possible.

## 3.2 Text Cleaning and Preprocessing

After digitization, the text underwent extensive cleaning and preprocessing to prepare it for NER analysis. This phase was crucial in enhancing the quality of the data and involved several key tasks:

- *\*Error Correction and Spell Checking\**: Given the prevalence of OCR errors, this step involved correcting misspellings, improper word breaks, and other inaccuracies that could affect the accuracy of NER. Automated spell-check tools, along with manual review, were employed to refine the text.
- *\*Normalization of Language Variants\**: Since the Puranas have been translated and transcribed into multiple languages, standardization of terms was essential. This process involved harmonizing synonyms, standardizing spellings across different translations, and aligning terms with their most commonly accepted forms in Hindu mythology.
- *\*Removal of Artifacts and Unnecessary Characters\**: The text was cleaned of non-relevant elements such as page numbers, footnotes, and special symbols, which could interfere with the NER process. Regex-based techniques were used to systematically remove these artifacts.
- *\*Tokenization and Sentence Segmentation\**: Text was broken down into smaller units such as sentences and tokens, facilitating more precise NER analysis. This segmentation also helped in managing large volumes of text and made the processing pipeline more efficient.

## 3.3 Named Entity Recognition (NER) Application

The core of the analysis involved applying Named Entity Recognition (NER) to the preprocessed text using the Stanza NLP library. Stanza, developed by Stanford, is a versatile toolkit that provides state-of-the-art NLP capabilities, including tokenization, part-of-speech tagging, and NER.

- *\*Configuration of Stanza Pipeline\**: The Stanza pipeline was configured to recognize entities specific to the Puranas, such as persons, locations, mythological events, and deities. This required tailoring the model to understand the context of ancient texts, which differ significantly from modern language usage.

- *\*Model Training and Adaptation\**: - *\*Fine-Tuning on Custom Datasets\**: To improve the model's performance, it was fine-tuned using datasets annotated with examples from Hindu mythology. This included curated lists of deity names, sacred locations, and specific mythological events, enhancing the model's ability to recognize these entities accurately. - *\*Transfer Learning Techniques\**: Transfer learning was applied to leverage pre-existing models trained on related texts, adapting them to the specific linguistic and

contextual requirements of the Puranas. This approach significantly reduced the time and resources needed for model training while improving recognition accuracy.

- **\*Batch Processing and Scalability\***: Given the voluminous nature of the Puranas, NER was implemented in a batch processing framework, allowing large sections of text to be processed in parallel. This scalability was essential for managing extensive texts and ensured that the analysis could be completed in a reasonable timeframe.

### 3.4 Contextual Disambiguation and Entity Classification

A critical component of the NER process was the contextual disambiguation of entities. Many names and terms in the Puranas have multiple meanings, and accurate classification required a deeper contextual understanding.

- **\*Rule-Based Post-Processing\***: Custom rule-based algorithms were developed to refine the initial classifications provided by Stanza. These rules were designed to handle specific ambiguities, such as distinguishing between deities and their associated symbols or between geographical locations and metaphorical uses of the same terms.

- **\*Semantic Analysis for Ambiguous Terms\***: Advanced semantic analysis techniques were applied to assess the context in which an entity appeared, helping to differentiate between different meanings. For example, "Kali" was classified based on its surrounding text as either the goddess, time, or darkness.

- **\*Cross-Referencing with Mythological Databases\***: To validate and enhance entity recognition, the identified entities were cross-referenced with established mythological databases and scholarly resources. This cross-referencing helped confirm the identities of less commonly known figures and provided a layer of verification for frequently appearing entities.

### 3.5 Analysis and Visualization of Extracted Entities

After entities were accurately identified and classified, the next step involved analyzing and visualizing the data to extract meaningful insights.

- **\*Relationship Mapping and Graph Analysis\***: The entities were mapped to visualize their relationships, such as kinship ties between deities, geographic connections between sacred sites, and links between mythological events. Graph-based analysis tools were used to create interactive diagrams that revealed the intricate web of connections within the Puranas.

- **\*Temporal and Spatial Analysis\***: Analysis also extended to understanding the temporal and spatial dynamics of the texts, identifying how certain deities or locations ap-



peared more frequently in specific contexts or timeframes. This helped uncover narrative patterns, such as the rise of certain gods during particular mythological eras.

- *\*Pattern Recognition and Thematic Exploration\**: Clustering techniques were employed to group entities based on thematic similarity, such as war deities, creation myths, or underworld locations. This thematic exploration allowed researchers to delve deeper into specific aspects of the Puranic narratives and compare them across different texts.

## 3.6 Validation, Feedback Loop, and Iterative Refinement

This report details the various cleaning steps applied during the NER-to-text conversion process to refine raw data and ensure its quality for further use. Each step focuses on removing noise, irrelevant information, or redundant content, thus improving the text for downstream tasks such as integration with digital archives and educational tools.

After the NER process, the extracted text often contained characters from non-Latin scripts, specifically in Devanagari, used for Hindi. To ensure the text was consistent and usable for further analysis, it was necessary to clean out non-Latin characters. For example, any Hindi or Devanagari script characters were removed using regular expressions to filter out non-ASCII content. This step was essential for maintaining a uniform text corpus in Latin script, particularly for English-language processing pipelines.

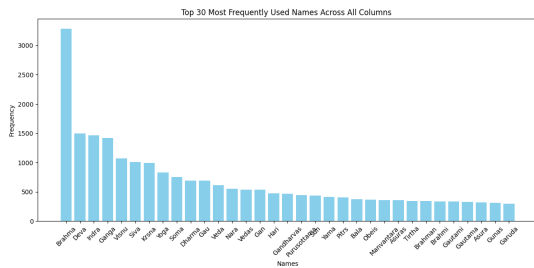
Following the removal of Devanagari characters, short words were removed. These are often non-meaningful words, especially single characters, stray letters, or punctuation marks that add little value to the text. To achieve this, words with fewer than two characters were excluded from the text to enhance the overall quality and relevance of the data.

### Significance of the Methodology

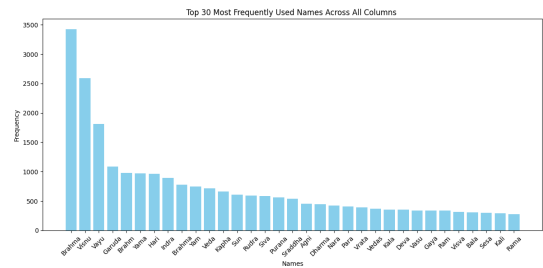
This methodology illustrates how combining traditional scholarship with cutting-edge technology can revolutionize the study of ancient texts. By integrating OCR and NER, this research not only enhances the accessibility and analysis of the Puranas but also demonstrates a scalable and adaptable framework that can be applied to other historical texts. The approach bridges the gap between the humanities and computational sciences, showcasing the power of technology in uncovering new insights into cultural heritage. This interdisciplinary model paves the way for future studies, enabling a more dynamic and interactive engagement with ancient narratives.

## 4. Data Analysis

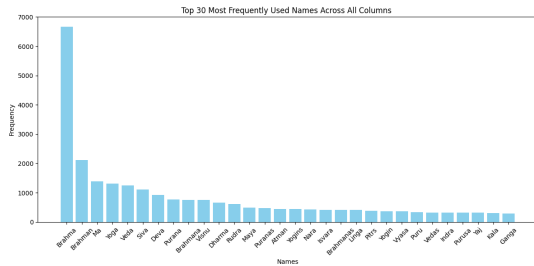
General Insights from the Charts: Each chart presents the Top 30 Most Frequently Used Names in datasets, likely from religious or mythological texts. The focus is on identifying the prominence of names and terms related to key figures, gods, and themes within these datasets.



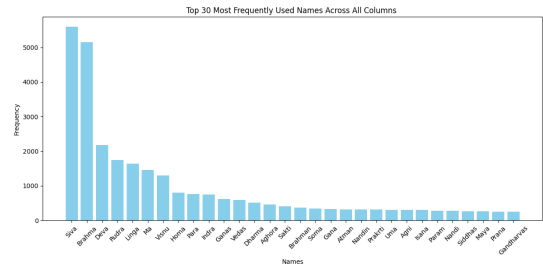
(a) BRAHMA



(b) GARUDA



(c) KURMA



(d) LINGA

Figure 4.1: Multiple Images Displayed Side by Side

### 4.1 Most Frequently Mentioned Entities

Siva, Brahma, and Visnu are consistently among the top names. These three deities are central to many Hindu scriptures, underscoring their theological importance. Deva and Rudra are also highly frequent, suggesting the repeated reference to gods and various manifestations, especially of Siva (Rudra being an earlier form). Secondary Frequently Used Names:

Linga (related to Siva worship), Homa (a ritualistic fire offering), and Vedas (the sacred texts) appear regularly, indicating the importance of ritual practices and texts

in the contexts these datasets are drawn from. Names like Ganesa, Dharma, and Sakti highlight the inclusion of figures and concepts related to morality, ritualistic power, and divine energies. Names with Moderate Frequency:

Soma (a ritualistic drink or deity), Agni (fire god), and Indra (king of the gods) appear in some charts with moderate frequency, emphasizing their roles in Vedic rituals and mythology. Atman (the concept of the soul) and Prakrti (nature) indicate philosophical references, adding depth to the texts beyond just divine entities. Philosophical and Textual References:

Dharma, Atman, and Param refer to spiritual concepts, showcasing the philosophical nature of some texts. Vedas and Puranas appear, underlining the recurring references to the foundational scriptures of Hinduism. Summary: The charts display a strong presence of Hindu deities, ritual terms, and philosophical concepts, revealing the religious, mythological, and spiritual nature of the datasets. The frequent mention of key figures such as Siva, Brahma, and Visnu suggests that the texts may revolve around their stories, worship, or related rituals. Moreover, the appearance of concepts like Dharma, Atman, and Prakrti points to the philosophical richness present in these texts.

## 5. Results

The application of Named Entity Recognition (NER) to the Puranic texts yielded a comprehensive extraction and classification of various named entities. These entities were systematically categorized into distinct types such as persons, locations, organizations, and geopolitical entities, which allowed for a structured exploration of the narratives contained within the Puranas. The results highlight the effectiveness of NER in unraveling the intricate network of figures, places, and events embedded in these texts, thus providing a new lens for examining the cultural, religious, and historical aspects of Hindu mythology.

### 5.1 Key Findings from the NER Process

1. **\*Comprehensive Categorization of Extracted Entities\*:** - The NER model effectively categorized entities into several types, providing a detailed classification that aligns with the narrative structure of the Puranas. **Person Entities:** A significant portion of the identified entities were persons, predominantly deities, sages, kings, and legendary figures central to Hindu cosmology. The identification of entities like Brahma, Vishnu, Shiva, Saraswati, and Parvati not only captured their roles but also contextualized their interactions and significance within the larger narrative framework. **Location Entities:** The recognition of key locations such as Kailash Mountain, Vaikuntha (the abode of Vishnu), and various sacred rivers like Ganga and Yamuna provided a geographical and spiritual map of the Puranic universe. These places are more than mere backdrops; they are active participants in the narratives, each carrying symbolic and theological significance. - **\*Mythological Objects and Artifacts\*:** Entities such as divine weapons (e.g., Sudarshana Chakra), sacred trees (e.g., Kalpavriksha), and mythical animals (e.g., Garuda) were identified, adding depth to the understanding of the mythological elements that populate the Puranic stories. - **\*Events and Rituals\*:** The NER process also recognized references to significant events like cosmic battles (e.g., the churning of the ocean), sacred rituals, and yagnas, providing insights into the ceremonial aspects of the Puranic traditions.

## 5.2 Cultural and Religious Insights from Extracted Entities

- *\*Understanding Divine Roles and Relationships\**: The analysis provided a clearer picture of the complex roles and relationships among deities, such as the Trimurti (Brahma, Vishnu, Shiva) and their consorts. This enabled the mapping of divine hierarchies and the identification of recurring motifs, such as the theme of divine intervention in human affairs. - *\*Symbolic Significance of Locations and Objects\**: Sacred places and objects, identified through NER, were shown to have layered meanings, representing physical, spiritual, and cosmic elements. For example, Mount Meru, frequently mentioned in the texts, is not only a geographical entity but also a cosmic axis and a symbol of stability and divine presence. - *\*Genealogical Mapping of Mythological Figures\**: The extraction of genealogical lines highlighted the connections between various deities, sages, and royal families. These mappings help in understanding the inheritance of divine qualities, powers, and responsibilities, offering a structured view of mythological heritage.

## 5.3 Thematic Patterns and Recurrences Across Texts

- *\*Identification of Core Themes\**: The systematic categorization of entities allowed for the identification of core themes prevalent across the Puranas. Themes such as creation, destruction, divine play (lila), and cosmic order were frequently associated with specific deities and events, revealing underlying narrative patterns. - *\*Comparative Analysis of Depictions\**: By comparing how certain deities and locations were portrayed in different Puranas, the research highlighted variations in emphasis and interpretation. For example, Vishnu's avatar narratives are more detailed in the Bhagavata Purana compared to the Vishnu Purana, reflecting a shift in theological focus over time.

## 5.4 Challenges and Complexities in Entity Recognition

- *\*Handling Ambiguities and Dual Roles\**: Entities often served multiple roles, creating challenges in classification. For instance, the term "Nandi" could refer to the bull of Shiva or a divine attendant, depending on the context. The results showed the need for advanced contextual analysis to resolve such ambiguities effectively. - *\*Recognition of Mythological Epithets and Synonyms\**: Deities with multiple names and titles, such as Krishna being referred to as Madhava, Govinda, and Keshava, posed challenges. The NER system's ability to recognize and group these variations was critical for consistent

analysis. - *\*Overlapping Categories\**: Some terms were difficult to classify due to their overlapping meanings. For example, "Indra" could refer to the king of gods, a generic term for a ruler, or a metaphorical representation of power. The results underscored the importance of context-aware models that can dynamically adjust to these nuances.

## 5.5 Broader Implications of the NER Results on Puranic Studies

- *\*Enhanced Textual Analysis\**: By converting unstructured Puranic narratives into structured data, NER facilitated a more nuanced textual analysis. Scholars could trace the evolution of certain themes, identify shifts in theological focus, and explore the changing portrayals of key figures over time. - *\*Digital Archiving and Preservation\**: The structured entity data supports the creation of searchable digital archives, making the Puranas more accessible to researchers and the public. This also aids in the digital preservation of ancient knowledge, ensuring that these texts remain available for future generations. - *\*Educational Applications\**: The results provide a foundation for developing educational tools such as interactive maps, genealogical trees, and thematic timelines that can engage students and the general public with Puranic narratives in an immersive and informative way.

## 5.6 Impact on Comparative Mythological and Religious Studies

- *\*Cross-Textual Comparisons\**: The structured entity data allows for systematic comparisons not only within the Puranas but also with other ancient texts, such as the Vedas and Upanishads. This can reveal shared motifs, narrative borrowings, and unique developments within Hindu mythology. - *\*Interdisciplinary Research\**: The data can be used in interdisciplinary research, linking mythology with historical, linguistic, and socio-cultural studies. For instance, identifying the prominence of specific deities in different historical periods can offer insights into the sociopolitical contexts of those times.

## 5.7 Technological Advancements in Textual Analysis

- *\*Demonstrating the Utility of NER for Ancient Texts\**: This study illustrates the potential of NER as a transformative tool in the digital humanities. It showcases how advanced NLP techniques can bring structure to complex ancient narratives, enabling new forms of analysis that were previously inaccessible. - *\*Future Directions in Model*

Development\*: The challenges encountered, such as contextual ambiguities and overlapping categories, point towards the need for more refined NLP models that are specifically tailored to ancient and mythological texts. This research lays the groundwork for developing these specialized models, pushing the boundaries of what can be achieved with NER in historical text analysis.

## 5.8 Conclusion

The NER analysis of the Puranas has provided substantial insights into the complex web of deities, locations, and events that characterize these ancient texts. By systematically extracting and categorizing key entities, this research not only deepens our understanding of Puranic narratives but also highlights the transformative potential of applying technology to ancient text analysis. The findings underscore the value of integrating computational methods with traditional scholarship, paving the way for new explorations of cultural and religious heritage through a technological lens. This approach not only preserves the richness of these narratives but also makes them more accessible, engaging, and analyzable for future generations.

## 6. Discussion

There are several existing Named Entity Recognition (NER) models, each with strengths in various domains, but they often fall short when applied to specialized and ancient texts like the Puranas. Below is a list of some of the widely used NER models, along with reasons they are not fully sufficient for the Puranic texts and why our method stands out.

### 6.1 SpaCy

- **\*Strengths\***: SpaCy is an industrial-grade NLP library with built-in support for various languages. It is highly optimized for modern languages and provides fast tokenization and NER models for recognizing common entities like PERSON, ORG, and LOC. - **\*Limitations for Puranic Texts\***: - **\*Modern Language Focus\***: SpaCy’s pre-trained models are designed for contemporary texts such as news articles or Wikipedia. They struggle to recognize historical or mythological entities. - **\*Lack of Specialized Entities\***: SpaCy does not have built-in models for recognizing specific categories like deities, mythological events, or sacred places, which are common in the Puranas. - **\*Handling of Multiple Scripts\***: Puranas are often written in Sanskrit or other regional languages like Tamil or Devanagari, which SpaCy’s NER models are not trained for.

### 6.2 Stanford NER

- **\*Strengths\***: Developed by Stanford NLP, this model is a popular choice for academic research. It supports multiple languages and is relatively easy to train on custom datasets. - **\*Limitations for Puranic Texts\***: - **\*Lack of Domain-Specific Training\***: While it supports custom training, the default model is not trained on ancient or mythological texts, making it less accurate in identifying specialized entities. - **\*Context Sensitivity\***: Stanford NER models are not highly optimized for contextual disambiguation, which is crucial when dealing with polysemous terms like “Kali” (which could mean the goddess, time, or darkness depending on context). - **\*Resource Intensive\***: Stanford NER tends to require significant computational resources and time to fine-tune, which can be a challenge when handling large corpora like the Puranas.



### 6.3 Hugging Face Transformers (BERT-based models)

- *\*Strengths\**: BERT and its variants, such as RoBERTa and DistilBERT, have become the state-of-the-art in NLP, with powerful contextual embeddings and pre-trained models for various tasks including NER. - *\*Limitations for Puranic Texts\**: - *\*Modern Dataset Bias\**: These models are generally pre-trained on large, modern datasets such as Wikipedia, news, and books, and are not suited for ancient, religious, or mythological content without significant retraining. - *\*Limited Handling of Historical or Religious Language\**: While BERT excels at understanding modern contexts, its performance degrades when applied to historical texts like the Puranas that involve archaic language, varied dialects, and culturally specific terminology. - *\*Training Complexity\**: While Hugging Face models are highly flexible and can be fine-tuned, they require extensive computational power and large domain-specific datasets to adapt to the Puranic context effectively.

### 6.4 Flair NLP

- *\*Strengths\**: Flair provides a framework for sequence labeling tasks and supports contextual string embeddings, which are particularly useful for NER. It is known for its ability to capture more nuanced relationships between words. - *\*Limitations for Puranic Texts\**: - *\*Focus on General Entities\**: Flair's pre-trained models focus on general entities (e.g., people, locations, organizations) and are not trained for recognizing mythological or religious entities specific to Hindu texts. - *\*Insufficient Historical Linguistic Adaptation\**: Flair's embeddings, while powerful for modern texts, do not capture the complexities of ancient languages and scripts, which often include diacritics, archaic forms, and script variations that are frequent in Puranas.

### 6.5 Polyglot NER

- *\*Strengths\**: Polyglot is a multilingual NER model that can handle multiple languages, making it useful for global or multi-lingual datasets. - *\*Limitations for Puranic Texts\**: - *\*No Specialized Models for Sanskrit or Regional Indian Languages\**: Although Polyglot supports multiple languages, its models for Indian languages like Sanskrit, Tamil, or Devanagari are not optimized for ancient scriptural contexts. - *\*Difficulty Handling Mythological Contexts\**: Polyglot models are designed to identify common, modern entities and are not well-suited to handling highly specialized domains like Hindu mythology.

## 6.6 Why These Models Are Not Sufficient for Our Use Case

- **\*Lack of Domain-Specific Training\***: Most existing models are trained on modern text corpora, such as news or web articles, and do not understand the mythological, historical, or religious context of the Puranas. They do not have the ability to recognize specific named entities such as Hindu deities, sacred locations, and mythological events. - **\*Handling of Multiple Scripts\***: Many of these models are designed for a single language or modern multilingual datasets, whereas the Puranas are written in a mix of Sanskrit, regional Indian scripts (like Devanagari, Tamil), and often feature diacritics and ligatures that are absent in modern languages. - **\*Contextual Ambiguity\***: The models mentioned above lack the ability to effectively disambiguate terms that have multiple meanings, a common issue in ancient texts. For instance, terms like “Shiva” could refer to the god, a devotee, or a symbolic concept in the Puranas, and without contextual understanding, general models will misclassify these terms. - **\*Lack of Cultural and Historical Sensitivity\***: Most models are not trained on religious or historical corpora and hence fail to grasp the unique cultural significance of entities in mythological texts. They are not designed to handle religious or philosophical concepts integral to the Puranas.

## 6.7 Why Our Method Stands Out

- **\*Fine-Tuning on Custom Datasets\***: Our method involves fine-tuning the NER models on a custom dataset specifically curated from Hindu mythology. This ensures that the model is familiar with the unique entities, terms, and relationships found in the Puranas, such as deities, specific places, and mythological events. - **\*Contextual Disambiguation\***: We have developed rule-based post-processing algorithms and semantic analysis techniques to disambiguate entities based on their contextual usage. For example, our method can accurately differentiate between “Kali” as a goddess, the concept of time, or darkness based on the surrounding text. - **\*Handling Multiple Scripts and Languages\***: Our methodology has been adapted to process multiple Indian scripts such as Devanagari, Tamil, and Sanskrit, allowing us to accurately digitize and analyze texts regardless of the script variations. - **\*Cross-Referencing with Mythological Databases\***: To enhance the accuracy of entity recognition, we cross-reference the identified entities with mythological databases and scholarly resources, ensuring that even obscure or lesser-known entities are correctly classified. - **\*Scalability and Parallel Processing\***: Our method is designed to process large volumes of text in a scalable manner. We implemented batch processing to manage the vast corpus of the Puranas, ensuring that the analysis is efficient without compromising accuracy. - **\*Iterative Refinement\***: Our process includes continuous

validation, feedback loops, and iterative model adjustments based on error analysis and expert review. This ensures that the final model is highly optimized for Puranic texts, addressing the specific challenges posed by these ancient and complex scriptures.

In conclusion, while existing NER models are powerful for general use cases, they lack the domain specificity, linguistic adaptability, and contextual sensitivity needed for analyzing ancient Puranic texts. Our customized approach, fine-tuned for the unique characteristics of these scriptures, provides a more accurate and reliable solution for extracting meaningful insights from these rich, historical narratives.

## 7. Conclusion

This thesis applied advanced Named Entity Recognition (NER) techniques to analyze the rich textual tradition of the Puranas. The Puranas, being a cornerstone of ancient Indian literature, are filled with complex narratives, diverse characters, and detailed geographical and cosmological descriptions. By leveraging Natural Language Processing (NLP) methods, particularly NER, this study aimed to extract meaningful insights from these ancient texts, which are known for their narrative depth and religious significance.

The NER analysis revealed a detailed network of persons, locations, and events that are central to the Puranic narratives. Specifically, entities such as [Shiva, Vishnu, Brahma] and places like [Mount Kailash, Vaikuntha] frequently appeared, reflecting the theological and cosmological focus of the Puranas. This automated identification of entities allowed for a structured understanding of the frequency and importance of key figures and places in the texts, thus providing a quantifiable approach to exploring the narrative structure and thematic content of these works.

### 7.1 Methodological Insights

The approach detailed in this thesis utilized tools such as easyocr and stanza for text extraction and named entity recognition, respectively. The accuracy of entity extraction was enhanced by these tools, allowing for a more refined categorization of deities, sacred locations, and mythological events central to the Puranic tradition. This methodological framework enabled a systematic analysis that would have been challenging to accomplish through manual methods alone, particularly given the vastness of the Puranic corpus.

By converting scanned images into machine-readable text, followed by NER-based analysis, the study was able to identify patterns that shed light on the theological and cultural fabric woven into the Puranas. The top 30 entities identified in the analysis were visualized through graphs, providing a clear, data-driven representation of the narrative focus of these texts.

## 7.2 Dominance of Theological Entities

As expected, the analysis showed that the majority of named entities pertain to deities like Shiva, Vishnu, and Brahma, reflecting the primary religious focus of the Puranas. Their repeated mention underscores the central role these figures play in the mythological and devotional framework of the texts.

## 7.3 Geographical Significance

Locations such as Mount Kailash, Vaikuntha, and other sacred spaces were frequently mentioned, illustrating the cosmological importance attributed to certain geographical regions in Hindu mythology. These places are often the settings for pivotal events, highlighting their symbolic and spiritual significance.

## 7.4 Recurrent Mythological Events

The NER approach helped uncover key events, such as the cosmic battles, the creation myths, and the divine interventions, which are central to the narratives of the Puranas. Events like the churning of the ocean (Samudra Manthan) or the battles between deities and demons form recurring themes that underscore the eternal cosmic struggle and balance.

## 7.5 Implications for Puranic Studies

The application of NER to the Puranas opens up new avenues for the analysis of ancient texts. Traditionally, the study of these texts has relied on qualitative methods of interpretation. However, by integrating computational techniques like NER, this research has demonstrated that it is possible to systematically uncover the underlying structures of these texts, offering a clearer understanding of their key themes and central figures.

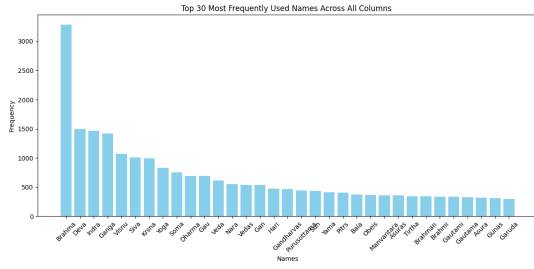
The results of this thesis underscore the potential for further computational studies of ancient texts. By expanding the NER framework to include other types of entities, such as philosophical concepts or rituals, future research could deepen our understanding of how the Puranas articulate not only the myths and deities of Hinduism but also the social, cultural, and religious practices embedded within them.

In conclusion, this thesis has shown that NER is a powerful tool for the analysis of religious and mythological texts like the Puranas. The automated extraction of key figures, locations, and events has provided a structured and data-driven approach to understanding the complex narratives of the Puranas. This methodology offers a new

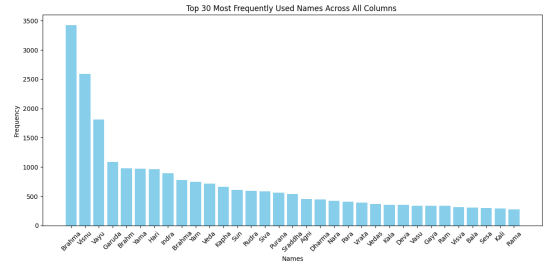
perspective for scholars of ancient literature, enabling them to analyze large texts with a level of precision and scalability that was previously unattainable.

Future research could build on this foundation by incorporating additional NLP techniques, such as sentiment analysis or topic modeling, to explore the emotional tone or thematic progression of these texts. Additionally, expanding the analysis to other Hindu scriptures or cross-textual studies could offer comparative insights into the broader corpus of ancient Indian literature.

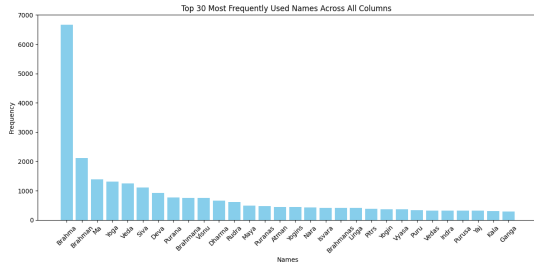
## 8. Appendix



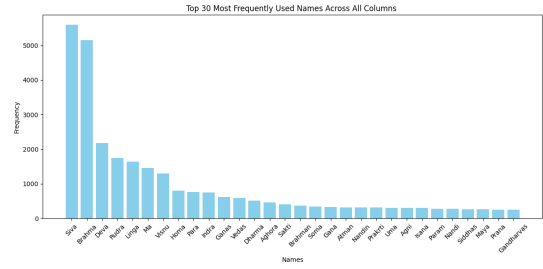
(a) BRAHMA



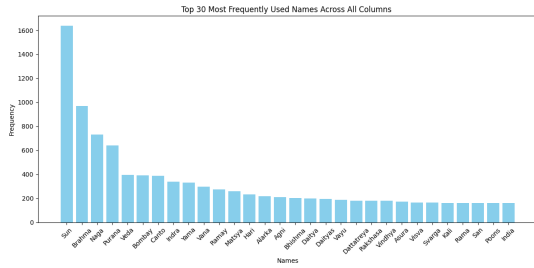
(b) GARUDA



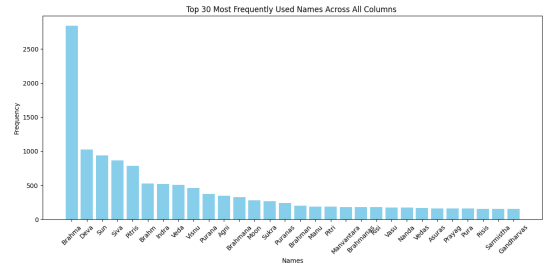
(c) KURMA



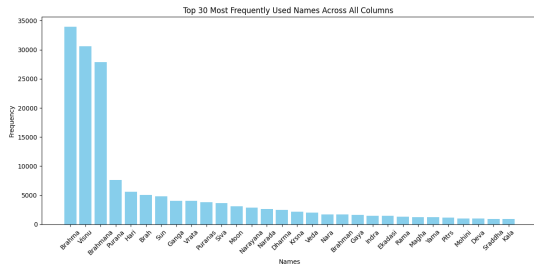
(d) LINGA



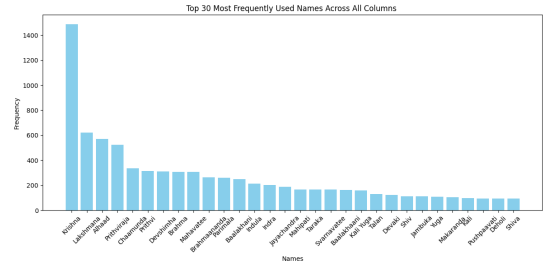
(e) markandey



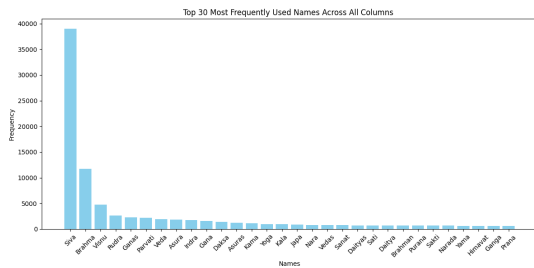
(f) atsyā



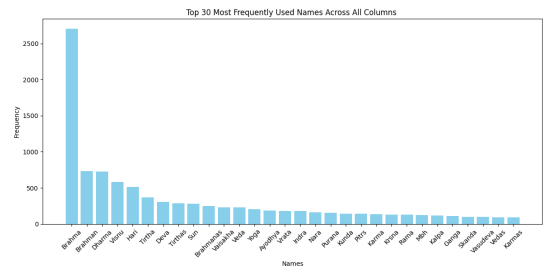
/captionnarad



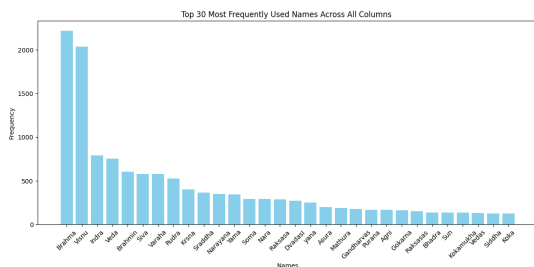
(g) pratisgara



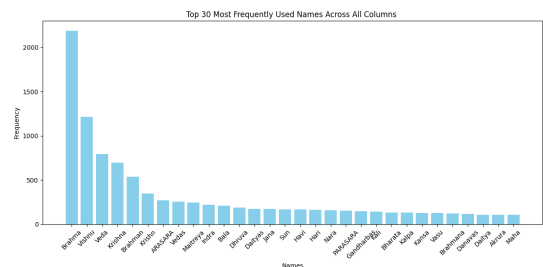
(h) shiv



(i) skandha



(i) varna



(k) vishnu



## Vishnu Puran Sample o/p

1	PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
2	Benuka	ADITI	Agastya	Abiksheta	Acvinas	Amavasya
3	Abhimani	Airavata	Agni	Achyuta	Adityas	Brahma
4	Abhjejt	AITREYA	Alakakanda	Adhidvaika	Airavata	BrahmaSavarni
5	Abhimani	ANASARA	Amazon	Adhomukha	Asaras	Divapara age Veda
6	Abhinanyu	Anasuya	Andithe	Adityas	Aptoryama	Divapara Yuga
7	Abhinaman	Ansa	Atratra	Agnibahu	Asita	Great Asura
8	Abhras	ARASARA	Bay	Agnistoma	Asuras	Great Hari
9	Abhutarajasas	Atharva Veda	Bay Bengal	Agrahayana	Bhuma	Great undecaying Vishnu
10	Abhyuthitawasa	AtharvaVeda	Bhagirathi	Ahanikara	Boar	Great Vishnu
11	Abikoluta	AURVA	Avanti	Airavata	Brahma	Kali Yuga
12	Accoutt Serpent Sesha	Avanti	BharataVarsha	Ajabithi	Brahmana	Knowledge Action Punishment
13	Achuta	Avira	Bhava	Ajamidha	Clepydra	Magha
14	Achyuta	Brahman	Bhurloka	Ajita	Danavas	Manvantara Auttama
15	Acvini Kumara	Bramha	Bhuvarloka	Ajmere	Destroyer	Manvantara Chakushusha
16	Adharma	Chapa	Brahma	Akura	Dhruva	Manvantara Raivata
17	Adhiseemakrishna	Darpa	Brahmaloka	Alaka	Ekdodistha	Manvantara Swarochisha
18	Adhwanyu	Ekdodistha	Brahman	Alakananda	Gandharvas	Manvantara Uttami
19	Adhyatmika	Elysium Press	Brama	Amaravati	Garuda	Paksha
20	Adina	Gandharbas	Budha	Ambasthas	Guyatri	Paura
21	Aditratha	Hastinapur	Capricorn	Amruttaya	Jaitra	Raivata Manvantara
22	Aditi	Ita	Chaitrarath forest	Anadhristi	Kala	Sayambhava Manvantara
23	Aditi	Ichama	Danavas	Anagha	Kallage	SECTION IV Narayana Kalpa
24	Aditya Dhatri	LARASARA	Danu	Anala	Mahakalpa	Swayambhava Manvantara
25	Accordingly Balarama	LURVA	Dhruva	Anamitra	Manus	Tarakamiya Taraka war
26	Agastya	LURVA	Dwapara	Anchileya	Maruts	Treta yuga
27	Agneyi	MAITYCIA	Dwipa	Andhaka	Matra	Tretaj yuga
28	Agni	MATREYA	Foulness	Andhatamisra	Mercury	Vaisakha
29	Agnidhra	Menu Swayambhava	Gandhamadana mountain	Andhra	Muhurta	Viaavasya Manvantara Brahma
30	Agnidhras	Munis	Gandharbas	Anga	Muhurtas	Vishnu Purana
31	Agnikuluma	NIDAGIA	Gandharva Tumburu	Angraka	Nadhika	YajurVeda
32	Agnimathara	PAKASARA	Ganga	Angrisc	Padma	Yoga

## Varna Puran Sample o/p

1	PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
2	Abhava	ABORI	Agni	Acchodaka	Agasti	Amavasya
3	Abhira	Amala	Amalaka	Achyuta	Adlesa	asterism Purvabhadrapada
4	Acamana	Bhandarkar	Aspara	Adharma	Asoka	Day Night
5	Acchodaka	Oriental Research Institute	Ardrhacandra tirtha	Aditya	Asura	Fortnight
6	Acyuta	Adityas	Arjuna	Adityas	Asvamedha	Month
7	AD Pusalkar	Agastyagita	Asadhoma	Afijana	Belva	Year
8	Adhama	Agnistoma	Asikunda	Agneya	Cane	DvadasI
9	Adhara	Alakananda	Asita	Agnidrahstra	Canopus	Dvapara yuga
10	Adideva	Asikunda	Asitod	Agnihotra	Citrastila	Ekladasi
11	Adilyatirtha	Astami	Asuras	Agnihotrastala	Dundubhi	eleventh Bhandkravana
12	Aditi	Asura	Atharva	Agniprabha	DvadasI	Indian Translation Series
13	Aditya	Bharata	Avanti	Agnisaras	DvadasI	Karttika festival
14	Aditya	Bharatiya Vidya Bhavan	Ayana	Agnisatyapada	Ekladasi	Magha
15	Adityas	Bombay	Ayanas	Agnitirtha	Ether	Navami
16	Agaru	Brahma	Badari	Agnivesya	Gakra	Pausa
17	Agasti	Brahma Valvarta Purana	Bhadra	Ahanikara	Garuda	Vasus
18	Agastya	Brahmin Suta	Bharata	Ahanyanya	Isavasya	
19	Agastyagita I Varaha	Brhaspati	Brahma	Aligiris	Kamandalu	
20	Agastyagita	BVB	Brahmaloka	Airavata	Kaumdra	
21	Agnyai	Caritrakola	Brahmasaras	Akhyavats	Krosodaka	
22	Agni	CarmanvatI	Brhaspati	Akura	Krsna	
23	Agnibahu	Caturdasi	Citrakuta	Aladhureya	Kumuda	
24	Agnidamstra	Citra	Deccan	Alakananda	Kurmas	
25	Agnidatta	Citragupta	Dhruvatirtha	Amara	Mercury	
26	Agnihotra	CityofYama	Durga	vati	Nilikas	
27	Agnij Asvins	Contd	Durva grass	Amarakantaka	Pada	
28	Agniman	CVS	Dvandvas	Amaravati	Paticaratrgama	
29	Agnimedha	Daivika	Dvipas	Amboadhara	Pasupati	
30	Agniprabha	Daksina	Ganas	Amrta	Prakrti	
31	Agnisaras	Dhanyadhenu	Gandaki	Ananta	Saktis	
32	Agnisatyapada	DvadasI	Ganga	Anasaka	Santipdtha	
33	Agnistoma	Gandharvas	Ganges	Andhakara	Satum	

## Skandha puran sample o/p

1	PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
2	A Stein	AstangaYoga	Aisvarya	Acyuta	Airavata	Aksaya
3	AB Keith	Asvamedha	Amavasya	Adhivedata	Amrta	Angaraka Caturthi
4	Abhaya	Avanti	Anavamala	Adityas	Apana	Asrama
5	Abja	BG X	Aries	Agnihotra	Asat	Bharata War
6	Acamaniya	BHP	Asana	Ahamkara	Asokas	Brahma
7	Acchoda	Brahmanas	Ayodhya	Airavata	Asrama	Brahman
8	Acyuta	Buddhism & Jainism	Bakula	Ajna Cakra	Astanga	Caitra
9	Adharma	Candala	Bhurloka	AksaraPurusas	Asthapura	Chapter Chapters
10	Adhisima Krsna	Cedi	Bhuvarloka	Amalaka	Bakulas	DvadasI day Vaisakha
11	Adi	Cellibates	Brahma	Amaravati	BG X	Dvapara Yuga
12	Aditi	Chastiser Madhu	Brahmaloka	Amrta	Brahma	Ekladast Vrata
13	Aditya	Citragupta	Brahmanda	Amu	Brahmandas	Festival Laksmi
14	Adrika	Cosmoses	Brahmanical Paramaha	Anarta	Cakratirtha	Great Departure
15	Agastya	Cyavana	Brahmapura	Anasuya	Candala	Japa
16	Agni	Danavas	Candrasahasra Vrata	Angras	Candrayana	Jatakunda Rama
17	Agnihotra	Dvipas	Chambal	Anjana	Cosmic Egg	Kalpa
18	Agnivotta	Ekanitika Dharma	Chenab	Antahkarana	Damana	Kalpa Adikalpa
19	Agnivattas	Elephants	Cosmic Egg	Anusasana	Discus	Kalpa Dharma
20	Ahalya	Chariots	Dandaka forest	Aranyakas	Dvipa	Lunar
21	Ahamkara	Cavalry Infantry	Dharma	Arbuda	Ego	Magha
22	Ahankara	Gayas	Dhyana	Arghya	Goblin	Muhurta
23	Ahimca	HARI	Durva grass	Arundhati	Gunas	Navaratri
24	Ahumkara	Homa	Dvipa	Aruni	Hari	Navaratri Festival
25	Ahuti	Ilili	East	Asadha	Havisa	Para
26	Aihiki	Jnana Yoga	Ekanitika Dharma	Asita	Jambira	Purnima
27	Airavata	Kala	Emblic Myrobalan	Brahmana	Jambu	Solar Year
28	Aja	Karmas	Ganga	Asoka	Kalakuta	Spring Season
29	Ajita	KB Davale Pub Bombay	Garuda	Asvalayana	Kalasa	Sukla Yajurveda
30	Ajnacakra	KriyaYoga	Ghosarka Tirtha	Asvattha	Kalasis	Sun Moon
31	Akasa	Lotus	Godavari	Asvina	Karandava	Tirtha Sahasradhara
32	Aksara	Mahapurusa	Gopratara Tirtha	Atharvaveda	Karma	Tirtha Vira
33	Aksara Purusas	Mantresvara	Govardhana mountain	Audumbara	Karmas	Treta Yuga Dharma
34	Aksarasadhanan	Mbh	Heavenly Ganga	Aurangabad	Karmikaras	Upavasa
35	Aksaya	Mbh Saunaka	Himalaya	Avantika I Puri	Liberator	Vaisakha

## Shiv puran sample o/p

1	PERSONS	ORGANIZATION	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
2	A G	Bhava	Acamana	Abhicarka	Acamana	Aghora inantra
3	A U M	Acamana	Acchoda lake	Abhira	Acit	Asura Mahisa
4	Abhasvaras Adityas	Acarya	Adhvas	Abhiseka	Adityas	Asura Taraka
5	Abhavayoga	AIHT	Aeon	Abu	Aghofastra	Avarana
6	Abhayamudra	Aisvarya	Africa	Acamana	Aghora	Avarana Puja
7	Abhicara	Aksauhini	Agni	Acamaniya	AgniI	Bharata war
8	Abhicarka	Alpa	Ahimsa	Acchavaka	Agnisoma	BHD Sec III
9	Abhicarya	Amba	Alaka Nanda	Acvara	Agniyajna	BhismaYudhisthira
10	Abhijit	Amogha	Amaravati	Adhaka	Aguru	Brahma Muhurta
11	Abhilasa	Andhaka	Amba	Adhara	Ahutis	Brahmacarya A'Yama
12	Abhilasastaka	Angiras	Amrta Amrta Soma	Adharma	Aksa	Caksusa Manvantara Daksa
13	Abhimanan Prayina	Anilas	Anabian Pennsula	Adharmamahisa	Aksatas	Caturdasi
14	Abhimukha	Anima	Anarta Desa	Adhidaivika	Amrta	Caturmasya festival
15	Abhirama	Anugraha	Anima	Adhikara	Anala	Dvapara Yuga
16	Abhiru	Apa	Antaratin	AdhvarayPratipratishatr	Ant	Great Charity
17	Abhiseka	Apara	Anugraha	Aditi	Aratnis	Great Dissolution
18	Abhramu	Aparna	Anani	Aditya	Arghya	Guha birth
19	Abu	Apomurti	Arbuda mountain	Afghanistan	Arjuna	Japa
20	Acala	Apya Yogins	Arctot District	Agarha	Asat	Japa Gayatri mantra Arghya
21	Acalesa	Arca	Ardra	Agastyas	Asramas	Japa Rudragita
22	Acalesvara	Arcis	Asana	Aghora	Astanga	Kailasa
23	Acamana	Arghya	Asana Darbha	Agneya	Astra	Kali Yuga
24	Acamaniya	Arunoda	Asi	Agni	Astramantra	Krta age Dvaipayana
25	Acancala	ASR Vol	Asia	Agnidhra	Astraraja	Krta Yuga
26	Acarya	Assembly	Asura	Agnihotra	Asura	Kurma Purana
27	Acetana	Astra	Asura Dusana	Agnisoma	Asuras	I War
28	Acintya	Astramantra	Asura Jalandhara	Agnistoma	Asvatthaman	Magha
29	Acitya	Asura Formerly	Asura Sumbha	Agnisvara	Atmans	Mahabharata war
30	Adaksina	Asvi	Asura Taraka	Agniyajna	Avarana	Manvantara
31	Adhaka	Atharvaveda	Asuras	Aguru	Avimuktasvara	Manvantara Satya Manu
32	Adhaki	Atman Prajna	Asuras Taraka	Ahalya	Badava	Marga

## Pratisgara Puran sample o/p

1	PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
2	Abde	Brahma Astra	Alha heartland	Aandhra	Agneya	Chandrabaar
3	Abhaya	Kabutaar	Arabian Sea	Abhinandana	Astra	Hemanta
4	Abhaya	Kauravamsas	Arya Dharma	Adiketu	Bell	Holi
5	Abhayashimha	Lakshmana	Atharva	Agnivamsas	Chariots	Jesta
6	Abhimanyu	Mahavatee	Banyan	Agnivansh	Mahavatee	Kali Yuga
7	Abhinandana	Prithviraja	Battlefield	Agra	Mlecchas	Kaliyuga
8	Abhinandana	Sun	Baudhashimha	Agrayaajak	Moolsharma	Kurukshetra war
9	Abhiri	Udup	Bhadrapada	Aheerij	Parrot	Maya war
10	Achamer	Valla Astra	Bhavishya Purana	Ajamer	Saturn	Shuklapaksha
11	Adharma	Vrij	Brahma	Ajmer	Shabree	Treta Yuga
12	Adivayamkara	Yoga Nidraa	Brahma Loka	Alhaad	Suryavarma	War Mahismati
13	Agahan	Kapota	Brindavan forest	Alolup	Svarnavatee	
14	Agama	Vela	Coast Narmada River	Anaadhrishta	Taraka	
15	Agneya		Coast Sindhu River	Andhra pradesh	Trident	
16	Agni		Desert	Anga	VallaAstra	
17	Agni Dev Gajasena		Deshraaj	Angda	Varuna	
18	Agni deva		Dwapara yuga	Antarvedee	Varunastra	
19	Agni deva Netrashimha		east Kapilashram	Aparaajita	Vayabya	
20	Agni Vamsa		Gandaki River	Arya Desh	Vayabya Astra	
21	Agni Vansh		Ganges	Aryadesa	Vayabyastra	
22	Agnihotri		Gulf Kach	Aryadesha	Veervardhan	
23	Agnivaan		Himalaya	Aryashimha		
24	Agnivamsas		Himalayas	Aryavarta		
25	Aheerin		Isha river	Asadha		
26	Ahlaad		Ishta region	Ashtaamee		
27	Ailavilin		Jupiter	Avantee		
28	Airavaat		Kailash Mountain	Ayocbaahu		
29	Ajodhya		Kali Yuga	Ayodhya		
30	Alf Hiltbeitel		Kalpakshetra	Baalakhaani		
31	Alhaad		Kapila	Baalakhani		
32	Alhaad		Kapilaashram	Baalheek		

## Narad Puran Sample o/p

1	PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
2	Adam	Angirasa	Anguttara Nikaya	Ahrubudhya	Arkapatavratra	Abhijit
3	Adhusa	Aperuka	Asanykas	Alaka	Arvidya	Agni Purana
4	Ajuthot	Bharatiya Vidya Bhavan	CandraTarabala	Arkaka	Jorian	Aitareya Brahmana
5	Agnis PurndraCh	Brahma	Bhurja patra	Antarveda	Doodasi	Akhyana Suktas
6	Aisvarya	CCVrata	Brahmarandhra	ApteSk	Kart	AksayaTraya day
7	Akriyat	Anandasrama Poon	Central	Asadha	Arckisvara	Amavasya
8	Acyuta	Candra	Anandasrama	Agni-three	Karikavratral	ancient Sutra Smrti
9	Akhanda	Aguistoma	Caspian Sea	ArunadayaVedha	Alligator	Anumati
10	Abhijit	Alpa	Akasa	Acchavata	Apana	Arhodaya Parva
11	Acalesvara	Airavata	Amarakantaka mountain	Adhoksaja	Amasas	Arunodaya time
12	AbhistaPrada	Amas	Aksaya Vata	Adhna	Apiss	Astaka days
13	Acaras	BrahmaValvaria	Ambu	Agastyesa	Gopadma	Astakas
14	Airavata	Bh P AP	Brahmanda Puranas	Apoklimas	Homas	Asvamedha Pandavas
15	Ajesa	Bhagavata	Caksu	Ardhendu	Dharma	Asvina Brahmana
16	Agni Tirtha	Caritas	Bhaskaracarya Lilavati	Ankurarpana	Dhatust	Ayurdya
17	Agnih	Brahmanda P	Bhavisy	Annavrata	Karavella	Bharata war
18	Akhyana Suktas	Bhaga	Caturdati	Aryama	Havis	Bhisma pancaka vrata
19	Ajatasatru	Asiatic Society Calcutta	Budha	Aradhacandraka	Dhatri	Bhismapanaka
20	Agnih	Asvina Su	Bhavigvoltara Purana	Annaprasana	Bhusa	BHP VIII
21	Adrih	Caitra	Asipatravana	Amuktabharana	Bull	Bindupasa
22	Agastya	Br Vaivarta P SC	Asvini	Ananga Visvarupa	Kali	Brahmacarya
23	Akasa	Bibliotheca Indica	Carusthali	Arthavada	Gemini	Brahmanda P ch
24	Airavata	Avanti	Brahmanda Purana	Apoklima	Fulminox	Budhastami day
25	Ahita	Bhavisyottara P	Brahma Tirtha	Apanarga	Dadimas	Caturdasi day
26	Agnaye Sucha	Adit	BengalOrissa	Andhra Pradesh	Dual	constellation Hasta
27	Ahamkara	AirsPrana	Belva	Anudatta	Agnistoma	Dasa
28	Abhavyasvat	Br Valvarta	Aghora Kalpa	Abhramu	Annakuta	Dasami day
29	Abhra	Arta	Aksayavata	Adhana	Gavyutis	Dasami day
30	Aindri	Sitarta	Brahmanda	Apastamba	Kendras	Dasami day
31	Aditi	Chandasiti	Asadddstar udhaismuktaiAmbika	Dipta	Dasami Tithi	
32	Akrtavrana	Bhakti & Karma	Central Asia	Asipatravana	Asat	Dasami Tithi
33	Agnihotras	Anila Hasta	Bharatryillya	Antariksha	Aghellions	Deathanniversary
34	Adat	Ardra	Annapurna	Ahnika	Krsna	
35	Adideva	Bharani Purvaphalguni	Aries	Alvagnistoma	Asrama	Dvayayuga
36	Acara	Arghya	Amavasya	Adittitirtha	Kheta	Durga puja festival
37	Akruah	AmdVratas	Dadhi ocean	AsokaSayanaVrata	Aratni	DvadasI day
38	Aditi	Citi	Arya	Amala District	Dhaivata	DvadasI day

## Matsya puran sample o/p

PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
1	Abhichira	avasya	Abhimant	Abhaya	Adhika war
2	Abhira	Agastya	Abhichira	Abhaya	Adhika war
3	Abhira	Agastya	Abhichira	Abhaya	Adhika war
4	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
5	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
6	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
7	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
8	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
9	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
10	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
11	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
12	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
13	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
14	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
15	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
16	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
17	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
18	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
19	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
20	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
21	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
22	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
23	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
24	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
25	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
26	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
27	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
28	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
29	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
30	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
31	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
32	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
33	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
34	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
35	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
36	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
37	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war
38	Abhichira	Alakavasa	Abhichira	Abhaya	Adhika war

## Markhandey Puran Sample o/p

PERSON	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
1	Abandon brohman dvija	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
2	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
3	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
4	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
5	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
6	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
7	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
8	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
9	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
10	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
11	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
12	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
13	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
14	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
15	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
16	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
17	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
18	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
19	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
20	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
21	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
22	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
23	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
24	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
25	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
26	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
27	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
28	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
29	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
30	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
31	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
32	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
33	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
34	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
35	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
36	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
37	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava
38	Abhiras	Abhiras	Abhisani M Bh	Adityas	Acti Malatimadhava

## Linga Puran sample o/p

PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
1	Abhaya	Tvavr	Abja	Abala	Abhavayoga
2	Abhavayoga	Acamana	Agastya	Abhacara	Asrama
3	Abhaya	Adityas	Aghora	Abhiseka	Autumn
4	Abhavahabhy	Aghamarsana	Aghorakalpa	Abhyuksana	Avarana : Sarvasati
5	Abhacara	Aghora	Agni	Abobalarudrabhasya ST	Agastyas : Vidhana
6	Abhijit	Aghora Mantra	Aisvarya	Aghora	Avarana areVajra
7	Abhinivesa	Agnivesa	Ambujala	Acamana	AvaranaGanadhipa
8	Abhiseka	Abhobala	Anga	Acamaniya	Bharata war
9	Abhivadya	Aisvarya	Aspasavya	Acyuta	Brahma
10	Abhyanga	Aksapada	Apayani	Adhaka	Coming Going
11	Abhyuksana	Altau	Aquanus	Adharma	Cosmic Egg
12	Acalevara	Amareasa	Arghya	Adharmaabht	Anala
13	Acamana	Amba	Asana	Adharmavasa	Anila
14	Acamaniya	Ambika	Asia	Adhibhautika	Anima
15	Acarya	Amsu	Asuras	Adhidaivika	Apana
16	Acaryas	Amsumat	Asvattha	Adhyatmika	Asparas
17	Acyuta	Angiras	Attashaa	Aditi	Asptoryama
18	Adgiras	Anima	Avara	Adityas	Arghya
19	Adhaka	Apa	Avarana	Adrsyanti	Arvaksrotas
20	Adharma	Apana	Avarana Ratipriya	Agama	Asramas
21	Adharmavati	Apara	Avarana	Aghamarsana	Astamangalas
22	Adhiroha	Aparna	AvaranaBinduka	Aghanayika	Astra
23	Adhoksajah	Asparas	AvaranaDhara	Aghora	Atthava
24	Adhavas	Asparas	Avimuktika region	Aghoresa	Avarana
25	AdhyahVisva	Arghya	Avimuktas vara	Agneya	Avimukta Absolution
26	Adhyatmika	Arghya	Bagamati river	Agneyi	Bakulas
27	Adhivasa	Dhupa	Abhahotra	Agni	Bhutadi
28	Adhyta	Arjyanam	Bandhujiva	Agniyala	Bija
29	Adidava	Asrama	Karavira	Agnipurana	Bindus
30	Adiprimeval	Assembly	Barada	Ahavaneya	Biija
31	Adikah	Astamangalas	Beas	Ahipatra	Boar
32	Adikah	Astamangalas	Beas	Ahipatra	Boar

## Kurma Puran sample o/p

PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
1	Abhaya	Tvavr	Abja	Abala	Abhavayoga
2	Abhavayoga	Acamana	Agastya	Abhacara	Asrama
3	Abhaya	Adityas	Aghora	Abhiseka	Autumn
4	Abhavahabhy	Aghamarsana	Aghorakalpa	Abhyuksana	Avarana : Sarvasati
5	Abhacara	Aghora	Agni	Abobalarudrabhasya ST	Agastyas : Vidhana
6	Abhijit	Aghora Mantra	Aisvarya	Aghora	Avarana areVajra
7	Abhinivesa	Agnivesa	Ambujala	Acamana	AvaranaGanadhipa
8	Abhiseka	Abhobala	Anga	Acamaniya	Bharata war
9	Abhivadya	Aisvarya	Aspasavya	Acyuta	Brahma
10	Abhyanga	Aksapada	Apayani	Adhaka	Coming Going
11	Abhyuksana	Altau	Aquanus	Adharma	Cosmic Egg
12	Acalevara	Amareasa	Arghya	Adharmaabht	Anala
13	Acamana	Amba	Asana	Adharmavasa	Anila
14	Acamaniya	Ambika	Asia	Adhibhautika	Anima
15	Acarya	Amsu	Asuras	Adhidaivika	Apana
16	Acaryas	Amsumat	Asvattha	Adhyatmika	Asparas
17	Acyuta	Angiras	Attashaa	Aditi	Asptoryama
18	Adgiras	Anima	Avara	Adityas	Arghya
19	Adhaka	Apa	Avarana	Adrsyanti	Arvaksrotas
20	Adharma	Apana	Avarana Ratipriya	Agama	Asramas
21	Adharmavati	Apara	Avarana	Aghamarsana	Astamangalas
22	Adhiroha	Aparna	AvaranaBinduka	Aghanayika	Astra
23	Adhoksajah	Asparas	AvaranaDhara	Aghora	Atthava
24	Adhavas	Asparas	Avimuktika region	Aghoresa	Avarana
25	AdhyahVisva	Arghya	Avimuktas vara	Agneya	Avimukta Absolution
26	Adhyatmika	Arghya	Bagamati river	Agneyi	Bakulas
27	Adhivasa	Dhupa	Abhahotra	Agni	Bhutadi
28	Adhyta	Arjyanam	Bandhujiva	Agniyala	Bija
29	Adidava	Asrama	Karavira	Agnipurana	Bindus
30	Adiprimeval	Assembly	Barada	Ahavaneya	Biija
31	Adikah	Astamangalas	Beas	Ahipatra	Boar
32	Adikah	Astamangalas	Beas	Ahipatra	Boar

## Garuda Puran Sample o/p

PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
1	Brahma	Acara	Acaladhrti	Bharati	Anvastaka days
2	mahagati	Achyranthos Aspera	Adharasakti	Abhaya	Abhisaka
3	Abhaya	Aditi	Adharasakti	Abhisahas	Asauca
4	Abhaya Mudra	Tsyas	Agni	Abhiseka	Adhaka
5	Abhaya	Advaita Yoga	Agurul	Abhijana	Bharata war
6	Abhayarista	Agni Officialis	Ajmoda Sindura	Abhyanga	Adhyatmika
7	Abhi	Agniteljas	Aiyasas west	Abja	Adayas
8	Abhi tva Deva	Ahuh Ca	Aksayya Udana	Agneya	Deluge
9	Abhidhivata	Ahusa	Amaja	Acamaniya	Devasura war
10	Abhijit	Madanantare	Amla	Acyuta	Ekadasi
11	Abhimana	Avyeya	Ananda mountain	Adhaka	Full moon day
12	Abhimin	Avanti	Anagarkrida	Adharma	Ganges Rare
13	Abhimanyu	Alhagimauroon	Ananta	Adhidaivika	Ahuri
14	Abhinysa Jvara	Amaras	Adhikarana	Anda	Grhya Sutras
15	Abhimanyatam	Amlaloka	Anrudidha	Adhisomaka	Hemanta
16	Abhiratth	Amrtadhara	Anupa	Adhivasa	Inferno
17	Abhisekani	Amrit India	Anusvara	Adhivasal	Kartiki Purnima
18	Abhiseka	Anodana	Apah	Adhivasana	Alkala
19	Abhistuta	Anustup	Apara	Adhoksaja	KauravaPandava war
20	Abhita	Apa	Aquaris	Adhomukha	Aksepana
21	Abhija	Apana	Arabian Sea	Aditi	Krta Yuga
22	Abhija	Apasangabha	Aravinda forest	Aditya	Alaktaka
23	Abhija	Asparas	ArghyaSvaha	Adra	Amagha
24	Abhija	Arcturus	Arjuna	Agartha	Mahabharatawar
25	Abhija	Arctis	Artesian	Agastya	Anna
26	Abhija	Arctis	Artesian	Agastya	Mandala Rveda
27	Abhija	Arctis	Artesian	Agastya	NaraNarayana
28	Abhija	Arctis	Artesian	Agastya	Apadana
29	Abhija	Arctis	Artesian	Agastya	New moon day
30	Abhija	Arctis	Artesian	Agastya	Pancratia
31	Abhija	Arctis	Artesian	Agastya	Primordial Gadadhara
32	Abhija	Arctis	Artesian	Agastya	Samvatsari Puja
33	Abhija	Arctis	Artesian	Agastya	Apantatraka
34	Abhija	Arctis	Artesian	Agastya	Apida
35	Abhija	Arctis	Artesian	Agastya	Sannipata Jvara
36	Abhija	Arctis	Artesian	Agastya	Sarad
37	Abhija	Arctis	Artesian	Agastya	Sarad
38	Abhija	Arctis	Artesian	Agastya	Sarad
39	Abhija	Arctis	Artesian	Agastya	Sarad
40	Abhija	Arctis	Artesian	Agastya	Sarad
41	Abhija	Arctis	Artesian	Agastya	Sarad
42	Abhija	Arctis	Artesian	Agastya	Sarad
43	Abhija	Arctis	Artesian	Agastya	Sarad
44	Abhija	Arctis	Artesian	Agastya	Sarad
45	Abhija	Arctis	Artesian	Agastya	Sarad
46	Abhija	Arctis	Artesian	Agastya	Sarad
47	Abhija	Arctis	Artesian	Agastya	Sarad
48	Abhija	Arctis	Artesian	Agastya	Sarad
49	Abhija	Arctis	Artesian	Agastya	Sarad
50	Abhija	Arctis	Artesian	Agastya	Sarad
51	Abhija	Arctis	Artesian	Agastya	Sarad
52	Abhija	Arctis	Artesian	Agastya	Sarad
53	Abhija	Arctis	Artesian	Agastya	Sarad
54	Abhija	Arctis	Artesian	Agastya	Sarad
55	Abhija	Arctis	Artesian	Agastya	Sarad
56	Abhija	Arctis	Artesian	Agastya	Sarad
57	Abhija	Arctis	Artesian	Agastya	Sarad
58	Abhija	Arctis	Artesian	Agastya	Sarad
59	Abhija	Arctis	Artesian	Agastya	Sarad
60	Abhija	Arctis	Artesian	Agastya	Sarad

## Bhrama Puran Sample o/p

	A	B	C	D	E	F
1	PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS
2	Abhaya	Acamana	Africa	Abhiras	Acamana	Aghamarsana Mantra
3	Abhijit	Adityas	Agastya	Abhiseka	Acara	Amanasya
4	Abhimana	Agasthya	Agni	Abjaka	Air Ether	Bharata var
5	Abhimanyu	Aghora Mantra	Aindhrate TransBrahmaputra	Abu	Amratata	Brahma
6	Abhiras	Aktavarna	Aja	Acamana	Amrta	Caksusa Manvantara
7	Abhiras Said	Amrta	Akasa Sayana	Acyuta	Anakas	Caksusa Manvantara Daksa
8	Abhiya	Amrtasangama	Alakananda	Adhaharas	Andakataha	Caksusa Manvantara Rks
9	Abhiyayakti	Amra	Amavasa	Adhidivakra	Anguli	Chapter
10	Abhiyavanya	Anavesti	Anrudhda	Adhomukha	Apana	Dissolution
11	Abhitarajas	Apa	Anjana mountain	Adikalpa	Asparas	Dvapara Age
12	Abjaka	Apara Vidya	Apath	Aditi	Anani	Gayatri Mantra
13	Abuddhiman	Aparajita	Arabian Sea	Agfghanistan	Argha	Indra festival
14	Acamana	Appaya	Arjuna	Agastya	Arghya	Jambuvudipa Supreme Being
15	Acyuta	Asamanjas	Asamanjas	Agastyas	Arjunas	Kaliyuga Dharma
16	Adantavinasana	Arka	Asamanjasa	Aghora	Arsa	Kanya Zodiac
17	Adharma	Assembly	Asat	Agneya	Anuna	Krita age
18	Adhikara	Astral Science	Asokasaptami	Agni city Tejovati	Asat	Krita Yuga
19	Adhis	Asamedhas	Asuras	Agnihotra	Asokas	Krita Yuga Dharma
20	Adhishtha	Asavathira	Atala	Agninivala	Asuras	Krita Yuga penance
21	Adhista	Avinis	Ayu	Agninivata	Asvathith	Magh
22	Adhivanya	Adhivanya	Badari mountain	Agnivithra	Asvitha	Magh nakshatra
23	Adhvanyu	Avanti	Bakula	Agniweya	Atmukkas	Mandala Eveda
24	Adhyatha	Avantis	Banas	Aguru	Avashbitha	Manvantara
25	Adi	Avighna	Banyan	Abhaya	Babhrus	Manvantara Caksusa Manu
26	Adipavanu	Bharata	Bhadra	AbhayaSangama	Badaras	Manvantara Tamasa Manu
27	Aditeh Sutha	Bias	Bhandira	Abhamkara	Badari	Mathura Kalpa
28	Aditi	Brahma	Bharata	Ahar	Bahuvaras	Namitika Yuga
29	Aditiyas	Siva	Bharata Dandaka forest	Ahavanya	Bhanu	Treta Yuga
30	Aditya	Cakustirtha	Bhimatandavamukha	Abhuka	Bhasman	Treta Yuga Dharma
31	Aditya Tirtha	Candrayana	Bhogavati	Aindava	Bhirgarajas	Valisvavata Manvantara
32	Adityas	Cedi	Bhuvavoloka	Aindra	Bhuzadi	Vedic Texts
33	Adri	Cedies	Black Sea	Airaka	Biijupuraka	Yoga
34	Adrika	Centre Manyu	Blue mountain	Airavata	Biijupurakas	Yuga

```
!apt-get install poppler-utils
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
poppler-utils is already the newest version (22.02.0-2ubuntu0.5).
0 upgraded, 0 newly installed, 0 to remove and 49 not upgraded.
```

```
!pip install pdf2image
```

```
Collecting pdf2image
  Downloading pdf2image-1.17.0-py3-none-any.whl.metadata (6.2 kB)
Requirement already satisfied: pillow in
/usr/local/lib/python3.10/dist-packages (from pdf2image) (10.4.0)
Downloading pdf2image-1.17.0-py3-none-any.whl (11 kB)
Installing collected packages: pdf2image
Successfully installed pdf2image-1.17.0
```

```
!pip install easyocr
```

```
Collecting easyocr
  Downloading easyocr-1.7.1-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: torch in
/usr/local/lib/python3.10/dist-packages (from easyocr) (2.4.1+cu121)
Requirement already satisfied: torchvision>=0.5 in
/usr/local/lib/python3.10/dist-packages (from easyocr) (0.19.1+cu121)
Requirement already satisfied: opencv-python-headless in
/usr/local/lib/python3.10/dist-packages (from easyocr) (4.10.0.84)
Requirement already satisfied: scipy in
/usr/local/lib/python3.10/dist-packages (from easyocr) (1.13.1)
Requirement already satisfied: numpy in
/usr/local/lib/python3.10/dist-packages (from easyocr) (1.26.4)
Requirement already satisfied: Pillow in
/usr/local/lib/python3.10/dist-packages (from easyocr) (10.4.0)
Requirement already satisfied: scikit-image in
/usr/local/lib/python3.10/dist-packages (from easyocr) (0.24.0)
Collecting python-bidi (from easyocr)
  Downloading python_bidi-0.6.0-cp310-cp310-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (4.6 kB)
Requirement already satisfied: PyYAML in
/usr/local/lib/python3.10/dist-packages (from easyocr) (6.0.2)
Requirement already satisfied: Shapely in
/usr/local/lib/python3.10/dist-packages (from easyocr) (2.0.6)
Collecting pycclipper (from easyocr)
  Downloading pycclipper-1.3.0.post5-cp310-cp310-
manylinux_2_12_x86_64.manylinux2010_x86_64.whl.metadata (9.0 kB)
Collecting ninja (from easyocr)
  Downloading ninja-1.11.1.1-py2.py3-none-
manylinux1_x86_64.manylinux_2_5_x86_64.whl.metadata (5.3 kB)
Requirement already satisfied: filelock in
```

```
/usr/local/lib/python3.10/dist-packages (from torch->easyocr) (3.16.1)
Requirement already satisfied: typing-extensions>=4.8.0 in
/usr/local/lib/python3.10/dist-packages (from torch->easyocr) (4.12.2)
Requirement already satisfied: sympy in
/usr/local/lib/python3.10/dist-packages (from torch->easyocr) (1.13.3)
Requirement already satisfied: networkx in
/usr/local/lib/python3.10/dist-packages (from torch->easyocr) (3.3)
Requirement already satisfied: jinja2 in
/usr/local/lib/python3.10/dist-packages (from torch->easyocr) (3.1.4)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.10/dist-packages (from torch->easyocr)
(2024.6.1)
Requirement already satisfied: imageio>=2.33 in
/usr/local/lib/python3.10/dist-packages (from scikit-image->easyocr)
(2.35.1)
Requirement already satisfied: tifffile>=2022.8.12 in
/usr/local/lib/python3.10/dist-packages (from scikit-image->easyocr)
(2024.8.30)
Requirement already satisfied: packaging>=21 in
/usr/local/lib/python3.10/dist-packages (from scikit-image->easyocr)
(24.1)
Requirement already satisfied: lazy-loader>=0.4 in
/usr/local/lib/python3.10/dist-packages (from scikit-image->easyocr)
(0.4)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from jinja2->torch->easyocr)
(2.1.5)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.10/dist-packages (from sympy->torch->easyocr)
(1.3.0)
Downloading easyocr-1.7.1-py3-none-any.whl (2.9 MB)
_____ 2.9/2.9 MB 32.0 MB/s eta
0:00:00
anylinux1_x86_64.manylinux2_5_x86_64.whl (307 kB)
_____ 307.2/307.2 kB 28.5 MB/s eta
0:00:00
anylinux_2_12_x86_64.manylinux2010_x86_64.whl (908 kB)
_____ 908.3/908.3 kB 56.1 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (281 kB)
_____ 281.3/281.3 kB 27.1 MB/s eta
0:00:00

from pdf2image import convert_from_path
import easyocr
import numpy as np
import PIL
from PIL import ImageDraw
import spacy
```





# **The Shiva Maha Purana**

Sage Shaunaka requested Sutaji to tell about the proper method of listening to Shiva Purana, so that mankind gets complete benefit. Sutaji replied: "First of all, an auspicious moment should be determined by an Astrologer. After that, friends and relatives should be invited, especially those who have the tendency of being away from such occasions. The sages and the virtuous people should be invited too. The 'Katha' must be held in sacred places like Shiva temple, any place of pilgrimage or in one's home after doing a Bhumi Pujan of the land where one intends to hold the Katha of Shiv Puran. The canopy should be well decorated."

"After making a resolution and doing worship of Ganapati - the destroyer of all hurdles and obstacles the Katha should be commenced. The person who is telling the 'Katha' should be facing north and all the listeners should sit facing east. The person who is telling the 'Katha' should be a scholar and should be capable of clearing all the doubts from the listener's mind. There should be no kind of distraction during the 'Katha' period. A devotee, who listens to the Katha leaving behind all of his worldly worries get complete benefits. A devotee should also make donations and offerings according to his capacity and capability otherwise he would become a wretched man. The Mantra 'OM NAMAH SHIVAYA' should be chanted through out the period of Katha."

## **RIGHT OF LISTENING TO SHIVA MAHA PURANA:**

An uninitiated person does not have the right to listen to the Katha of Shiva Maha Purana. So a man desirous of listening to Shiva Maha Purana should first get initiated.

## **AUSTERITIES TO BE FOLLOWED:**

A devotee who has taken the vow of listening to the Katha of Shiva Maha Purana, should follow a celibate life. He should sleep on the floor and should have only fruits in the diet. He can have his normal dinner only after the completion of the Katha. The diet should be fresh and pure. He should also

```

bounds = reader.readtext(np.array(images[4]), min_size=0,
slope_ths=0.2, ycenter_ths=0.7,height_ths=0.6, width_ths=0.8,
decoder='beamsearch', beamWidth = 10)
bounds

[[[149, 156], [1353, 156], [1353, 218], [149, 218]],
 'Shivalinga. From that day onwards the Shivalinga became famous. ',
 0.8181146533921518),
 [[156, 264], [866, 264], [866, 316], [156, 316]],
 '1.5 The Deities go to Kailash Mountain',
 0.9918164917516095),
 [[151, 362], [1474, 362], [1474, 418], [151, 418]],
 'Nandikeshwarji narrates the tale of battle between Lord Brahma and
Lord',
 0.7710677394693577),
 [[157, 417], [295, 417], [295, 461], [157, 461]],
 'Vishnu.',
 0.630442373467233),
 [[152, 511], [1489, 511], [1489, 574], [152, 574]],
 '"Once while travelling lord Brahma reached the abode of Lord
Vishnu: He',
 0.7952309516264272),
 [[153, 565], [1420, 565], [1420, 626], [153, 626]],
 'saw Lord Vishnu: He saw lord Vishnu resting o Shesh-Nag and being',
 0.6829291875659096),
 [[152, 622], [1544, 622], [1544, 674], [152, 674]],
 'attended by Garuda and other attendants. When Brahmaji saw that
Vishnu did',
 0.7839014462707197),
 [[151, 670], [1441, 670], [1441, 730], [151, 730]],
 'not get up to receive him, he became very angry. Very soon, Verbal
dual',
 0.8697074231872134),
 [[149, 721], [1511, 721], [1511, 781], [149, 781]],
 'erupted between them. It became so severe that a battle was fought
between',
 0.8144849757932107),
 [[149, 772], [1470, 772], [1470, 829], [149, 829]],
 'them, which continued for very long time. All the deities arrived
from the',
 0.5858529140237657),
 [[151, 820], [1541, 820], [1541, 881], [151, 881]],
 'heaven to watch the battle. They became very worried when they saw
no sign',
 0.7102462624593933),
 [[151, 872], [1536, 872], [1536, 933], [151, 933]],
 'of battle coming to an end: They decided to go to lord Shiva, to
seek his help.',
 0.6459547963425764),
 [[156, 977], [831, 977], [831, 1029], [156, 1029]],

```

'1.6 Anaal-Stambh (The Pillar of Fire)',  
0.7329948601958487),  
([[152, 1078], [1517, 1078], [1517, 1138], [152, 1138]]),  
'"Though Lord Shiva knew everything, but still feigning ignorance,  
he asked',  
0.7370291970648212),  
([[151, 1127], [1446, 1127], [1446, 1189], [151, 1189]]),  
'about the well beings of the world. The deities told him about the  
battle,',  
0.9371692306588729),  
([[152, 1184], [878, 1184], [878, 1236], [152, 1236]]),  
'fought between Brahmaji and Vishnuji: "',  
0.7299636400023428),  
([[152, 1284], [1514, 1284], [1514, 1340], [152, 1340]]),  
'"Lord Shiva then sent his one hundred Ganas to pacify both of them:  
He too',  
0.9053913334510594),  
([[155, 1334], [1534, 1334], [1534, 1394], [155, 1394]]),  
'went there accompanied by mother Parvati, boarded on a chariot.  
When Lord',  
0.5300140871768328),  
([[151, 1386], [1470, 1386], [1470, 1440], [151, 1440]]),  
'Shiva reached there, he saw that Brahmaji and Vishnuji were about  
to use',  
0.9108064110605142),  
([[149, 1430], [1467, 1430], [1467, 1496], [149, 1496]]),  
'their deadly weapons- Maheshwar and Pashupat respectively. Fearing  
the',  
0.4536554641319501),  
([[150, 1485], [1453, 1485], [1453, 1548], [150, 1548]]),  
'destruction, which these deadly weapons might have caused, Lord  
Shiva',  
0.6884850950023326),  
([[152, 1539], [1516, 1539], [1516, 1596], [152, 1596]]),  
"manifested himself in the form of 'Analstamba' (pillar of fire)  
between them:",  
0.6349666954298028),  
([[152, 1592], [1504, 1592], [1504, 1646], [152, 1646]]),  
'Brahmaji and Vishnuji had already released their weapons- Maheshwar  
and',  
0.9212098123606701),  
([[150, 1636], [1454, 1636], [1454, 1701], [150, 1701]]),  
'Pashupat: Both the weapons fell into that pillar of fire and got  
destroyed',  
0.7765577939612495),  
([[153, 1739], [1479, 1739], [1479, 1805], [153, 1805]]),  
'"Brahmaji and Vishnuji were very surprised to see the pillar of  
fire, which',  
0.8269090621648364),



```

    ([[155, 1792], [1443, 1792], [1443, 1852], [155, 1852]],
    'was so enormous in size that it reached the sky and penetrated down
the',
    0.598389747234682),
    ([[154, 1846], [1532, 1846], [1532, 1898], [154, 1898]],
    "earth. Vishnuji transformed himself into a boar and went to the
'Patal' (nether",
    0.7325215944588511),
    ([[152, 1898], [1502, 1898], [1502, 1950], [152, 1950]],
    "world) to find the base of that 'Pillar of fire'. But he was
unsuccessful in his",
    0.7064392155463264),
    ([[152, 1950], [582, 1950], [582, 2004], [152, 2004]],
    'attempt and came back:',
    0.83835629155542)]

```

```

templist = []
for i in range(len(bounds)):
    templist.append(bounds[i][1])
templist

```

```

['Shivalinga. From that day onwards the Shivalinga became famous. ',
'1.5 The Deities go to Kailash Mountain',
'Nandikeshwarji narrates the tale of battle between Lord Brahma and
Lord',
'Vishnu.',
'"Once while travelling lord Brahma reached the abode of Lord Vishnu:
He',
'saw Lord Vishnu: He saw lord Vishnu resting o Shesh-Nag and being',
'attended by Garuda and other attendants. When Brahmaji saw that
Vishnu did',
'not get up to receive him, he became very angry. Very soon, Verbal
dual',
'erupted between them. It became so severe that a battle was fought
between',
'them, which continued for very long time. All the deities arrived
from the',
'heaven to watch the battle. They became very worried when they saw
no sign',
'of battle coming to an end: They decided to go to lord Shiva, to
seek his help.',
'1.6 Anaal-Stambh (The Pillar of Fire)',
'"Though Lord Shiva knew everything, but still feigning ignorance, he
asked',
'about the well beings of the world. The deities told him about the
battle,',
'fought between Brahmaji and Vishnuji: "',
'"Lord Shiva then sent his one hundred Ganas to pacify both of them:
He too',
'went there accompanied by mother Parvati, boarded on a chariot. When

```

```
Lord',  
'Shiva reached there, he saw that Brahmaji and Vishnuji were about to  
use',  
'their deadly weapons- Maheshwar and Pashupat respectively. Fearing  
the',  
'destruction, which these deadly weapons might have caused, Lord  
Shiva',  
"manifested himself in the form of 'Analstamba' (pillar of fire)  
between them:",  
'Brahmaji and Vishnuji had already released their weapons- Maheshwar  
and',  
'Pashupat: Both the weapons fell into that pillar of fire and got  
destroyed',  
'"Brahmaji and Vishnuji were very surprised to see the pillar of  
fire, which',  
'was so enormous in size that it reached the sky and penetrated down  
the',  
"earth. Vishnuji transformed himself into a boar and went to the  
'Patal' (nether",  
"world) to find the base of that 'Pillar of fire'. But he was  
unsuccessful in his",  
'attempt and came back:']
```

```
for i in range(len(images)):  
    bounds = reader.readtext(np.array(images[i]), min_size=0,  
slope_ths=0.2, ycenter_ths=0.7,height_ths=0.6, width_ths=0.8,  
decoder='beamsearch', beamWidth = 10)  
    templist = []  
    for i in range(len(bounds)):  
        templist.append(bounds[i][1])  
    with open("/content/myfile.txt", "a") as file:  
        # Iterate over the list and write each element to the file  
        for item in templist:  
            file.write(item + "\n")
```

```

import os
import glob
import stanza
import pandas as pd

# Initialize the Stanza pipeline for English with NER processing
nlp = stanza.Pipeline(lang='en', processors='tokenize,ner')

def read_text_from_file(file_path):
    with open(file_path, 'r', encoding='utf-8') as file:
        return file.read()

def process_file(file_path):
    # Read the text from the file
    text = read_text_from_file(file_path)

    # Process the text with the NER pipeline
    doc = nlp(text)

    # Store named entities of different types in separate sets
    entities = {
        "PERSONS": {entity.text for entity in doc.ents if entity.type == 'PERSON'},
        "ORGANIZATIONS": {entity.text for entity in doc.ents if entity.type == 'ORG'},
        "LOCATIONS": {entity.text for entity in doc.ents if entity.type == 'LOC'},
        "GEOPOLITICAL ENTITIES": {entity.text for entity in doc.ents if entity.type == 'GPE'},
        "PRODUCTS": {entity.text for entity in doc.ents if entity.type == 'PRODUCT'},
        "EVENTS": {entity.text for entity in doc.ents if entity.type == 'EVENT'}
    }

    # Convert the entities to a DataFrame
    df = pd.DataFrame(dict([(k, pd.Series(list(v))) for k, v in entities.items()]))

    # Generate the output file path with the same name as the input file
    base_name = os.path.basename(file_path)
    name_without_ext, _ = os.path.splitext(base_name)
    output_path = f'C:/Users/Admin/Desktop/purana text/{name_without_ext}_ner_output.xlsx'

    # Store the output in an Excel file
    df.to_excel(output_path, index=False)

    print(f"Named entities from {file_path} saved to {output_path}")

# Define the folder containing the .txt and .md files
folder_path = 'C:/Users/Admin/Desktop/purana text'

# Use glob to find all .txt and .md files in the folder
file_paths = glob.glob(os.path.join(folder_path, '*.txt')) + glob.glob(os.path.join(folder_path, '*.md'))

# Process each file
for file_path in file_paths:
    process_file(file_path)

```

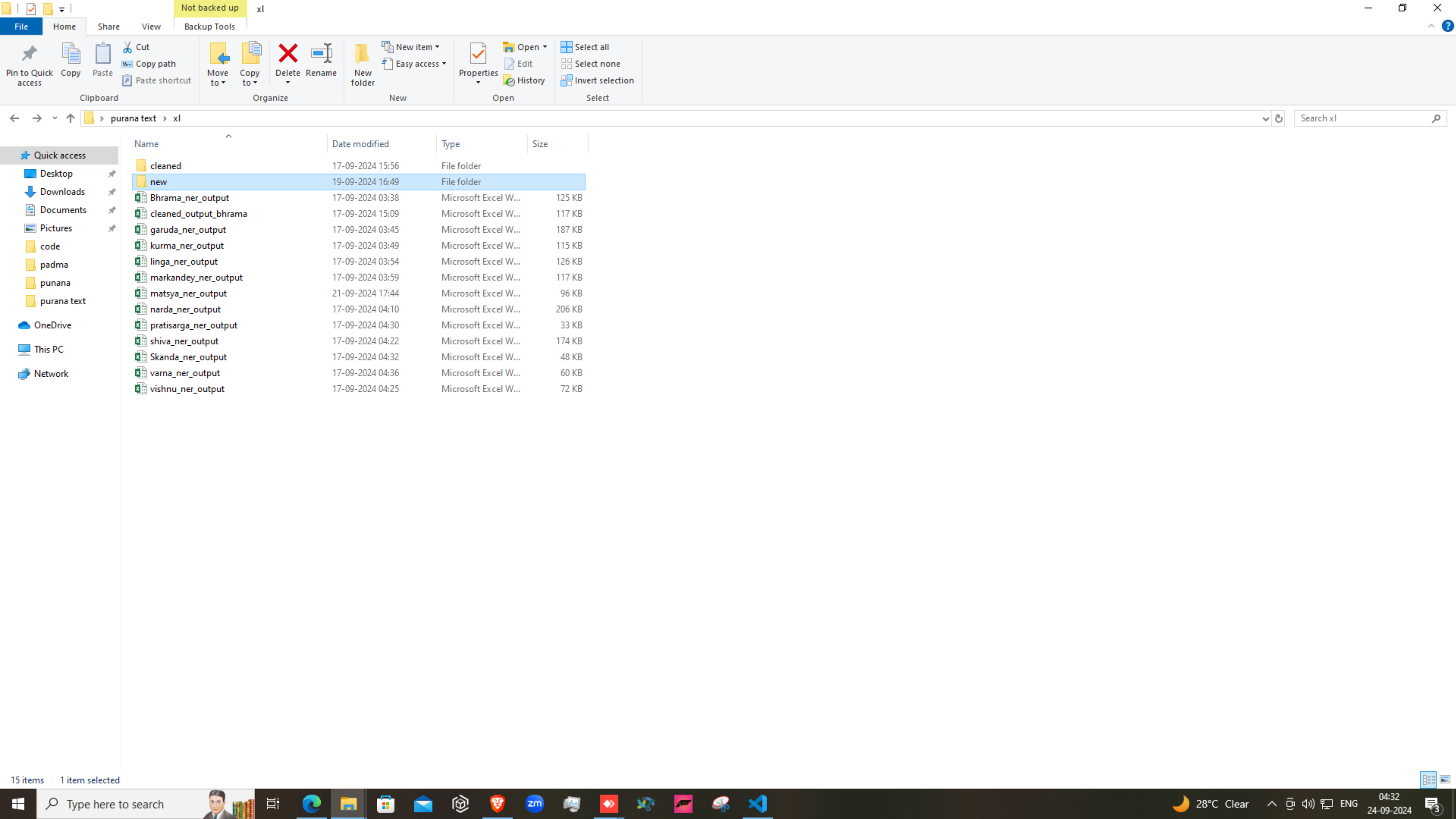
```
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/code/next.py
2024-09-24 04:24:40 INFO: Checking for updates to resources.json in case models have been updated. Note: this behavior can be turned off with download_method=None or download_method=DownloadMethod.REUSE_RESOURCES
Downloading https://raw.githubusercontent.com/stanfordnlp/stanza-resources/main/resources_1.8.0.json: 386kB [00:00, 31.9MB/s]
2024-09-24 04:24:41 INFO: Downloaded file to C:\Users\Admin\stanza_resources\resources.json
2024-09-24 04:24:41 WARNING: Language en package default expects mwt, which has been added
2024-09-24 04:24:41 INFO: Loading these models for language: en (English):
```

```
=====
```

Processor	Package
-----	
tokenize	combined
mwt	combined
ner	ontonotes-ww-multi_charlm

```
=====
```

```
2024-09-24 04:24:41 INFO: Using device: cpu
2024-09-24 04:24:41 INFO: Loading: tokenize
C:\Users\Admin\Desktop\vs_studio\env\Lib\site-packages\stanza\models\tokenization\trainer.py:82: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for more details). In a future release, the default value for `weights_only` will be flipped to `True`. This limits the functions that could be executed during unpickling. Arbitrary objects will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted by the user via `torch.serialization.add_safe_globals`. We recommend you start setting `weights_only=True` for any use case where you don't have full control of the loaded file. Please open an issue on GitHub for any issues related to this experimental feature.
  checkpoint = torch.load(filename, lambda storage, loc: storage)
2024-09-24 04:24:42 INFO: Loading: mwt
C:\Users\Admin\Desktop\vs_studio\env\Lib\site-packages\stanza\models\mwt\trainer.py:170: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for more details). In a future release, the default value for `weights_only` will be flipped to `True`. This limits the functions that could be executed during unpickling. Arbitrary objects will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted by the user via `torch.serialization.add_safe_globals`. We recommend you start setting `weights_only=True` for any use case where you don't have full control of the loaded file. Please open an issue on GitHub for any issues related to this experimental feature.
  checkpoint = torch.load(filename, lambda storage, loc: storage)
2024-09-24 04:24:42 INFO: Loading: ner
C:\Users\Admin\Desktop\vs_studio\env\Lib\site-packages\stanza\models\ner\trainer.py:197: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for more details). In a future release, the default value for `weights_only` will be flipped to `True`. This limits the functions that could be executed during unpickling. Arbitrary objects will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted by the user via `torch.serialization.add_safe_globals`. We recommend you start setting `weights_only=True` for any use case where you don't have full control of the loaded file. Please open an issue on GitHub for any issues related to this experimental feature.
  checkpoint = torch.load(filename, lambda storage, loc: storage)
C:\Users\Admin\Desktop\vs_studio\env\Lib\site-packages\stanza\models\common\pretrain.py:56: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for more details). In a future release, the default value for `weights_only` will be flipped to `True`. This limits the functions that could be executed during unpickling. Arbitrary objects will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted by the user via `torch.serialization.add_safe_globals`. We recommend you start setting `weights_only=True` for any use case where you don't have full control of the loaded file. Please open an issue on GitHub for any issues related to this experimental feature.
  data = torch.load(self.filename, lambda storage, loc: storage)
C:\Users\Admin\Desktop\vs_studio\env\Lib\site-packages\stanza\models\common\char_model.py:271: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for more details). In a future release, the default value for `weights_only` will be flipped to `True`. This limits the functions that could be executed during unpickling. Arbitrary objects will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted by the user via `torch.serialization.add_safe_globals`. We recommend you start setting `weights_only=True` for any use case where you don't have full control of the loaded file. Please open an issue on GitHub for any issues related to this experimental feature.
  state = torch.load(filename, lambda storage, loc: storage)
2024-09-24 04:24:43 INFO: Done loading processors!
Named entities from C:/Users/Admin/Desktop/purana text\Bhrama.txt saved to C:/Users/Admin/Desktop/purana text1/Bhrama_ner_output.xlsx
□
```



Name	Date modified	Type	Size
cleaned	17-09-2024 15:56	File folder	
new	19-09-2024 16:49	File folder	
Bhrama_ner_output	17-09-2024 03:38	Microsoft Excel W...	125 KB
cleaned_output_bhrama	17-09-2024 15:09	Microsoft Excel W...	117 KB
garuda_ner_output	17-09-2024 03:45	Microsoft Excel W...	187 KB
kurma_ner_output	17-09-2024 03:49	Microsoft Excel W...	115 KB
linga_ner_output	17-09-2024 03:54	Microsoft Excel W...	126 KB
markandey_ner_output	17-09-2024 03:59	Microsoft Excel W...	117 KB
matsya_ner_output	21-09-2024 17:44	Microsoft Excel W...	96 KB
narda_ner_output	17-09-2024 04:10	Microsoft Excel W...	206 KB
pratisarga_ner_output	17-09-2024 04:30	Microsoft Excel W...	33 KB
shiva_ner_output	17-09-2024 04:22	Microsoft Excel W...	174 KB
Skanda_ner_output	17-09-2024 04:32	Microsoft Excel W...	48 KB
varna_ner_output	17-09-2024 04:36	Microsoft Excel W...	60 KB
vishnu_ner_output	17-09-2024 04:25	Microsoft Excel W...	72 KB



FileHomeInsertPage LayoutFormulasDataReviewViewTell me what you want to do...Sign inShare

Cut

Copy

Format Painter

Calibri11A A

Wrap Text

Merge & Center

General

Conditional Formatting

Format as Table

Normal

Bad

Good

Neutral

Calculation

Check Cell

Explanatory ...

Input

Linked Cell

Note

Insert

Delete

Format

AutoSum

Fill

Clear

Sort & Filter

Find & Select

Clipboard

Font

Alignment

Number

Styles

Cells

Editing

A1PERSONS

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	PERSONS	ORGANIZATIONS	LOCATIONS	GEOPOLITICAL ENTITIES	PRODUCTS	EVENTS													
2	Pratardana	Vcdas	the Paṇḍya Kingdom	Nanda	Madgus	the Treta Yuga													
3	Nanda	Adityas	South West	Sūkara	Pippalas	the Gayatrī Mantra													
4	Pancajanya-Kṛṣṇa's	Kurukṣetra, Pañcāla, IMatsya, Ma	Siberia	Yakṣas	Bahuvāras	the Aghamarṣaṇa Mantra													
5	Rukmin Kṛṣṇa	Kāsmīras	the Mandara mountain	Balahaka	Vadava	Chapter 51													
6	Garuda Tirtha Brahmā	Post	Mathurā	Yuvanasva	Bījapūraka	the Yoga 17													
7	Soma's	Havi	the Gautami Ganga	Vrṇḍavana	Karṇapūras	Kṛta2 Yuga penance													
8	Srī Kṛṣṇa	Daṇḍaka	the Leo zodiac	Sehoa	Caru	the Kṛta Yuga Dharma													
9	May Soma	Hotravartis	Hara	Kuhakas	Pārthivas	Manvantara													
10	Glory of the Holy River Nāra	Aśvatīrtha	Mānuṣa	Karnikara	Manvantara	Kṛta yuga													
11	Sages	the Supreme Soul	the Muñja grass	Karavīra	Pratisancara	the Kanya Zodiac													
12	Vidyadhara	Cola	Dandaka forest	Pivara	Plaksa	the Kṛta Yuga													
13	Yakṣas	Tarpana to Devas, Pitrs	the Govardhana Tirtha	Ajamukharasa	Visnu's	the Vaivasvata Manvantara													
14	Quieten Kṛtya	Manvantara	Siddhas	Aśvinī	Boar	the Dvapara Age													
15	Jyoti	Jivañjīvakas	Ocean	Susmanta	Purusottama	Yuga													
16	Rajatanabha	Vāḷakhilyas	Vanga	Madhva	Saindhavāyanas	Mandala of the Rgevda													
17	Śvetamādhava	the Maker of the universe	the Naimisa forest	Varāha	Bhrigarajas	Brahma													
18	Sarasvatīsangama	Pramathas-Ganas	Vindhya Mountains	Paila	Pomegranate	the Caksusa Manvantara													
19	Daksa-Prajāpati	Salvation	South-Eastern	Marici	Lakucas	the Mathura Kalpa													
20	Muni	Para Vidya	Asamañjasa	Hraum	Trident	the Bharata war													
21	Yuvanasva	Vedanta	Vedāntas	Madhuvāhinī	Latāvallis	Magha naksatra													
22	Valmiki Rāmāyaṇa	Puskara	northern Varsa mountains	Vidyutprabha	Pūgaphalas	Manvantara O													
23	Vrṇḍavana	Pūtanā	Capricorn	Kovidāra	Gangā	Caksusa Manvantara Rks													
24	Saṅkara	Holy Centre Manyu	Alakananda	Konkaṇas	Acamana	the Treta Yuga Dharma													
25	Yati	Purusottama	Samkhya	Su-rastra	Ghrta	Treta Yuga													
26	Nandīsvara	Vaidyuta	Pisces	Maitreya	Vāstuka	the Naimittika Laya													
27	Satyaketu	Ila	Mitra	Viṣṇu	Plakṣas	the Kaliyuga Dharma													
28	Ekaparna	Trident	Banas	Kṛṣṇa	Argha	Magha													
29	Vikala	Sunanda10	the Eastern Sea	Sūktiman	Aśvayantra	Dissolution													
30	Clay	Pitr Tirtha	Nimbasa	Arimardana	Pindasa	the Vedic Texts													
31	Sarmiṣṭhā	the Vaṣaṭkāra Mantras	Jambu Pippala	Kalingas	Karañjas	the Kṛta age1													
32	Varāha	Nivṛtti	Vamana	Samkarsana	Sun-god	Amavasya													
33	Artha	Pravara-Sangama	the Himalaya mountain	Arsa	the Vehicle Garuda	Indra's festival													
34	Prākṛtika	Avighna	The Varsa mountains	Pariplava	Paitamahatirtha	the Manvantara of Caksusa Manu													
35	Raksas	Acamana	the mountain Meru	Amsuman	Hotra	the Jambudvipa the Supreme Being													
36	Marici	Tradition Nara	Sarala	Madhavas	Badaras	the Caksusa Manvantara Daksa													
37	Vidyās	Twentytwo Thousand Holy Centre	Nila	Bala	Cuckoos	the Manvantara of Tamasa Manu													
38	Nagas	Samkhya and Yoga	Upper Regions	Nirrti's	Datyūhas														
39	Dvainaavana	MW	Sat	Vaikuntha	Pink Sindhuvarakas														

Sheet1

Ready

28°C Clear04:3324-09-2024

```

import os
import pandas as pd
import re
import unicodedata

# Function to normalize text
def normalize_text(text):
    if isinstance(text, str): # Apply normalization only if the value is a string
        # Normalize diacritics to closest Latin equivalent
        text = unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8')

    # Define specific patterns and words to remove
    removals = [
        r'\b(Sri|sri|King|Raja)\b[\s\*]*', # Remove 'Sri', 'sri', 'King', 'Raja' with optional spaces or '*' after
        r'\bGlory\b', r'\bHoly\b', r'\bSage\b', r'\bFortynine\b', r'\bPaulastya Tirtha\b',
        r'\bDescent\b', r'\bMarriage\b', r'\bEarth\b', r'\bdead\b', r'\bOn Namo\b', r'\bSarva\b',
        r'\bLord\b', r'\bOm\b', r'\bAm\b', r'\bDynasty\b', r'\bGreatest\b', r'\bDevas\b',
        r'\bchief\b', r'\bDevi\b', r'\bHrai\b'
    ]

    # Remove specific words or patterns
    for pattern in removals:
        text = re.sub(pattern, '', text, flags=re.IGNORECASE)

    # Remove punctuations or special characters: '.', '*', '(', ')', '{', '}', '[', ']'
    text = re.sub(r'[.\*()\{\}\[\]]', '', text)

    # Remove the number '1' at the end of words (like 'word1' to 'word')
    text = re.sub(r'l\b', '', text)

    # Strip extra spaces
    text = re.sub(' +', ' ', text.strip())
    return text

# Input and output folder paths
input_folder = r"C:\Users\Admin\Desktop\purana text\xl\cleaned"
output_folder = r"C:\Users\Admin\Desktop\purana text\xl\new"

# Ensure the output directory exists
if not os.path.exists(output_folder):
    os.makedirs(output_folder)

# Loop through all Excel files in the input folder
for file_name in os.listdir(input_folder):
    if file_name.endswith(".xlsx"): # Process only .xlsx files
        input_path = os.path.join(input_folder, file_name)
        output_path = os.path.join(output_folder, file_name)

        # Load the Excel file into a DataFrame
        df = pd.read_excel(input_path)

        # Apply normalization to every column in the DataFrame
        df = df.applymap(normalize_text)

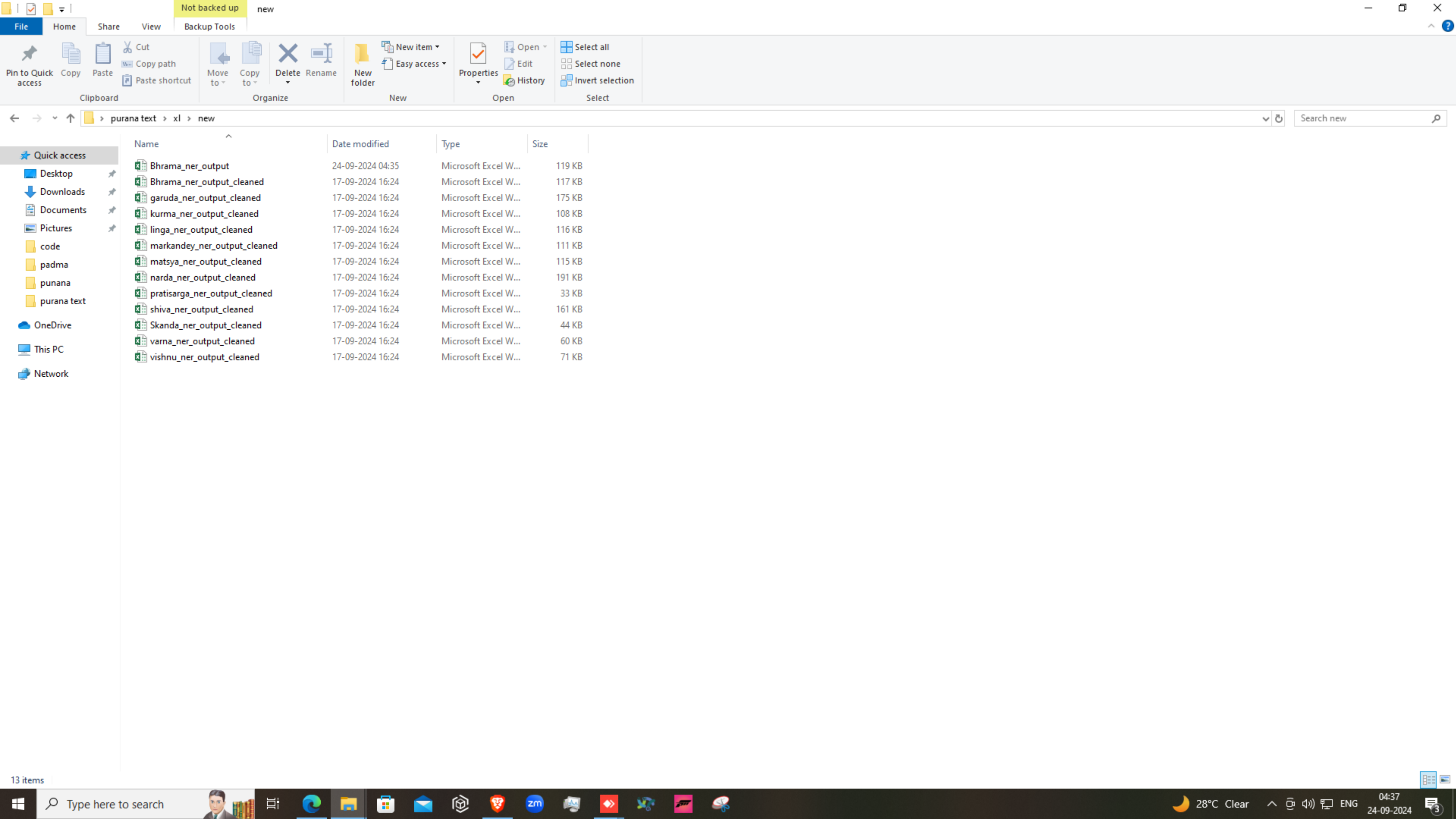
        # Save the corrected DataFrame back to a new Excel file with the same name in the output folder
        df.to_excel(output_path, index=False)

        print(f"Processed and saved: {file_name}")

```

```
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
c:\Users\Admin\Desktop\vs_studio\ckeaning_bhrama\latin:53: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
    df = df.applymap(normalize_text)
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
c:\Users\Admin\Desktop\vs_studio\ckeaning_bhrama\latin:53: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
c:\Users\Admin\Desktop\vs_studio\ckeaning_bhrama\latin:53: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
c:\Users\Admin\Desktop\vs_studio\ckeaning_bhrama\latin:53: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
c:\Users\Admin\Desktop\vs_studio\ckeaning_bhrama\latin:53: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
    df = df.applymap(normalize_text)
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
PS C:\Users\Admin\Desktop\vs_studio> & c:/Users/Admin/Desktop/vs_studio/env/Scripts/python.exe c:/Users/Admin/Desktop/vs_studio/ckeaning_bhrama/latin
c:\Users\Admin\Desktop\vs_studio\ckeaning_bhrama\latin:53: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
c:\Users\Admin\Desktop\vs_studio\ckeaning_bhrama\latin:53: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
    df = df.applymap(normalize_text)
Processed and saved: Bhrama_ner_output.xlsx
PS C:\Users\Admin\Desktop\vs_studio>
```





Name	Date modified	Type	Size
Bhrama_ner_output	24-09-2024 04:35	Microsoft Excel W...	119 KB
Bhrama_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	117 KB
garuda_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	175 KB
kurma_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	108 KB
linga_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	116 KB
markandey_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	111 KB
matsya_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	115 KB
narda_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	191 KB
pratisarga_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	33 KB
shiva_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	161 KB
Skanda_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	44 KB
varna_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	60 KB
vishnu_ner_output_cleaned	17-09-2024 16:24	Microsoft Excel W...	71 KB

