Final Project Notebook

DS 5001 Text as Data | Spring 2025

Metadata

- Full Name: Vishwanath Premanand Guruvayur
- Userid: qtf7du
- GitHub Repo URL: https://github.com/vishugp/Mahabharata_ETA
- UVA Box URL: https://virginia.box.com/s/a5e1x8io85kwmz2jfi4pole80xuwhe8a

Overview

The goal of the final project is for you to create a **digital analytical edition** of a corpus using the tools, practices, and perspectives you've learning in this course. You will select a corpus that has already been digitized and transcribed, parse that into an F-compliant set of tables, and then generate and visualize the results of a series of fitted models. You will also draw some tentative conclusions regarding the linguistic, cultural, psychological, or historical features represented by your corpus. The point of the exercise is to have you work with a corpus through the entire pipeline from ingestion to interpretation.

Specifically, you will acquire a collection of long-form texts and perform the following operations:

- **Convert** the collection from their source formats (F0) into a set of tables that conform to the Standard Text Analytic Data Model (F2).
- **Annotate** these tables with statistical and linguistic features using ETA libraries such as NLTK (F3).
- **Produce** a vector representation of the corpus to generate TFIDF values to add to the TOKEN (aka CORPUS) and VOCAB tables (F4).
- **Model** the annotated and vectorized model with tables and features derived from the application of unsupervised methods, including PCA, LDA, and word2vec (F5).
- **Explore** your results using statistical and visual methods.
- **Present** conclusions about patterns observed in the corpus by means of these operations.

When you are finished, you will make the results of your work available in GitHub (for code) and UVA Box (for data). You will submit to Gradescope (via Canvas) a PDF version of a Jupyter notebook that contains the information listed below.

Some Details

- Please fill out your answers in each task below by editing the markdown cell.
- Replace text that asks you to insert something with the thing, i.e. replace `(INSERT IMAGE HERE)` with an image element, e.g. `![](image.png)`.
- For URLs, just paste the raw URL directly into the text area. Don't worry about providing link labels using `[label](link)`.
- Please do not alter the structure of the document or cell, i.e. the bulleted lists.
- You may add explanatory paragraphs below the bulleted lists.
- Please name your tables as they are named in each task below.
- Tasks are indicated by headers with point values in parentheses.

Raw Data

Source Description (1)

Provide a brief description of your source material, including its provenance and content. Tell us where you found it and what kind of content it contains.

The source material for my project is the Mahabharata corpus, obtained from sacred-texts.com, a reputable archive of religious and mythological texts.

This version is a public domain English translation by Kisari Mohan Ganguli, completed in the late 19th century. It encompasses all 18 Parvas (books) of the epic, providing a complete narrative of one of ancient India's most significant texts. The content covers a vast range of themes such as mythology, philosophy, politics, war, family conflict and ethics and is presented through stories, discourses, and dialogues including the Bhagavad Gita. The prose translation closely follows the structure and meaning of the original Sanskrit, making it a rich and coherent source for text analysis.

Source Features (1)

Add values for the following items. (Do this for all following bulleted lists.)

- Source URL: https://sacred-texts.com/hin/maha/index.htm
- UVA Box URL: https://virginia.box.com/s/ylcbexuohyz8ceq1cft0l420lkvpyhtb
- Number of raw documents: 18
- Total size of raw documents (e.g. in MB): 14.5 MB
- File format(s), e.g. XML, plaintext, etc.: plaintext

Source Document Structure (1)

Provide a brief description of the internal structure of each document. That, describe the typical elements found in document and their relation to each other. For example, a corpus of letters might be described as having a date, an addressee, a salutation, a set of content paragraphs, and closing. If they are various structures, state that.

Each document corresponds to one of the 18 parvas (books) of the Mahabharata, each of which contains multiple Upa-parvas (sub-books or chapters). Each Upa-parva is further divided into sections, which consist of paragraphs typically structured as English translations of Sanskrit verses. These verses often follow a poetic format and convey narrative, dialogue, or philosophical discourse within the epic.

Parsed and Annotated Data

Parse the raw data into the three core tables of your addition: the `LIB`, `CORPUS`, and `VOCAB` tables.

These tables will be stored as CSV files with header rows.

You may consider using `|` as a delimitter.

Provide the following information for each.

LIB (2)

The source documents the corpus comprises. These may be books, plays, newspaper articles, abstracts, blog posts, etc.

Note that these are *not* documents in the sense used to describe a bag-of-words representation of a text, e.g. chapter.

- UVA Box URL: https://virginia.box.com/s/mmcb2l34x77en2k87l2cbyrh47vmpjs8
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/01_create_F2.ipynb

- Delimitter: |
- Number of observations: 18
- List of features, including at least three that may be used for model summarization (e.g. date, author, etc.): raw_title, timeline, main_theme, broad_theme, key_focus
- Average length of each document in characters: 620,384

CORPUS (2)

The sequence of word tokens in the corpus, indexed by their location in the corpus and document structures.

- UVA Box URL: https://virginia.box.com/s/a7ep2x9g2aj4vw78rt9kjmu4eguw53cn
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/01_create_F2.ipynb
- Delimitter: `|`
- Number of observations: 2,500,000
- OHCO Structure (as delimitted column names): book_id|chap_id|sec_id|para_num|sent_num|token_num
- Columns (as delimitted column names, including `token_str`, `term_str`, `pos`, and `pos_group`): pos_tuple|pos|token_str|term_str|pos_group|term_len

VOCAB (2)

The unique word types (terms) in the corpus.

- UVA Box URL: https://virginia.box.com/s/f61qiaac0s51rzkrgbjrwpoe3fi3ce5z
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/01_create_F2.ipynb
- Delimitter: `|`
- Number of observations: 30,682
- Columns (as delimitted names, including `n`, `p`', `i`, `dfidf`, `porter_stem`, `max_pos` and `max_pos_group`, `stop`): n|n_chars|p|i|s|h|stop|stem_porter|stem_snowball|stem_lancaster|max_pos|max_pos_group|n_pos_group|cat_pos_group|n_p
- Note: Your VOCAB may contain ngrams. If so, add a feature for `ngram_length`.
- Top 20 significant words in the corpus by DFIDF: vimokshana, jayadhratha, upamanyu, kirata, athlete, sthulakesa, jatugriha, mountainfestival, vakavadha, rituparna, paushya, arjunavanavasa, utanka, pramadvara, ratio, dasarnakas, exit, assiduously, parana, accusation

Derived Tables

BOW (3)

A bag-of-words representation of the CORPUS.

- UVA Box URL: https://virginia.box.com/s/8y851qllyz5nh2i44cuh4dacatcmqbvm
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/02_TFIDF.ipynb
- Delimitter: `|`
- Bag (expressed in terms of OHCO levels): book_id, chap_id (OHCO[:2])
- Number of observations: 245,040
- Columns (as delimitted names, including `n`, `tfidf`): book_id|chap_id|term_str|n|tf|tfidf

DTM (3)

A represenation of the BOW as a sparse count matrix.

- UVA Box URL: https://virginia.box.com/s/rlhiab3wno4crd2zzscts8xyvzeg2vhs
- UVA Box URL of BOW used to generate (if applicable): https://virginia.box.com/s/8y851qllyz5nh2i44cuh4dacatcmqbvm
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/02_TFIDF.ipynb
- Delimitter: `|`
- Bag (expressed in terms of OHCO levels): book_id|chap_id (OHCO[:2])

TFIDF (3)

A Document-Term matrix with TFIDF values.

- UVA Box URL: https://virginia.box.com/s/nv8bkqnerwgh7zoo9kja92rw3oigy4xs
- UVA Box URL of DTM or BOW used to create: https://virginia.box.com/s/8y851qllyz5nh2i44cuh4dacatcmqbvm
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/02_TFIDF.ipynb
- Delimitter: `|`
- Description of TFIDIF formula ($\LaTeX$ OK): Sum for TF and Standard IDF

$$ \text{TFIDF}_{t,d} = \left(\frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \right) \cdot \log_2 \left(\frac{N}{\text{DF}_t} \right) $$

Reduced and Normalized TFIDF_L2 (3)

A Document-Term matrix with L2 normalized TFIDF values.

- UVA Box URL: https://virginia.box.com/s/72u8aqo5sxlet2wusagthwo4849jqhlk
- UVA Box URL of source TFIDF table: https://virginia.box.com/s/nv8bkqnerwgh7zoo9kja92rw3oigy4xs
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/02_TFIDF.ipynb
- Delimitter: `|`
- Number of features (i.e. significant words): 3080
- Principle of significant word selection: I have shortlisted terms based on their entropy (dh) values. Only terms with dh values above the 90th percentile (top 10 percentile most informative terms) are considered significant.

Models

PCA Components (4)

- UVA Box URL: https://virginia.box.com/s/4vkmqap2xxju345zm78xm7x0n5sos2eb
- UVA Box URL of the source TFIDF_L2 table: https://virginia.box.com/s/72u8aqo5sxlet2wusagthwo4849jqhlk
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/03_PCA.ipynb
- Delimitter: `|`
- Number of components: 10
- Library used to generate: Scikit-Learn - Decomposition
- Top 5 positive terms for first component: `shafts carwarriors pierced sanjaya army` (Depicting War)

- Top 5 negative terms for second component: `deities vidura penances vyasa brahman` (Depicting peace and spirituality)

PCA DCM (4)

The document-component matrix generated.

- UVA Box URL: https://virginia.box.com/s/gvyru078aaoshrm16a3hdkqla4gc5cxj
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/03_PCA.ipynb
- Delimitter: `|`

PCA Loadings (4)

The component-term matrix generated.

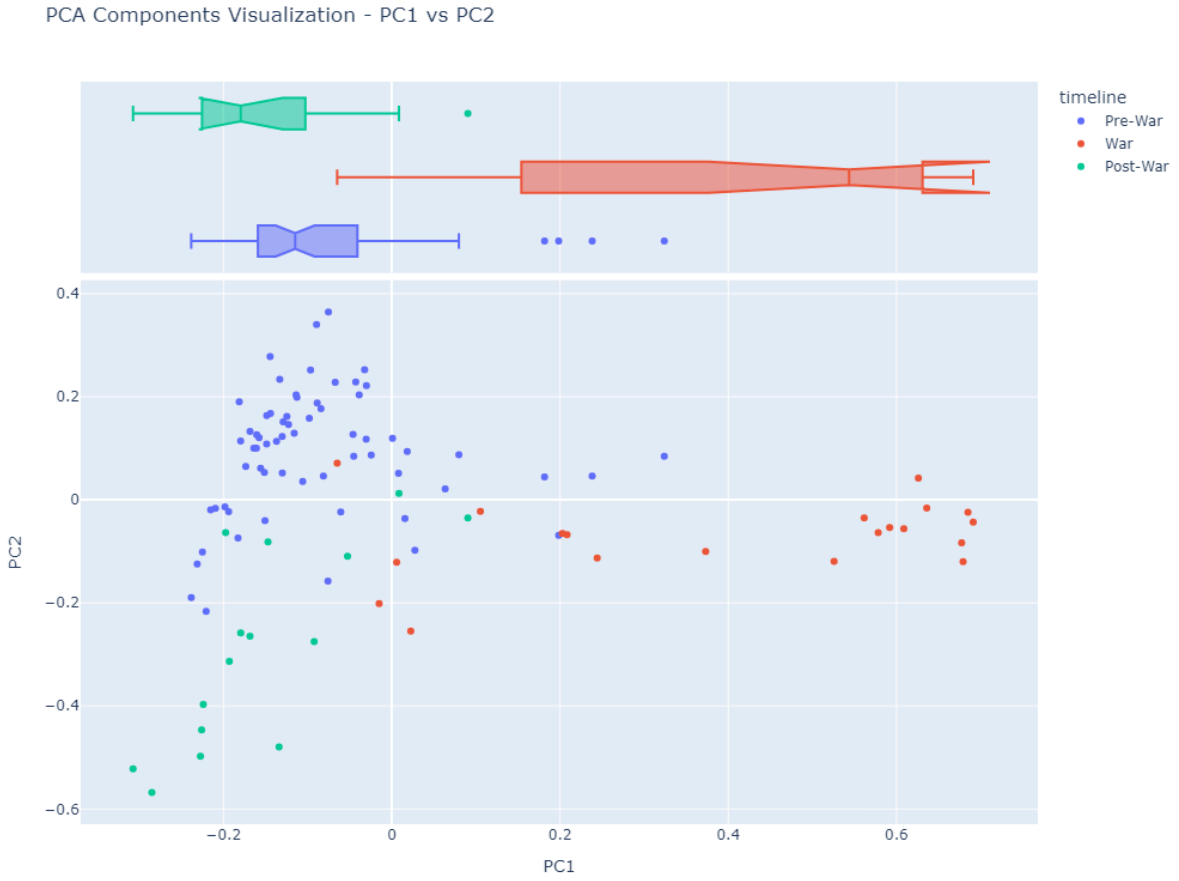
- UVA Box URL: https://virginia.box.com/s/ej8lmbqx00p5s3040gomr0e53mzbv2ok
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/03_PCA.ipynb
- Delimitter: `|`

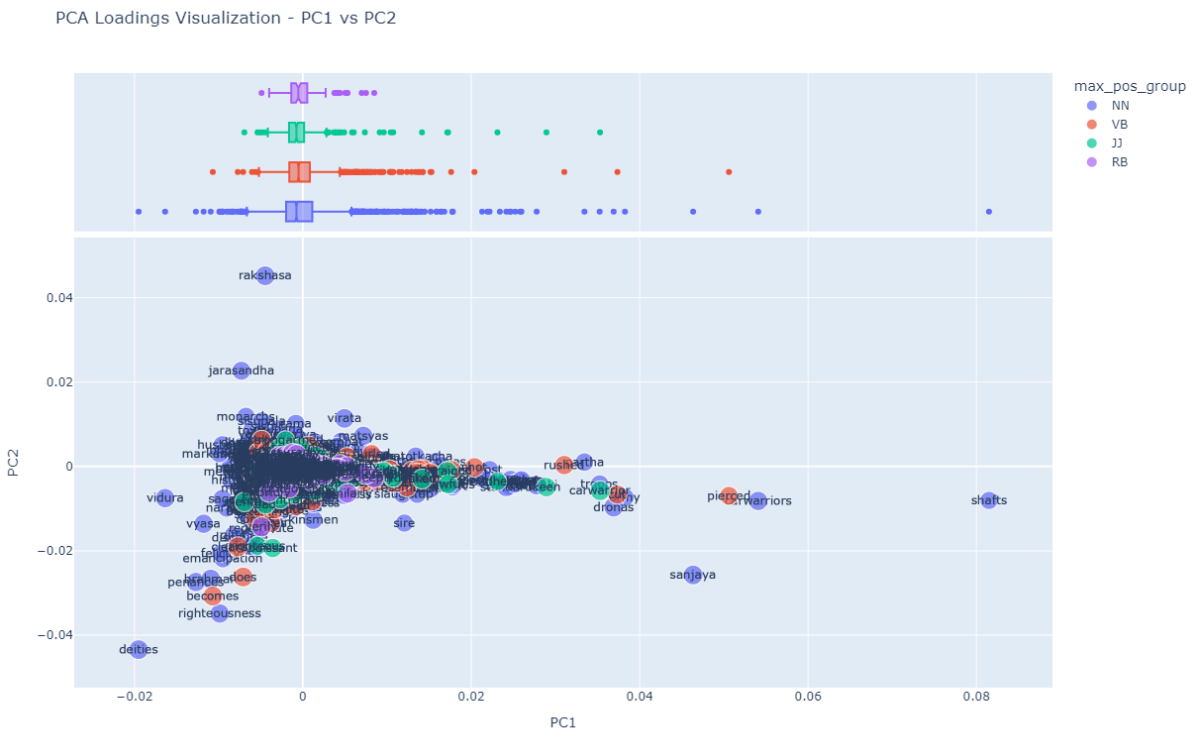
PCA Visualization 1 (4)

Include a scatterplot of documents in the space created by the first two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)





Briefly describe the nature of the polarity you see in the first component:

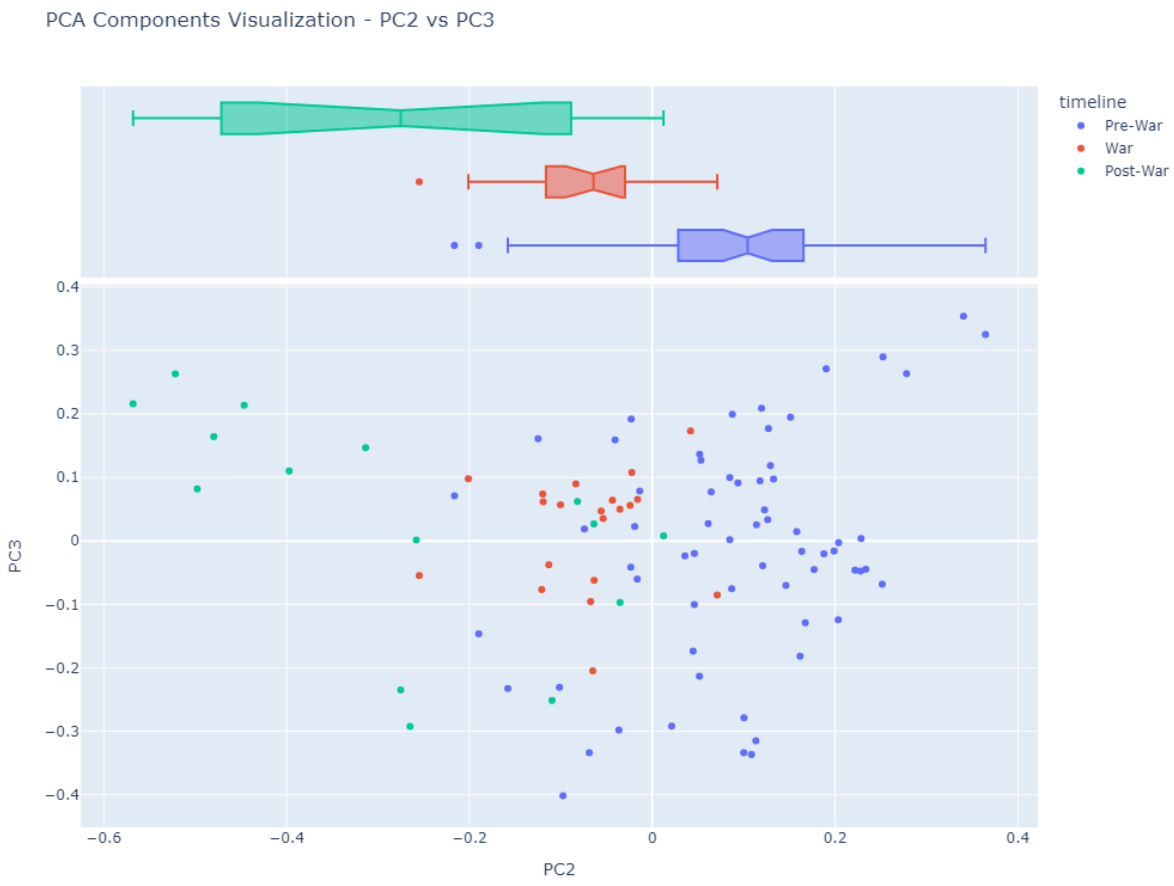
The first component very well distinguishes between the war and non-war parts of the saga. We see the War Timeline in the Positive PC1 whereas the Pre and Post War test in the Negative PC1. Looking at the loadings, we observe ancient war related terms like shafts, carwarriors, pierced, rushed, army, etc. in the positive PC1. On the other hand, we see words like deities, righteuosness, penances, brahman, etc. on the negative PC1 which is starkly oppositve from war and describes spirituality and peace.

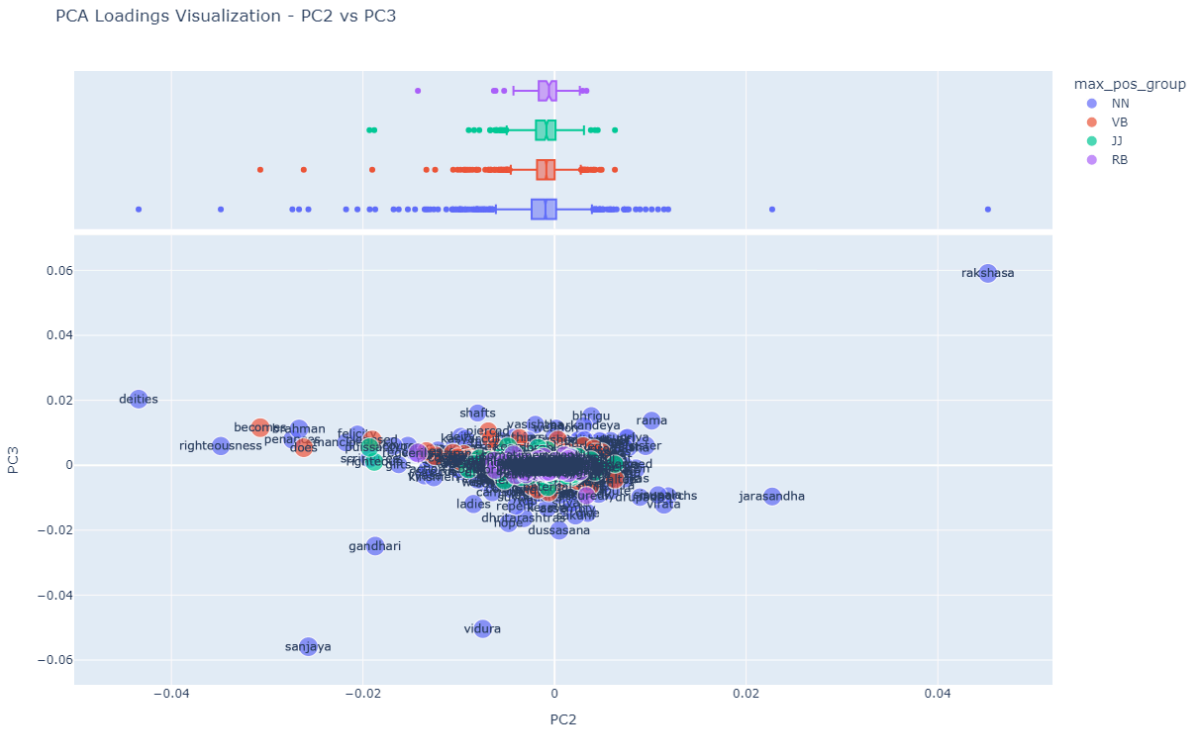
PCA Visualization 2 (4)

Include a scatterplot of documents in the space created by the second two components.

Color the points based on a metadata feature associated with the documents.

Also include a scatterplot of the loadings for the same two components. (This does not need a feature mapped onto color.)





Briefly describe the nature of the polarity you see in the second component:

The second component does something similar to the first component but this time it separates the Pre-War parts in the positive PC2 and the War and Post War text in the negative PC2. Looking at the loadings, it is interesting to note that the positive PC2 words are rakshasa, jarasandha, virata, sisupala, etc. which are antagonistic characters in the story whereas the negative PC2 again has words like deities, righteuosness, penances, brahman, etc. while this time these can be interpreted as supportive/protagonistic characters.

LDA TOPIC (4)

- UVA Box URL: https://virginia.box.com/s/b1fj1n789v9rfm5yitozin34qeshp50e
- UVA Box URL of count matrix used to create: https://virginia.box.com/s/we7sh2hlvfutett35r6clnw5r6qdqx4s
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/04_LDA.ipynb
- Delimitter: `|`
- Libary used to compute:
  - ScikitLearn - (CountVectorizer from sklearn.feature_extraction.text, LatentDirichletAllocation from sklearn.decomposition)
- A description of any filtering, e.g. POS (Nouns and Verbs only):
  - Removing stop words and archaic english words like 'thou', 'thee', 'hath', 'thy', 'art', 'ye', 'hast'
- Number of components: 10
- Any other parameters used: Max 100 Iterations, n_terms = 1008
- Top 5 words and best-guess labels for topic five topics by mean document weight:
  - T00: words race battle virtue kings
  - T01: acts knowledge mind creatures body
  - T02: weapons weapon energy gods celestials
  - T03: kings race monarch wealth city
  - T04: person life duties wealth world

LDA THETA (4)

- UVA Box URL: https://virginia.box.com/s/zbwlqftu5tikpgrlw3u7mz9b04xymjry
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/04_LDA.ipynb
- Delimitter: `|`

LDA PHI (4)

- UVA Box URL: https://virginia.box.com/s/gd5filhnc6gmp3g61dru5maolc20xkvb
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/04_LDA.ipynb
- Delimitter: `|`

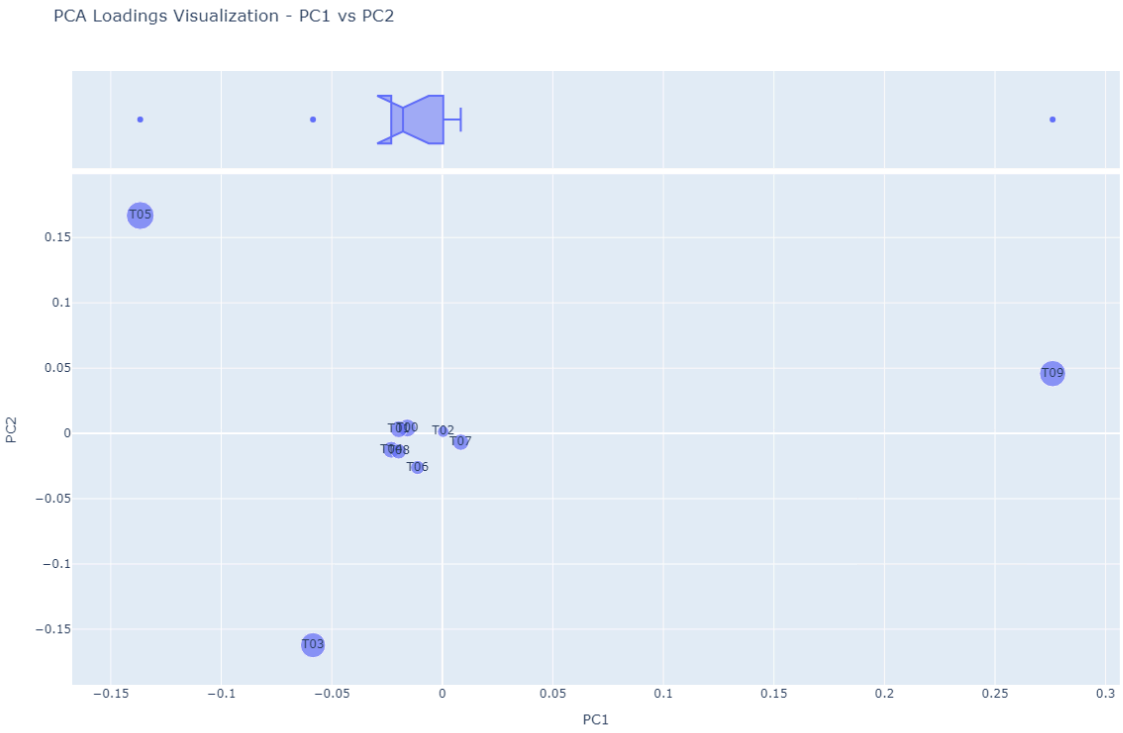
LDA + PCA Visualization (4)

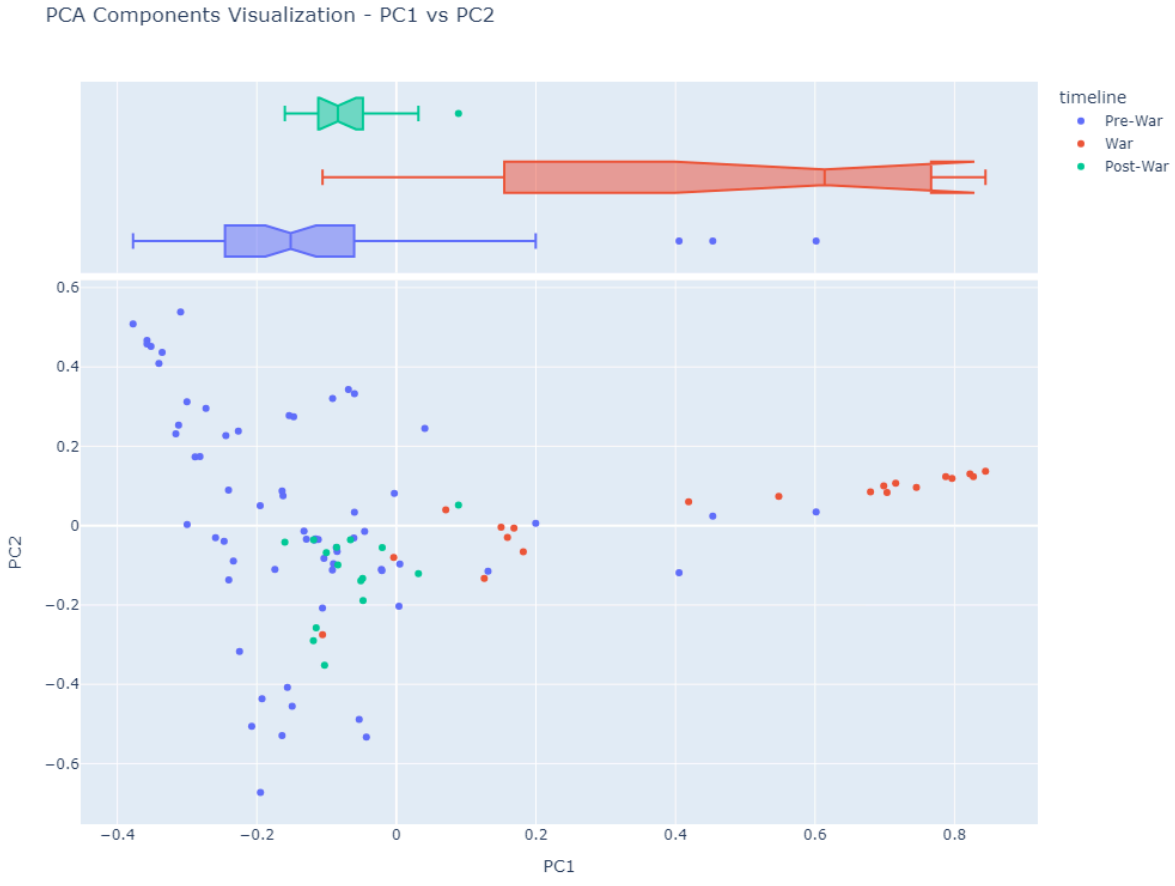
Apply PCA to the Theta table and plot the topics in the space opened by the first two components.

Size the points based on the mean document weight of each topic (using the THETA table).

Color the points basd on a metadata feature from the LIB table.

Provide a brief interpretation of what you see.





The PCA on the Theta Table gives us interestingly similar results to our previous PCA analysis. We get the topics which are more related to battle and war in the positive first Principal Component like Topic 9, 7 and 2 whereas on the other extreme PC we have topics 5 and 3 which relate to domesticity and kingdom related topics.

The PCS when augmented with the timeline of the of the narration also shows that Positive PC1 is all War related timeline whereas negative PC1 is Pre and Post War narration.

Sentiment VOCAB_SENT (4)

Sentiment values associated with a subset of the VOCAB from a curated sentiment lexicon.

- UVA Box URL: https://virginia.box.com/s/ijykn00uike9gcs0ddqmium39gqlumr1
- UVA Box URL for source lexicon: https://virginia.box.com/s/vo8t6payde8znrk3eatye8g5fb0r5cr2
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/05_Sentiment_Analysis.ipynb
- Delimitter: `|`

Sentiment BOW_SENT (4)

Sentiment values from VOCAB_SENT mapped onto BOW.

- UVA Box URL: https://virginia.box.com/s/bu6vw0sjtutnllvk3xh60k52my33iaxk
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/05_Sentiment_Analysis.ipynb
- Delimitter: `|`

Sentiment DOC_SENT (4)

Computed sentiment per bag computed from BOW_SENT.

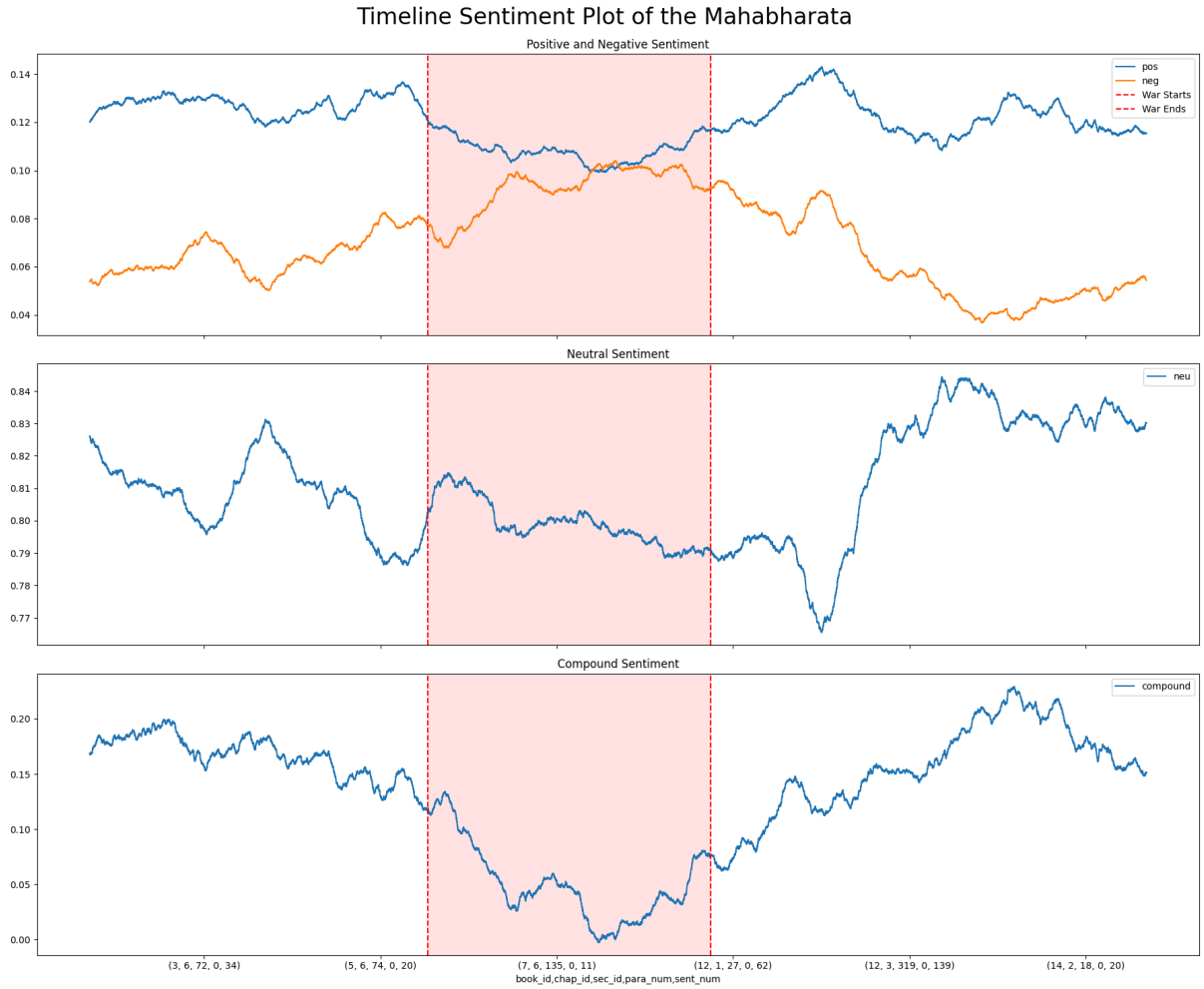
- UVA Box URL: https://virginia.box.com/s/p1kjetqqmycpcsrj6j9ohfdpmszh4wgu
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/05_Sentiment_Analysis.ipynb
- Delimitter: `|`
- Document bag expressed in terms of OHCO levels: OHCO[:2] book_id, chap_id

Sentiment Plot (4)

Plot sentiment over some metric space, such as time.

If you don't have a metric metadata features, plot sentiment over a feature of your choice.

You may use a bar chart or a line graph.



VOCAB_W2V (4)

A table of word2vec features associated with terms in the VOCAB table.

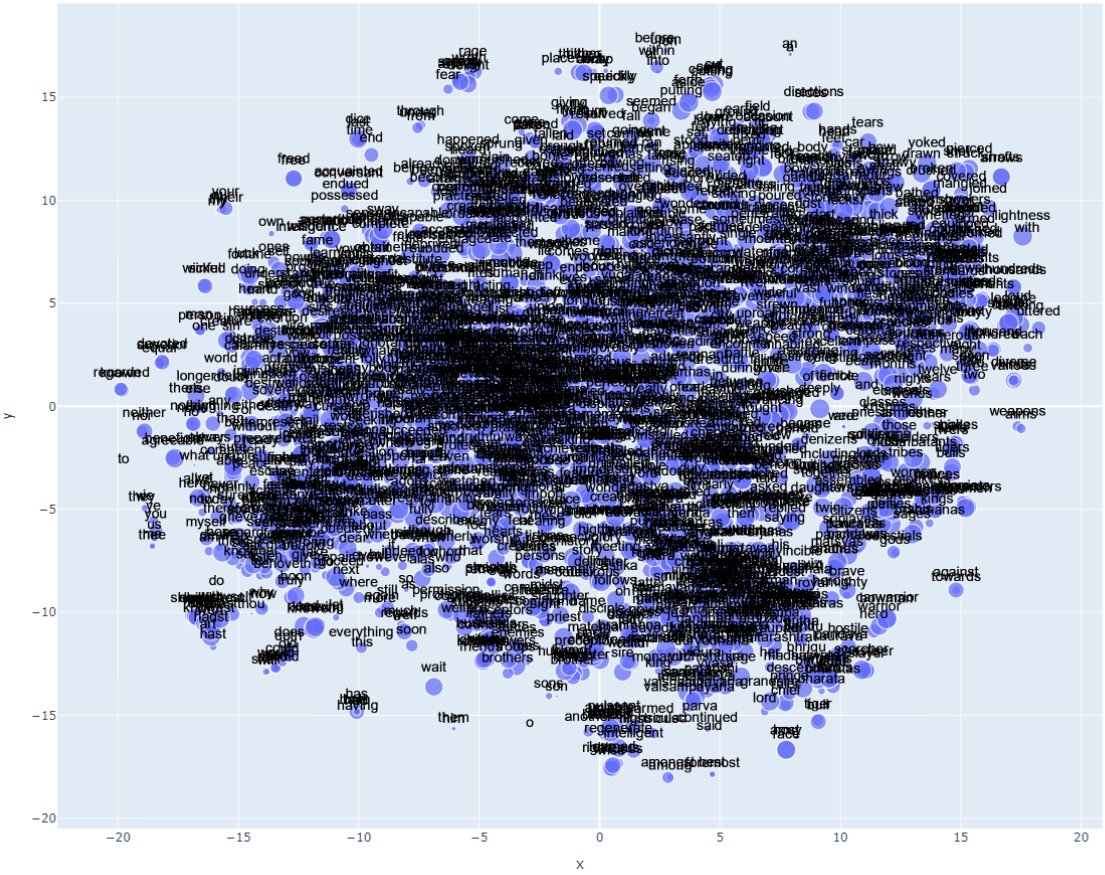
- UVA Box URL: https://virginia.box.com/s/jhz84zi1yoaqtjnyg64uwvlqn7deb9rb|
- GitHub URL for notebook used to create: https://github.com/vishugp/Mahabharata_ETA/blob/main/notebooks/06_Word2Vec.ipynb
- Delimitter: `|`
- Document bag expressed in terms of OHCO levels: OHCO[:2] - book_id, chap_id
- Number of features generated: 200
- The library used to generate the embeddings: gensim (word2vec from models and Dictionary from corpora)

Word2vec tSNE Plot (4)

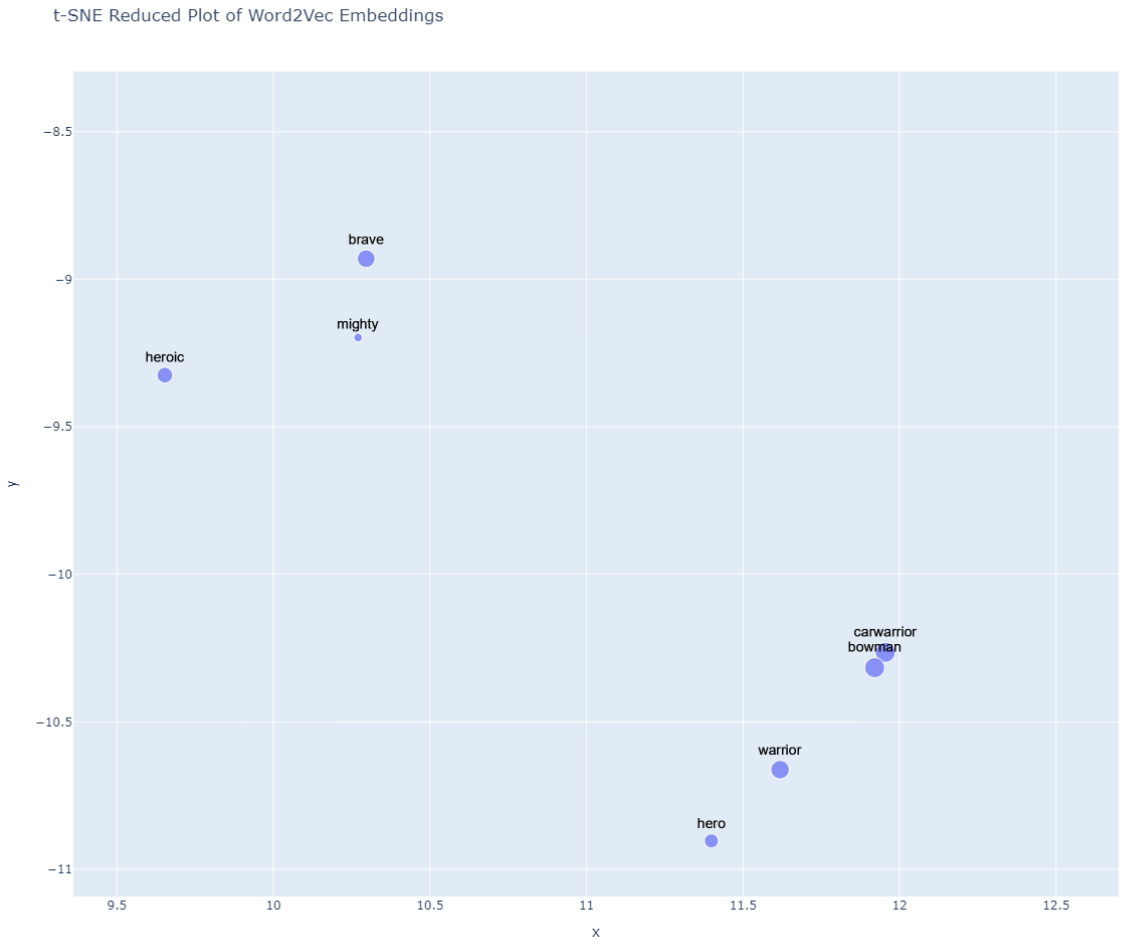
Plot word embedding featues in two-dimensions using t-SNE.

Describe a cluster in the plot that captures your attention.

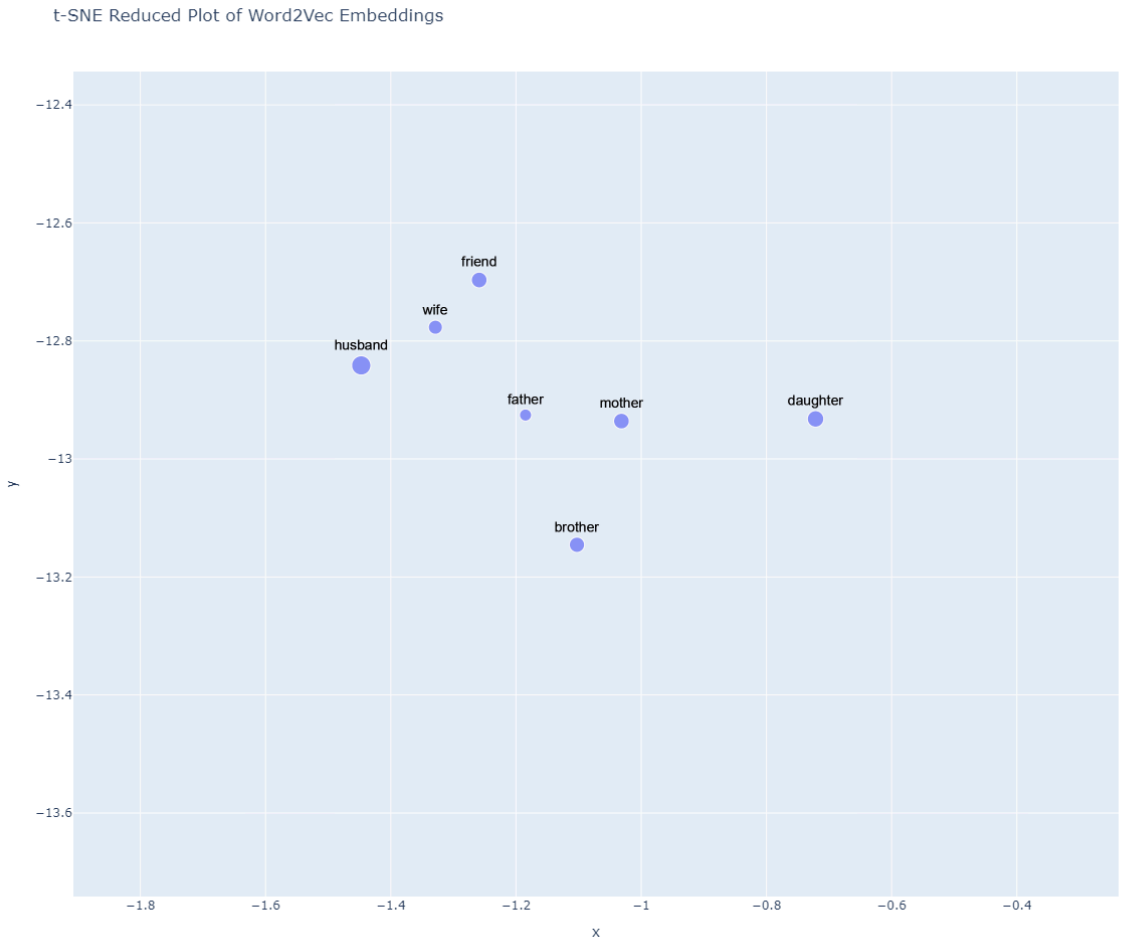
t-SNE Reduced Plot of Word2Vec Embeddings



It's good to see that all emotions cluster together in the plot



All warrior terms appear together



All human relational terms appear together which is nice.

Riffs

Provde at least three visualizations that combine the preceding model data in interesting ways.

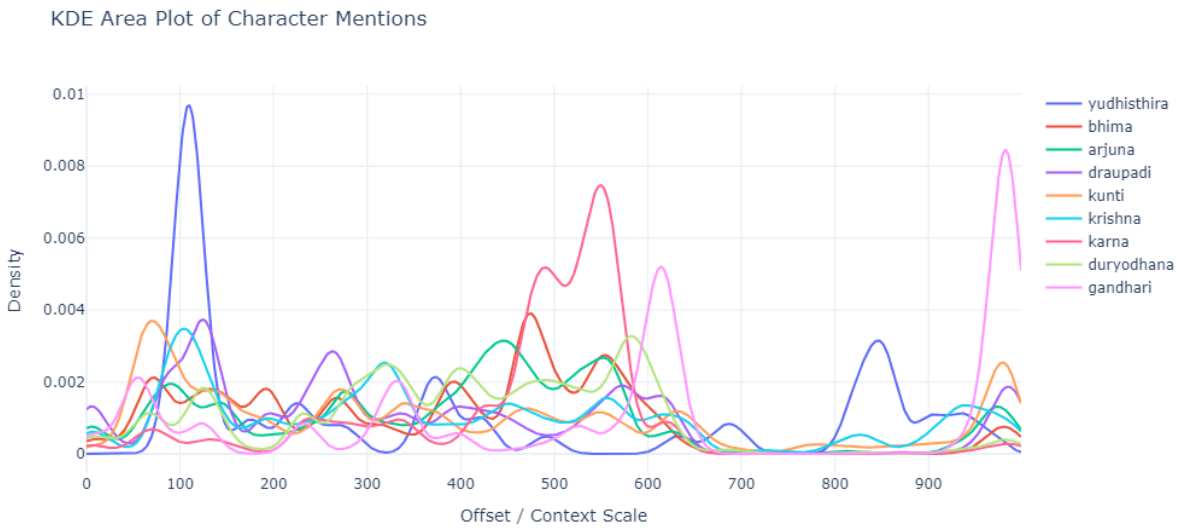
These should provide insight into how features in the LIB table are related.

The nature of this relationship is left open to you -- it may be correlation, or mutual information, or something less well defined.

In doing so, consider the following visualization types:

- Hierarchical cluster diagrams
- Heatmaps
- Scatter plots
- KDE plots
- Dispersion plots
- t-SNE plots
- etc.

Riff 1 (5)

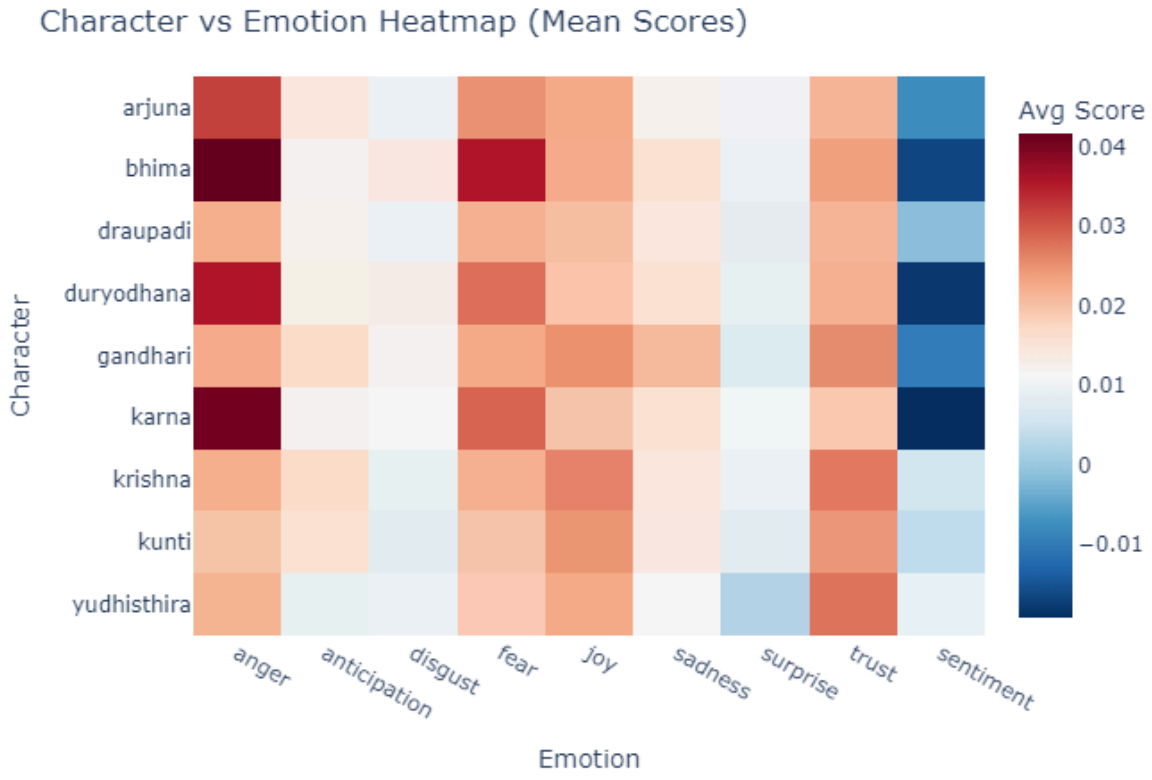


The above plot illustrates the frequency of appearances of key characters throughout the Mahabharata.

Notably, Yudhishthira who is the eldest of the Pandavas, shows high density in both the Pre-War and Post-War Phases, reflecting his pivotal role in shaping the events leading to and following the Kurukshetra War. In contrast, Karna who is the warrior born to the Sun God, has a concentrated appearance primarily during the War Phase ending with in his death during the war.

Gandhari, the mother of the Kauravas (antagonists) exhibits two distinct peaks in her presence one immediately after Karna's death and another at the end of the epic when her curse on Krishna comes to fruition and leads to the annihilation of his entire bloodline.

Riff 2 (5)



The heatmap visualization presents the mean emotional scores for key characters in the Mahabharata, revealing several noteworthy patterns.

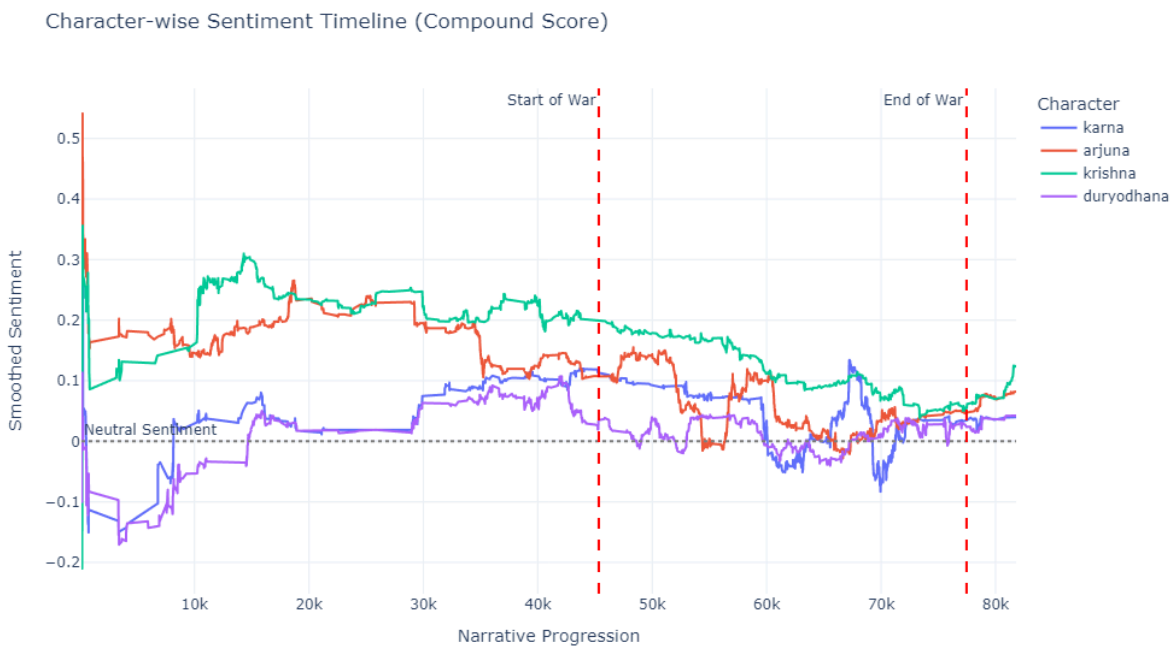
Characters associated with stronger anger include Bhima, Karna, and Duryodhana which aligns with their narrative roles and temperaments. Bhima and Karna also register the highest fear scores, suggesting significant internal conflict despite their warrior status.

Joy appears most prominently in Yudhishthira, Krishna, and Karna demonstrating emotional complexity even among antagonistic characters. Trust scores are highest for Krishna and Yudhishthira which is consistent with Krishna's divine guidance role and Yudhishthira's reputation for honesty.

Overall sentiment scores are most negative for Bhima, Karna, and Duryodhana, reflecting their tragic character arcs throughout the epic. Draupadi displays moderate values across emotional categories, representing her multifaceted experiences throughout the narrative.

This quantitative analysis interestingly supports the traditional interpretations of these characters while also providing numerical evidence of their emotional complexity.

Riff 3 (5)



This above plot shows how the sentiment in the text is associated with the character in the scene.

In this emotional analysis of the Mahabharata, we see how the protagonist pair (Krishna and Arjuna) and antagonist friends (Duryodhana and Karna) have their sentiment lines dramatically intertwine during the war, especially around Karna's controversial and unjust death.

Before the war, Krishna stayed mostly positive while others fluctuated. Duryodhana began negatively but improved over time. During battle, everyone's emotions became unstable and started to overlap, showing how war blurred the lines between heroes and villains. Karna's sentiment drops dramatically around his death scene (65-70k mark), when he was unfairly killed while defenseless—a moment that challenges our sense of right and wrong. After the war, all characters' emotions settle near neutral territory, suggesting that in the end, neither side found true happiness in victory or defeat.

Interpretation (4)

Describe something interesting about your corpus that you discovered during the process of completing this assignment.

At a minumum, use 250 words, but you may use more. You may also add images if you'd like.

Working with the Mahabharata corpus for this assignment was a very engaging experience. The text is incredibly rich covering the drama across families before, during, and after the war in a way that feels surprisingly realistic given its mythological context. The original corpus was quite clean, but to make the most of it structurally, I had to add a few lines to the text files so that I could define chapters by their Upaparva boundaries which corresponds to the Chapter level between 18 books to 2000+ sections. This made it easier to use chap_id as a key level in the OHCO and get meaningful Bag-of-Words groupings at the chapter level—which turned out to be very effective.

All the methods I applied gave very interesting results! For example, LDA topic modeling gave clearly interpretable clusters. Each topics specifically captured unique themes like battle, grief, heroism, spirituality, materialism which are very reflective of the core themes of the epic.

Another highlight was using t-SNE on the Word2Vec embeddings. The spatial clustering of related concepts was very clear, I demonstrated some examples in the tSNE section before. Playing around with the Completing analogy and Similarity finding was interesting too.

Sentiment analysis was also fun to explore, especially seeing how the emotional tone shifted across chapters and characters. Seeing how the War section shifted the sentiment of antagonists and protagonists closer was intriguing.

Overall, this project made me appreciate how text analytics and data science can really help unpack and understand large, complex corpora like the Mahabharata. It was both technically interesting and creatively satisfying to work on. Once I get more time, I will definitely try to take this project further. One direction I'm particularly interested in is applying Named Entity Recognition (NER) to map character relationships and locations more systematically. This could help track interactions and movements across the storyline, which are currently embedded in narrative text. It would also be interesting to try something which incorporates deeper temporal analysis or even attempt network graphs of character co-occurrence!

Thank you for such a thoughtfully designed course, it has truly empowered me to explore text data with purpose and confidence!