Ancient Texts NLP & Classification Project

Insights from Counts, Frequencies, and Weights

All code at https://github.com/izsolnay/Ancient_NLP

ISSUE / PROBLEM

Can modern tools provide insights into ancient text corpora?

Objective

Develop a reliable machine learning model which can classify texts into genres

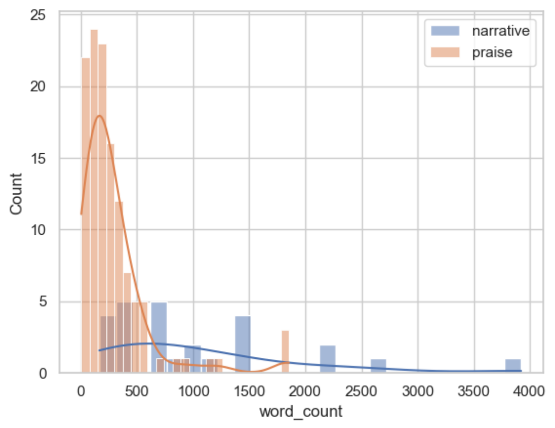
Steps for this stage

- Prepared text for word count, frequency, and weight analyses
- Introduced new features `‘lemmatized‘`, `‘word_count‘`, `‘token_count‘`, and `‘split_tokens‘`
- Generated visuals
- Extracted top bi- and tri-grams per text
- Investigated top bi- and tri-grams by `‘B_category’`, `‘God’`, and `‘Person’`

IMPACT

Text lengths vary considerably

- After lemmatization, most genres had on avg well below 1000 words per text
- Genres ‘City laments’ and ‘Narratives featuring heroes’ had highest word counts with median counts above 1500
- Praise texts were on average much shorter that narrative texts



To address

- Grossly varying text lengths
- Previous insight that some genres have only 1 or 2 texts attributed to them

RESPONSE

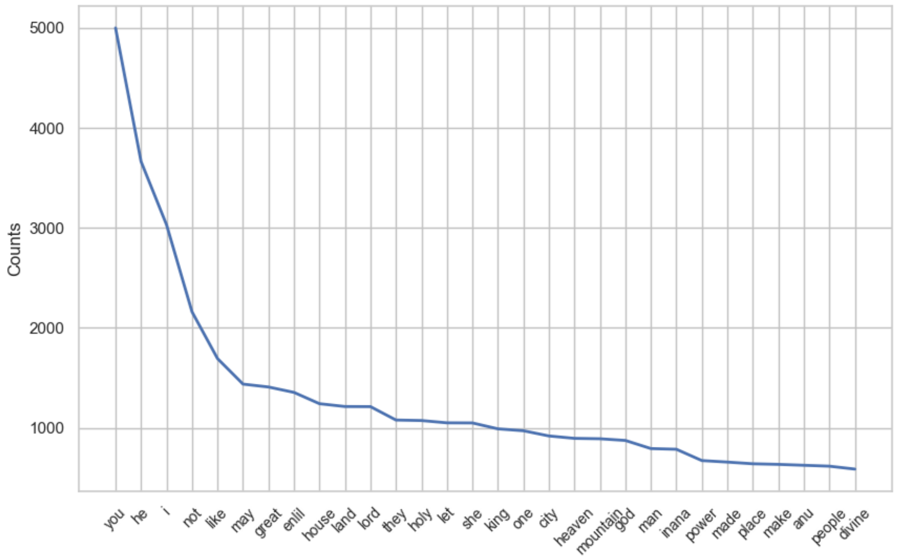
Sample of top words by count
(excluding pronouns)

- Narrative genres: 'Inana', 'Enlil, ' and 'lord'
- Praise texts: 'may', 'great', 'lord,' and 'like'
- Goddess Inana: 'heaven,' and 'holy'
- God Ninurta: 'Lord', 'Enlil,' and 'mountain'

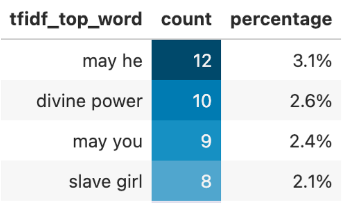
- The moon as Nanna: 'king,' 'great,' and 'holy'
- The moon as Suen: 'heaven', 'divine', 'power', 'light', and 'earth'
- The moon as Nanna-Suen: 'house,' and 'Enlil'

KEY INSIGHTS

Top words by count for all texts were 'you', ‘I', 'he', 'not', and 'like'

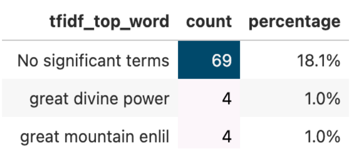


Top bi-grams



tfidf_top_word	count	percentage
may he	12	3.1%
divine power	10	2.6%
may you	9	2.4%
slave girl	8	2.1%

Top tri-grams



tfidf_top_word	count	percentage
No significant terms	69	18.1%
great divine power	4	1.0%
great mountain enlil	4	1.0%

'divine power’ notable top weighted term in bi- and tri-grams

- In 35 top tri-grams and 10 top bi-grams of all texts
- In 25 ‘Royal praise poetry and hymns to deities on behalf of rulers’
- In 4 trigrams of texts featuring person Ishme-Dagan
- In 5 trigrams of texts featuring the goddess Ninisina