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

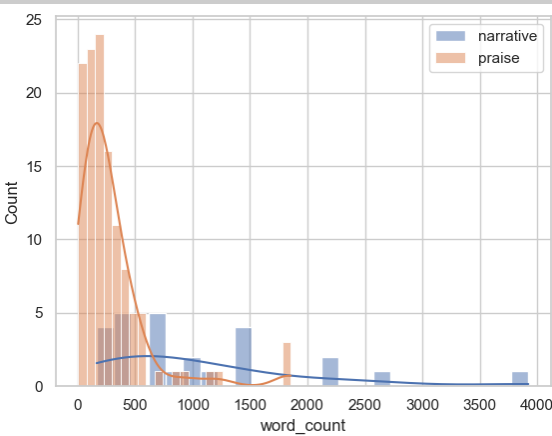Develop a reliable machine learning model which can classify texts into genres

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
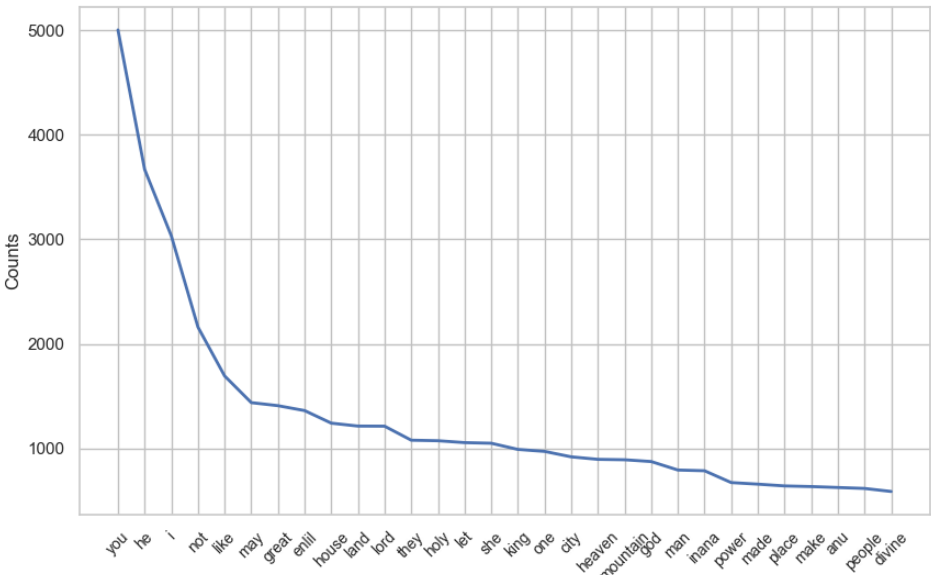
RESPONSE

Sample of top words by count (excluding pronouns)

- Narrative genres: 'like' and 'Inana'
- Praise texts: 'great', 'holy', 'may', and 'land'
- The goddess Inana: 'Inana' and 'heaven'
- The god Ninurta: 'Ninurta', 'Lord', and 'Enlil'

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

KEY INSIGHTS

Top bi-grams

tfidf_top_word	count	percentage
may he	13	3.4%
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	68	17.8%
great divine power	5	1.3%
great mountain enlil	5	1.3%
look upon city	4	1.0%

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

➢ In 34 top tri-grams and 10 top bi-grams of all texts
➢ In ~20 '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