Homework 1

Maya K. Hess

September 2017

1 Topic

My dataset will facilitate cross cultural comparison regarding goddesses with different associations (e.g., fertility, night). For the end product I would like to include as much visual content as possible, so I am downloading the images included on the pages for future use. I chose this topic largely arbitrarily with a vague goal of eventually exploring keyword tf-idf information across the aggregated categories.

The root of the 'data' folder contains the information individual goddess wikipedia entries, stored in JSON files named by their page IDs. Each entry contains data as in the sample files included. There are also two larger files, 'associations.json' and 'cultures.json', containing the category contents of the association categories and culture categories, though these duplicate the information available in the 'data/categories' folder's JSONs (again named by IDs). I have included both the page IDs and page titles for ease of human reference, though this is redundant for automated use. 'data/images' contains the downloaded images scraped from the MediaWiki file pages referenced in the goddess wiki pages. This work comprised the bulk of my extra-API scraping learning. Some dead references exist, as for example links to files incorrectly classified as images; these will need to be checked for future use.

There are approximately 1200 goddesses included, 92 categories, and 1700 images.