Progress Report

Kate Li

2024-11-17

Project Overview

Package Name and Purpose

Given a list of racialized terms and a corpus of literary texts processed by BookNLP, CulturalMapNLP takes in the list of terms to assess trends in racial representation across a body of literary works. It does this by looking at the list of each term's frequency across the corpus, trends in the terms' collocates, and the manners through which characters are racialized. (Note this function assumes BookNLP has already parsed corpus files, and that all outputs produced by a parse are stored within an inputted directory. As such, given processing by BookNLP has been performed, this package would not benefit from additional integration with BookNLP.)

Progress Summary

Completed Work

kateli

Kate has written functions that allow package users to read in, standardize, and consolidate character data. Specifically, given a set of files produced by a BookNLP parse of a corpus (which we term here as "BookNLP files"), Kate has written functions that allow for the streamlining of a) extraction of character-relevant data for each text from a directory containing all BookNLP files for an entire corpus, and b) consolidation of this character-relevant data into an aggregated data frame that can be stored in the user's global environment. Major aspects of the documentation (e.g., parameters, intended uses, examples) have been written. No tests have been developed.

sttruong

Sang has written functions that allow for the execution of tf-idf and collocate analysis on the output of a BookNLP parse. In particular, Sang has written functions that allow users to 1) divide texts within a directory of BookNLP into subgroups by connecting referenced texts within the corpus to text metadata (e.g., novel year of publication, nationality of author, etc.), which allows users to determine if any group-based trends in word frequency (measured by tf-idf) emerge, and to graphically visualize any trends which emerge on a group-level basis.

Remaining Work

kateli

Kate still needs to find ways to more efficiently process data, in the case of large inputs. For example, if the corpus is large (e.g., 1000+ texts) or the size of each constituent text within the corpus is large (e.g., 80K+ tokens), then existing functions in the package may stall or take a long time to run. More efficient ways to run or process this data are needed, in the case of large corpuses. In addition, Kate needs to investigate additional functions that can be added to the package to ensure increased interpretability of package outputs. For example, if many collocates are produced from the collocate analysis and it is difficult to visualize/ideate what relations exist between collocates, Kate needs to determine what methods of dimensionality reduction (e.g., tSNE, PCA, etc.) might be most effective at allowing for intuitive user interpretation.

sttruong

Sang still needs to research lists of racialized terms, which can be used as input parameters for key functions (e.g., the collocate/tf-idf analysis) or cross-references for outcomes (e.g., modifiers used to describe key characters across the corpus). Sang might also be interested in expanding existing metadata on file to see if corpus subgroups can be made with more granularity (e.g., if the corpus is divided on the basis of the author's self-identified race rather than on the basis of nationality), which may involve further activities like web scraping. In addition, Sang will continue ideating on how character-based outcomes produced by BookNLP (e.g., key modifiers or possessions attributed to characters) can be integrated into more conventional analysis streams (such as the methods of tf-idf/collocate analysis).

Issues and Solutions

Technical Challenges

The team has not encountered any major technical difficulties, likely because many of the tasks completed so far have already been performed to some extent (e.g., Kate already has BookNLP parses and experience with consolidating outputted files from prior research, and Sang's analysis incorporates functions from existing text mining packages). However, the team anticipates difficulties later down the line, especially as large-scale computing comes into the picture (e.g., when the team attempts scaling up to larger-scale corpora).

Collaboration Challenges

This package is being developed primarily in conjunction with Kate's research activities at Stanford's Literary Lab. As such, Sang (who does not have existing affiliation with LitLab) is not as familiar with data that would be inputted specifically for this package (e.g., corpus format/processing methods like BookNLP), and does not have the same intuition for outcomes that might be of interest to literary scholars. However, because Sang has a background in NLP/text mining, he is very comfortable with administering and developing computational functions to assist in package development. He is also capable of recommending specific functions/dependencies and improving existing codeflow, to make package development more efficient. As a result, work has been divided so Kate can structure how the package can be built to maximize effectiveness, whereas Sang has the ability to provide feedback on package development and give technical feedback where necessary.

Changes to Project Plan

Scope Modifications

No major changes have been made to the project scope. However, we are considering incorporating additional measures, such as web scraping, to expand the metadata associated with a corpus.

Updated Timeline

We have decided to move one of our milestones, which is testing on Large-Scale Corpora, a week earlier (Nov 30th to Nov 25th). This is because we now anticipate challenges with large-scale data processing (e.g., if the corpus is large) and want more time to build in features regarding memory management/more efficient data processing in the likely case of large corpora.