Project proposal

Project Group 8: Baiheng Chen, Kai Cui, Tuocheng Chen, Yuda Ding, Ziyi Song

Code to get data

Here is the code to get data from the Gutenberg ebooks project. We use the `wget` command to download the data. The code is written in bash and is run on the cluster.

- submit.sh

```
#!/bin/bash

rm -rf slurm_out;mkdir -p slurm_out
rm -rf data;mkdir -p data

sbatch --output="slurm_out/slurm-%A_%a.out" \
        --error="slurm_out/slurm-%A_%a.err" \
        --array 1-3000 getData.sh
```

- getData.sh

```
#!/bin/bash

base_url="http://aleph.gutenberg.org"
n=$SLURM_ARRAY_TASK_ID
start=$(echo "20 * ($n - 1) + 10001" | bc)
end=$(echo "20 * $n + 10000" | bc)

download_file() {
    dir1=$(printf "%d" $(($1/10000)))
    dir2=$(printf "%d" $(($1%10000/1000)))
    dir3=$(printf "%d" $(($1%1000/100)))
    dir4=$(printf "%d" $(($1%100/10)))

    for suffix in "" "-0" "-8"; do
        url="${base_url}/${dir1}/${dir2}/${dir3}/${dir4}/${1}/${1}${suffix}.txt"
        wget --spider $url

        if [ $? -eq 0 ]; then
            wget -P data $url
        break
        fi
    done
}
for i in $(seq $start $end);do
    download_file $i
done
```

About our dataset

Dataset Overview

We are using text data from e-books, sourced from here. We have extracted about 60,000 e-books for certain statistical analyses. We used 3000 jobs for parallel downloading, ultimately obtaining 59,180 TXT files, totaling 22GB.It would be an intersting and challenging work to implement some valuable analysis with HPC on such a huge dataset.

Descriptions of the variables

All we have are text variables. Specifically, we have a fixed format head of each ebook including some basic information, here's an example:

- Title: Apocolocyntosis
- Author: Lucius Seneca
- Release Date: November 10, 2003 [EBook #10001]
- [Date last updated: April 9, 2005]
- Language: English
- Character set encoding: ASCII

And the whole books will go after the heads.

Statistical Methods

- What are the main styles of these books, more specifically, if one is more interested in a specific book, what other books can we recommend to him/her? (Text Style Analysis)
- Is there a more popular style during a specific period of time?
- For a specific author, how does his/her style change across time?
- What is the difficulty level of reading different books? (Can be studied by counting sentence length and vocabulary complexity)
- What are the emotional tendencies of different books?

Computational Steps for Each Method

Text Style Analysis for Book Recommendation

Extract stylistic features such as sentence length, vocabulary diversity, use of adj/adv, etc. Then apply clustering algorithms like K-means to group books with similar styles. We will parallelize by book in the feature extraction part and cluster with aggregated data.

Popularity of Styles Over Time

Organize the books by their publication date. Apply the style analysis from the starting point. Use time series analysis to detect trends. We will parallelize by decade or other time divisions suitable, analyzing each period in parallel.

Style Change Across Time for Specific Authors

Filter the books by the specific author. Then apply text style analysis for each book. Use regression analysis to track changes in style over time. The parallelization will be by book, with an extra step to order the results by time.

Difficulty Level of Reading Different Books

Calculate readability scores (such as the Flesch-Kincaid Grade Level or Gunning Fog Index) which consider sentence length and word difficulty. Consider different levels of difficulty level. We will parallelize this by computing the scores for multiple books at once, then aggregate the data.

Emotional Tendencies of Different Books

Count the frequency of words associated with different emotions. Construct emational analysis of each book. Apply clustering algorithms. Like in 1, the parallelization happens by book, with the results aggregated for later clustering.

Link to the Github repository

You can find our project here: https://github.com/kcui23/stat605_final_project/ Copy Link