\* This file describes code and data associated with Chapter 5 (“Discourse and Character”) of *The Values in Numbers: Reading Japanese Literature in a Global Information Age*.

**------- Data Files -------**

1) **“Data\Fiction\_Meta.xlsx”** – Metadata for all “Fiction” texts used in analysis. Roughly 1,900 titles are included in the corpus, including 74 in-copyright works that were digitized by hand. The bulk, however, come from the Aozora Bunko digital archive.

2) **“Data\Kindai\_Meta.xlsx”** – Metadata for the “Kindai Magazine” corpus used in the analysis. This corpus was provided by the National Institute for Japanese Language and Linguistics for research use and I do not have permission to share the underlying text data in any form.

3) **“Corpora\FictionW2VPreProcessed\”** – This folder contains all the “Fiction” texts used in the analysis in a pre-processed format (e.g., lemmatized, character names filtered out). I include the full corpus in this format so that the “Cluster Detection” procedure (see below) can be repeated using different parameter settings.

4) **“Corpora\AozoraFictionTokenized\”** – Contains all **out-of-copyright** “Fiction” works used in analysis, stored in a pre-tokenized format. **\* Note: Results of analysis cannot be reproduced with only these texts because in-copyright files are not included.**

5) **“Corpora\UnidicLemmaMerged\”** – Contains all **out-of-copyright** “Fiction” works used in the analysis, stored in lemmatized and tokenized format, with racial identifiers merged into unified terms as described in the chapter. **\* Note: Results of analysis cannot be reproduced with only these texts because in-copyright files are not included.**

6) **“results\_fic\_bootstrap\”** – This folder contains the 20 word-embedding models used as part of the bootstrapping procedure described below. It also contains the raw and processed output of the Cluster Detection procedure after running it on the “Fiction” corpus. “Fiction\_Clusters.xlsx” contains the re-formatted output of the procedure, and is automatically generated from a script in “PrePostProcessingCode.ipynb” notebook. “Fiction\_Clusters\_Summary.xlsx” contains this same information, manually reformatted for ease of analysis and production of semantic grids.

7) **“results\_kindai\_bootstrap\”** – This folder contains the 20 word-embedding models used as part of the bootstrapping procedure described below. It also contains the raw and processed output of the Cluster Detection procedure after running it on the “Kindai Magazine” corpus. “Kindai\_ Clusters.xlsx” contains the re-formatted output of the procedure, and is automatically generated from a script in “PrePostProcessingCode.ipynb” notebook. “Kindai\_Clusters\_Summary.xlsx” contains this same information, manually reformatted for ease of analysis and production of semantic grids.

8) **“WordLists\”** – This folder contains various word lists used in the analysis, including manually curated stopwords lists for the “Fiction” and “Kindai Magazine” corpora; a list of proper names generated from the magazine corpus in order to exclude them from the word embedding models (hand checked); a list of character names generated from the fiction corpus in order to exclude them from word embedding models (also hand checked); and a list of place names that includes colonial cities and larger political entities that fell under Japanese imperial control. The latter is useful for exploring the frequency of reference to imperial locales in both corpora.

9) **“ClusterContextsFiction\”** – This folder contains output from the analysis of identified clusters (e.g., voice) as they occur around specific racial identifiers (e.g., Native) in the “Fiction” corpus. For each context, metadata for the source text is provided along with the corresponding original passage. Also included in this folder is a summary of the contexts in which words from the voice cluster occur with words for “Native.”

10) **“ClusterContextsKindai\”** –This folder contains output from the analysis of identified clusters (e.g., face) as they occur around racial identifiers (e.g., Westerner) in “Kindai Magazine” corpus. For each context, metadata for the source text is provided along with the corresponding original passage.

**------- Code Files -------**

1) **“Ch5\_Code.R”** – An R file with scripts for reproducing all visualizations in the chapter, as well as some extended analysis of temporal trends in the data.

2) **“PrePostProcessingCode.ipynb”** – A jupyter notebook used for running various pre and post processing tasks on the corpora used in analysis.

3) **“CharacterAnnotationCode.ipynb”** – A jupyter notebook used to extract and annotate character names in the “Fiction” corpus.

4) **“Cluster\_Detection\_Fic.py”** – A python file which identifies significant semantic clusters based on comparing rates of co-occurrence with one racial identifier versus another in “Fiction” corpus, as described in the body of the chapter. Several parameters can be set when running this analysis, including the threshold to use for semantic similarity; the number of words to include in analysis based on overall frequency; the window size for observed contexts; the p-value threshold to use when determining significance; and the percentage of overlap that is allowed when selecting the clusters to test for significance.

5) **“Cluster\_Detection\_Kindai.py”** – A python file which identifies significant semantic clusters based on comparing rates of co-occurrence with one racial identifier versus another in “Kindai Magazine” corpus, as described in the body of the chapter. Parameters are the same as above. This file is provided for reference only, as the underlying text data cannot be shared.

6) **“utils.py”** –A python file containing various utility functions used in the Cluster Detection and bootstrap creation python files.

7) **“word2vec\_bootstrap\_fic.py”** – A python file that creates N bootstrap word embedding models on a corpus of pre-processed texts. The bootstrap models are used to control for variation in the word embedding process, as described in the chapter. Bootstrap models for the “Fiction” corpus are included in this repository.

8) **“word2vec\_bootstrap\_kindai.py”** – A python file that creates N bootstrap word embedding models on a corpus of pre-processed texts. Bootstrap models are not provided for the “Kindai Magazine” corpus.

**------- Corpus Processing and Analysis Steps -------**

Listed below are the steps taken to carry out the analysis described in this chapter.

1) All corpora were stripped of paratextual information and older *kanji* variants were normalized to the newer forms.

2) As part of processing, all texts were tokenized and words were reduced to their *lemma* forms using the Unidic dictionary. Code for extracting the lemma forms is in the jupyter notebook file “PrePostProcessing.ipynb” and requires installation of the MeCab python library and the Unidic dictionary for contemporary Japanese.

3) General word frequencies, as well as specific frequency lists for character and proper names, are generated from the lemmatized corpora. Unidic provides additional support for this task as it identifies proper nouns.

4) For the purposes of analysis, variation in racial identifiers is minimized by replacing variants with a single, unified term (e.g., “Japanese”). The code for doing this is included in the jupyter notebook “PrePostProcessing.ipynb,” and the subsequently “merged” form of the files is stored in the “Corpora\UnidicLemmaMerged\” folder. All subsequent analysis relies on these modified texts.

5) The “word2vec\_bootstrap” files generate 20 bootstrap word embedding models for each of the corpora, excluding high-frequency proper names and character names. The “Cluster\_Detection” files are then used to build semantic clusters according to specific parameters and to then test if a cluster is significantly associated with one racial identifier versus another. The whole procedure is described in greater detail in the body of chapter 5. Results of the cluster detection process are output to a file whose name corresponds to the parameters used for building the semantic clusters (e.g., “kfree\_thresh0.69\_numtest5000\_window20...”). The output file lists all significant clusters for each pair of racial identifiers compared (i.e., “Korean” versus “Japanese”). For every word in a cluster is given 1) the relative frequency of the word across all occurrences of the first racial identifier (e.g., “Korean”) and 2) the relative frequency of the word across all occurrences of the second racial identifier (e.g., “Japanese”). The first value is used in the following step as a proxy for the strength of a particular semantic cluster (i.e., the value is summed across all words in that cluster).

6) The raw output from the Cluster Detection procedure is reformatted (using code snippet from “PrePostProcessing.ipynb”) and manually cleaned to create a Summary spreadsheet. This allows for easier analysis and the generation of semantic grids using the “Ch5\_Code.R” file.

7) Once significant clusters are identified, use a code snippet from “PrePostProcessing.ipynb” to retrieve passages where words from a cluster appear with a specific racial identifier. Examples of this output can be found in the “ClusterContexts” folders. In addition to showing contexts in their lemmatized form, this code also tries to retrieve the original passage from each text. Note that the current code does not always retrieve the correct original passage and so manual search may also be necessary.

8) A similar code snippet was used to retrieve passages where words from a cluster appear with a verified character name. Prior to this, it is necessary to identify and annotate character names in selected texts using code in “CharacterAnnotationCode.ipynb.”