ISMLA 17/18 Exercise 03: Investigating Kanji Overlap in Movie Subtitles

1 Project Setup

- Include the dependency on UIMAFit to a new Maven project.
- Add `subtitles-collection.jar` to the project classpath.
- Add `kanji-frequencies.jar` to the project classpath.

2 Implementation Steps (All Relevant Information on the Slides)

1. Write a UIMAFit analysis engine which annotates kanji (compounds) above a certain frequency rank. The engine must be configurable by language, paths to character and token frequency lists, and a rank threshold.
2. Set up a UIMAFit pipeline consisting of our `ParallelSubtitlesReader` and two instances of your Kanji annotator configured to operate on the Chinese and Japanese language versions.
3. Call your pipeline via `SimplePipeline.iteratePipeline(...)` to get access to the JCas object for each movie, and store counts of the annotated Kanji in each movie and language version.
4. Run the infrastructure with zero minimum rank thresholds on both languages, investigate the shared Kanji returned for a two-language pair of the same movie, and a second one of different movies, and use your findings to motivate the use of a cutoff.
5. Use the stored counts to compute the separation quality statistic.
6. Maximize the value of the statistic by setting different threshold values.
7. Reinvestigate the shared Kanji for the best values, and comment on whether the situation improved.
8. Repeat all the previous steps for compounds instead of single Kanji (only slight modifications to the analysis engine are necessary).

3 Submitting your result

Come to one of us **as a group** to show and explain your output and code:

- Björn's office on Fri, November 24, 2-3 pm
- before or after the lecture on Mon, November 27 (half an hour each)
- before or after the exercise on Wed, November 29 (half an hour each)
- Johannes' office on Thu, November 30, 1-6 pm (one day later than usual)