UNIVERSITY OF TÜBINGEN, DEPARTMENT OF GENERAL LINGUISTICS

Homework 1: Readability Formulas

Zarah Weiss, Xiaobin Chen, and Detmar Meurers

Due November 6th, 2019, 10am

1 IMPLEMENT THE FLESCH-KINCAID FORMULA

$$Grade = 0.39\frac{words}{sentences} + 11.8\frac{syllables}{words} - 15.59 \tag{1.1}$$

1.1 Input and Output Requirements

Implement a program that automatically calculates the Flesch-Kincaid readability formula for English texts. Use the version of the formula given above.
Your program should take the following input arguments (in this order):

1. An input directory containing plain text files containing a single text each
2. The name of the output csv file to which the results should be saved

Your program should output a single csv file (as specified by the user in the input) with the following columns:

File the file name

N.Sentences the total number of sentences in the text file

N.Words the total number of words in the text file

N.Syllables the total number of syllables in the text file

Words.Per.Sentence the average sentence length in words

Syllables.Per.Word the average word length in syllables

Grade.Score the Flesch-Kincaid grade level

1.2 Resources

You may write your code either in Java or in Python 3. Use one of the following APIs for sentence segmentation and tokenization:

- **OpenNLP** version 1.9.1 using the corresponding off-the-shelf model for English[1]
- **Stanford CoreNLP** version 3.9.2 using the corresponding off-the-shelf model for English[2]
  - Note: **don't** use the model English (KBP)
- **NLTK** version 3.4.5[3]
  - Use the nltk.tokenize.punkt sentence segmentizer and
  - the nltk.tokenize.word_tokenize tokenizer with the pretrained English model.
- **spaCy** version 2.2[4]
  - use the English model en_core_web_sm

Since these tools may differ in the segmentation they produce, we will evaluate the output of your code against results that we obtain using the same tool using the model and version specified here. If you choose to use another tool, we will match the results against the one that yields the most similar results to yours.

1.3 Submission

Submit your executable code via moodle in the format specified under *General Submission Requirements* at the end of this document. The submission deadline is

- **November 6th, 2019 at 10am**

1.4 Hints

- Do not include punctuation marks in your word counts.
- A good way to count syllables is to count the number of vowels. But be aware of diphthongs and the silent e at the end of words.
- It is sufficient to obtain a reasonable approximation of the actual number of syllables in a text.

2 Use your Code on the given Data Set

Run your code on the following test data and include the resulting csv under the name **results-task2.csv** in your submission.

- Test data in **newsela-sample.zip**
- Meta data in **articleInfo.txt**

The data is a subset of the Newsela data, a leveled news data set.[5] Note that not all files named in the meta data are included in the data sample.
Important: You do not need to use the meta information in this assignment, but if you are curious you can match your readability scores against the newsela reading scores. If you want to map the scores to grade levels, you can use this conversion guide:

- `https://support.newsela.com/item/supportArticle/grade-to-lexile-conversion`

[1] `https://opennlp.apache.org/download.html`
[2] `https://stanfordnlp.github.io/CoreNLP/download.html`
[3] `https://www.nltk.org/install.html`
[4] `https://spacy.io/`
[5] `https://newsela.com/`

3 Know Your Code

You need to be able to explain your pipeline in a few words in front of the class. We will choose a student at random to elaborate on their code. Relevant information are for example:

- Which components did you use in your pipeline?
- In which order are they executed?

4 General Submission Requirements

4.1 Submission

Homework can only be submitted via moodle before the end of the respective deadline. Your submission must always include a README file named **README.txt** including the following information:

- Full names and email addresses of all students who contributed to and want credit for this homework.
- The programming language use used
- A full list of the NLP APIs you used
- The name of your main file that should be called for execution of your code.

4.2 Programming Languages

- We assume that you program in either Java or Python 3. If you have a good reason to use another programming language, talk to us about it early as possible.
- If you use Java, make sure to use Maven to handle your dependencies (`https://maven.apache.org/`). Submit an executable jar file as well as your source code.
- If you use Python 3, make sure to use a virtual environment to handle your dependencies. If you are not familiar with virtual environments, you may finde the following links useful:
  - `https://www.geeksforgeeks.org/python-virtual-environment/`
  - `https://docs.python-guide.org/dev/virtualenvs/`
- Make sure that your code can be executed. If we cannot run your code, it counts as a failure.