Julius Caesar's *Gallic Wars*

This document provides a glimpse of natural language processing using Python and specialized libraries for classical languages.

Core question

By the middle of the first century BC Julius Caesar conquered Gaul. His own testimony in the *Commentaries on the Gallic Wars* names at least 33 major enemies of his campaigns. In Caesar's account, who does he consider as his fiercest rivals?

I am operating under the assumption that the number of times Caesar mentions his foe is proportional to just how important he is considered as an enemy. Thus, a figure mentioned often in the text is "fiercer" than one seldom mentioned.

Data source

Fortunately, many editions of Caesar's text are freely available online, but as always the quality and suitability of each for processing by natural language tools and Python vary considerably. For example, the original Latin is available at Project Gutenberg, the Perseus Project and the Internet Classics Archive. For my purposes, the plain text files have required only minimal transformation:

1. Download the text as separate books from one of the more reputable sources (e.g. of the three mentioned, the last is hosted by MIT)
2. Remove all header and footer material
3. Remove end of line breaks, except those that separate the text into paragraphs
4. Remove any numbering, so the remaining text consists only of text

One step I have not taken, but would like to complete once I gain some proficiency with the tools, is to *normalize* the text. That is, it would be very useful to verify that the edition has used consistent orthography for the Latin language, following modern standards for *i* and *j* and *u* and *v*, for example.

Data processing

The Classical Language Toolkit is the resource of choice for processing texts in Latin. Once the target text is read in, a number of operations are available, such as word and sentence segmentation, tokenization, lemmatization (i.e. finding the "root" word), counting vocabulary, finding unique words, and measuring lexical diversity.

Because I'm interested in word frequency—for example, how many times does the name Dumnorix appear in Book 2—the procedure should be simple: separate the target text into word tokens, lemmatize them, search for and count the number of returns for the character in question. Then compared to other major characters, determine who is "the winner" by number of mentions in that

particular book.

But, while ordinary verbs, adjectives, etc. were rendered to their root words by CLTK, proper names appear to be left in their original form. That is to say, the lemmatizer ignores Dumnorix and instead returns the original inflected forms Dumnorige, Dumnorigem, Dumnorigis, etc. Therefore, in order to accurately count the number of mentions, it is necessary to search the text for each form separately and sum the results.

Conclusion

It comes as no surprise that Vercingetorix is by far Caesar's staunchest opponent in the *Gallic Wars*, with 46 mentions in Book 7, followed by Ariovistus, with 42 mentions in Book 1, Ambiorix, with 38 mentions in Books 5 and 6, and so on. These results could just as well be had by opening a text editor and counting search returns, but the obvious power of natural language processing becomes apparent once we go beyond simple counting and begin to tackle larger analytical tasks.