Language Identification and the *textcat* Package in R

Shane R. Freeborn, Robert F. Gray

INFO 523 (Fall 2021), University of Arizona, Tucson, Arizona, USA

Identifying the language of a given text has long been a problem, even before the information age. It once required a human who knew the language to perform this task. With the advent of computation, new approaches to this issue were developed, and the interested party no longer needed to find a human with specific knowledge to identify the language of a text. In some languages, it simply involves recognizing an abundance of glyphs or symbols unique to that language. However, what if a single alphabet is used to write hundreds of languages (for example, the Latin alphabet). The unique-character-counting approach cannot work in this case, and we must turn to other methods. Much work in this field was done in the 1990s and early 2000s, using methods derived from machine learning [2]. The first viable example being Cavnar and Trenkle's N-gram-based approach [1]. A reduced version of this was developed by Hornik et al. [2], and a language-identification package for R was created in 1997 by van Noord [3]. The authors of this project first explore the creation of language-specific N-gram profiles using R and text identification through the minimization of a given distance metric. The textcat package [3] is then put to the test, and an assessment of its identifying potential is made. Lastly, recommendations of identification approaches are made based on the type of writing system involved.

References

[1] W. Cavnar, J. Trenkle. (1994). N-Gram-Based Text Categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161–175. Las Vegas, US.

[2] K. Hornik, P. Mair, J. Rauch, W. Geiger, C. Buchta, & I. Feinerer. (2013). The textcat Package for n-Gram Based Text Categorization in R. *Journal of Statistical Software*, *52*(6)

[3] G. van Noord. (1997). “TextCat.” URL: http://odur.let.rug.nl/~vannoord/TextCat.