Star Wars Blog

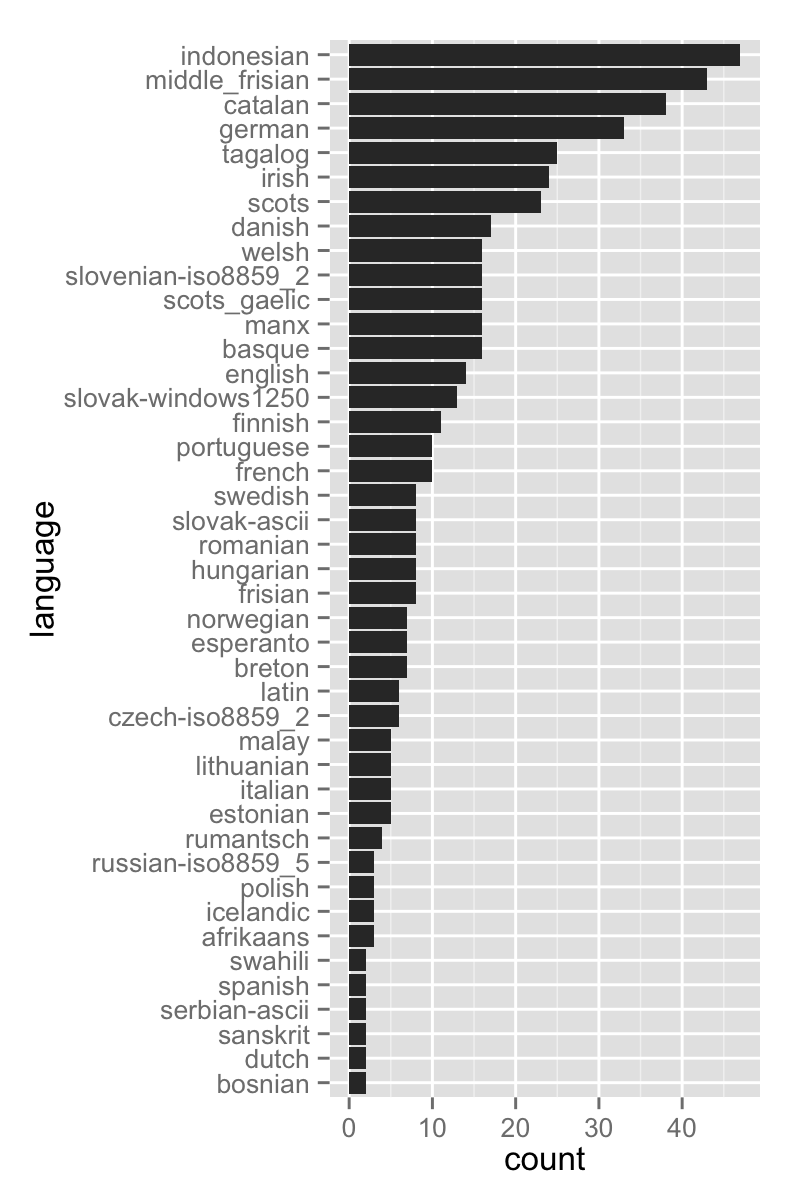
Richard Carter & Roman Popat

2 December 2015

(Introduction missing...)

We analysed the names of 513 Star Wars characters taken from a wikipedia page. These include names from the series of Star Wars books. We used a method of text categorisation commonly used to predict which language a piece of text belongs to. The text is chopped up into small bits, for example the word 'Star' can be chopped up into the letters S-T-A-R but also the letter combinations ST, TA, AR, STA and TAR. We then count up how many times each letter combination occurs. The rate at which letter combinations occur in a piece of text is compared to the rate at which the same letter combinations occur in different languages. Using this 'n-gram' method or model, we can say that a particular piece of text is most like a particular language, in the case of the word 'Star' this is icelandic.

We performed this process for the 513 Star Wars characters we could get hold of. In each case we derived a language classification and totalled the number of characters that matched each language (results below).



...

(possible caveat to end on...) Keep in mind that these algorithms are designed to categorise text documents that contain several paragraphs if not hundreds of pages. They can also be used to categorise technical documents into different subjects, categories, or academic disciplines. The accuracy is quite compromised when attempting to use a single word or a two word name.