NLP Project - Finding an author's nationality by spelling

Karl Rosengren
Intelligent Systems
Universidad Politecnica de Madrid

Keywords: Satoshi Nakamoto, NLP, Brittish, American, spelling

ABSTRACT

The project consisted of trying to find the nationality behind the mysterious pseudonym Satoshi Nakamoto which is the creator of Bitcoin. By analysing forum posts written from 2009-2011 with regards to differences in spelling between American and Brittish english. At the end, I found a great difference in the occurence of Brittish and American spelling. The definition I found on American spelling occured nearly six times more than that of the Brittish spelling.

CONTENTS

1 ABOUT THE PROJECT

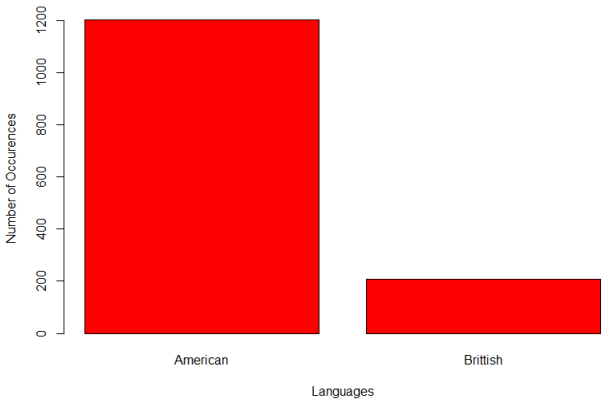
First of all I gathered my datasets which consisted in putting together a dataset of all forumposts written by Satoshi Nakamoto in the years 2009-2011, it was not good work, but it had to be done since I found no complete textfile of them all. As I did this I was still thinking about how to find the differences between the two types of the English language. I started researching ways to differ them and found two lists of usage of words that differed between the languages, e.g ”Lorry” and ”Truck” in Brittish and American. However, my dataset for the two languages consisted of only between 40-50 words if I were to take a guess. When I completed my first draft of the project the two small datasets still gave me no hits for all forum posts. I have not looked into it by hand so I don't know if the program was bad but I deemed it that this way would not function with such a small dataset of the two languages.

With this in mind I changed my approach to differences in words for certain meanings to that of differences in spelling and through defining the differences

Brittish	American
our	or
ise	ize
yse	yze
ence	ense

These differences would lead to different spellings of the words: ”Colour/Color”,”Apologise/Apologize”,”Analyse/Analyze” or ”Defence/Defense”.

My new approached proved to yield much stronger results and a result that is taking me towards two different conclusions As I've included in my program the script ends with plotting the following Bargraph to visualize the difference in occurence between the two types of spelling. In the provided bargraph we can see that the American type of spelling occurs nearly six times more than that of the Brittish way of spelling. What does that tell us?



2 ANALYSIS

I've come up with two ways to analyze the outcome of my program and the first is that; clearly, Satoshi Nakamoto is American and there might be some words that contain my defined substrings of Brittish even in the American language, which would explain the 200 times ”our, ise, yse, ence” occurs.

The second way to interpet this is that around $1/7$ times of writting another person is writing the answers on the forum than the other $6/7$ when an American person is writing. To decide on an absolute possibility is hard with this outcome and anything is possible, it is also the case that Satoshi Nakamoto could be a foreigner with english as his/her second language and has just learned multiple ways of writing in english.