Identifying Text Language Through Machine Learning

José Carlos Duarte

*duartejc2001@hotmail.com*

August 2016

Introduction

There are many different languages in the world. More than we can imagine. But, what is different between them? Only the words?

There are many differences between the languages including the use of different characters, terminators, fonts, grammar and other aspects.

Knowing the differences is possible to create engines to detect the language finding particular details for that language.

This is the idea of this work.

Solution approach

In this work I decided to take 4 types of information from text files:

1. First char of each word.
2. Last char of each word.
3. Length of each word.
4. Number of repeated cases, where 1, 2 and 3 are the same.

With this information I built a table with 4 dimensions where every word of the text file would be an entry.

It is an R4 space, where every word of the text file is a point. So for each text file we have a different distribution of points in this universe.

This way, it is possible to find similar distribution of points between two texts in same language.

To evaluate I have clustered the files, starting with same cluster points and then calculated the distance between these coordinates, comparing with other languages, expecting that the points resulted from files of same language will be closer than other languages files will.

Scope and limits

In this study 5 languages were used, all languages using Roman alphabet. The valid characters are restricted to **A**(65) to **Z**(90), and **@**(64) using this last one to group any extended character. An extended character are letters with diacritics that are not found in the English language, such as “é”, “ö” or “ç”. Since for these languages it is not very usual to start or finish with extended characters, this approach should not change significantly the results.

The list of languages is: French, German, Spanish, Portuguese and English. Italian was used in technical compare since English files were not available for this study.

Note: In appendix, a table linking characters to its decimal code is available.

Differences between languages

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **First character X Last character**  In this graphs showing relation between first and last characters is possible to notice some differences between the languages. No word in Portuguese ends by character F(70) to K(75). English has more variety in terminators than other languages. French has more words ending in Z(90) than other languages. Spanish and Portuguese points distribution are not so different. |
|  |  |
|  |  |
|  | **Last character**  These histograms further corroborate the previous conclusion: Portuguese and Spanish words rarely ends by chars 70 to 75 (F to K) or 85 to 90 (U to Z). In German texts, 85 to 90 (U to Z) rarely occur as last letter. In French, 70 to 75 (F to K) occurs less than other termination. English has words ending with almost any letter. |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **Occurrences**  The number of occurrences of the same type of word (starting with same character, ending with same character and having same length) has the same distribution for all languages. This makes sense even because in this case we have same content translated for all languages. But is possible to see that in English this kind of occurrence should be a little less comparing with other languages. |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **Word Size**  The distribution of word size seems different for some of these languages, mainly for German. This language and Portuguese has the longest words for this sample For all languages, the distribution follows the normal distribution except for German which is skewed right. |

Study case 1: Cooking recipes texts

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | In this case, nine cook recipes were used, were seven were used to feed the learning and two for test. The results in the charts were obtained getting the clusters for each group starting from same five pre-defined points. The results for Spanish, German and English were according expectations. French was closer to English but French point was almost same distance. Portuguese was closer to Spanish. The similarity between the languages can explain this. The fact that the recipe web site is from Portugal recipes can explain why the Portuguese point was less closer comparing with other languages. The original text written in Portuguese may contain more diverse use of words from vocabulary than translated texts. |

Study case 2: Texts from online newspapers

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | In this case, four articles from different newspaper websites in the worldwide were used, with three articles to feed the learning and one for testing. The chartss were obtained using same clusterization process described before. Again, the results for Spanish, German and English were expected and for French and Portuguese, the points are near of the closest points. |

Study case 3: Technical documents texts

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | In this case, thirty two XML files from SQL management pack were parsed to Text files and divided in two groups. First group with twenty five files was used to feed the learning while other group, with seven files was used to test. The result was the least significant of the three study cases. Only French and German show the expected results. The fact that in technical documentation we have many words not translated like product names can explain this result. Words in English may distort the data points driving to diverged results. |

Conclusion

The method has a good logic that can work to detect the language with a margin of error. The error margin should increase for similar languages, small texts and technical texts and mixing different topics, like sports and food, since the terms used are different.

A bigger sample in the learning database should help to reduce the margin of error. Other options could be to include more information like number of specific characters in a word, to obtain more contrast between the languages included in the study.

To support all Unicode characters also should be a big improvement to increase the sharpness and also to support more languages.

However, the main benefit of this idea is to allow language detection without the need to resort to a dictionary.

Tools used in this work

Three scripts in **R** were responsible for all the data shown in this work.

**Final\_hw.r** is the main script and **Kmeans\_4d.r** and **KMeans\_Helper.r** support the Cluster process in 4 dimensions.

To convert the text files to a CSV in a format with four columns so the R script could consume, it was developed a program called **parse\_txt.exe**, written in **C**. This program was also used to extract the translated strings from Management Pack XMLs to a text file.

References

The cooking recipes used in this work were extracted from the following website:

[**http://www.tasteit.pt/en**](http://www.tasteit.pt/en)

The news articles used in this study were extracted from the following websites:

[**http://www.seattletimes.com**](http://www.seattletimes.com)

[**http://www.lefigaro.fr**](http://www.lefigaro.fr)

[**http://www.faz.net**](http://www.faz.net)

[**http://www.uol.com.br**](http://www.uol.com.br)

[**http://www.clarin.com**](http://www.clarin.com)

The Management Pack files used in this study were extracted from Microsoft SQL Server 2014.

**Character / Decimal code table**

> for (i in 64:90)

+ {

+ if (i==64) print("Char Dec")

+ print(paste0(rawToChar(as.raw(i))," ",as.character(i)))

+ }

[1] "Char Dec"

[1] "@ 64"

[1] "A 65"

[1] "B 66"

[1] "C 67"

[1] "D 68"

[1] "E 69"

[1] "F 70"

[1] "G 71"

[1] "H 72"

[1] "I 73"

[1] "J 74"

[1] "K 75"

[1] "L 76"

[1] "M 77"

[1] "N 78"

[1] "O 79"

[1] "P 80"

[1] "Q 81"

[1] "R 82"

[1] "S 83"

[1] "T 84"

[1] "U 85"

[1] "V 86"

[1] "W 87"

[1] "X 88"

[1] "Y 89"

[1] "Z 90"

Index

Introduction …………………………………………………………………………………………… page 2

Solution approach ……………………………………………………………………………… page 3

Scope and limits ………………………………………………………………………………… page 3

Differences between languages ……………………………………………… page 4

First Character X Last Character distribution …… page 4

Last Character distribution …………………………………………………… page 5

Number of occurrences distribution ………………………………… page 6

Word size distribution ………………………………………………………………… page 7

Study case 1: Cook recipes texts ……………………………………… page 8

Study case 2: Texts from web newspapers …………………… page 9

Study case 3: Technical documents texts …………………… page 10

Conclusion ………………………………………………………………………………………………… page 11

Tools used in this work ……………………………………………………………… page 12

References ………………………………………………………………………………………………… page 12

Character / Decimal code table …………………………………………… page 13

Index ……………………………………………………………………………………………………………… page 14