Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

On Continent and Script-Wise Divisions-Based Statistical Measures for Stop-Words Lists of International Languages

Jatinderkumar R. Saini\textsuperscript{a,c,*} and Rajnish M. Rakholia\textsuperscript{b,c}

\textsuperscript{a}Narmada College of Computer Application, Bharuch 392 011, Gujarat, India
\textsuperscript{b}S. S. Agrawal Institute of Computer Science, Navsari 396 445, Gujarat, India
\textsuperscript{c}R K University, Rajkot, Gujarat, India

Abstract

The data for the current research work was collected for 42 different International languages encompassing 3 continents viz. Asia, Europe and South America. The data comprised of unigram model representation of lexicons in the stop-words lists. 13 scripting systems comprising Arabic, Armenian, Bengali, Chinese, Cyrillic, Devanagari, Greek, Gurmukhi, Hanja & Hangul, Kana, Kanji, Marathi, Roman (Latin) and Thai were considered. Based on a comprehensive analysis of statistical measures for Stop-words lists, it has been concluded that Asian languages are mostly self-scripted and that the average number of stop-words in Asian languages is more than those in European languages. In addition to various important and other first research results, a very important inference from the current research work is that the average number of stop-words for any given language could be predicted to be 200.

© 2016 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of organizing committee of the Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016).

Keywords: Function-Words; Language; Natural Language Processing (NLP); Script; Stop-Words.

1. Introduction

Owing to the increased availability of computing power and awareness as well as the need of processing naturally spoken languages by people, the domain of Natural Language Processing (NLP) has come up with a wide scope of research. Often, the research areas of NLP overlap with the other areas of computing including Text Mining (TM) and Artificial Intelligence (AI). These areas of NLP, TM and AI, often even with other areas like Computational Linguistics (CL) and Data Mining (DM), collectively encompass various tasks and sub-tasks including Concept Mining, Information Extraction, Information Retrieval, Stemming, Lemmatization, Search Engine Indexing, Text Summarization, Part-of-speech (POS) Tagging, Wordnet development, Text Analysis and Text Classification, to name a few.

Almost all of the tasks and sub-tasks of NLP and its intra and inter-related areas require that during pre-processing phase or as required, stop-words be removed from the text corpus before further processing. In computing, according to Wikipedia\textsuperscript{31}, they are the words which usually refer to the most common words in a language. Neither is there a universal list of stop-words, common for all languages, nor is there any standardized list of stop-words for each

\textsuperscript{*}Corresponding author. Tel.: +91-9687689708.
E-mail address: saini--expert@yahoo.com

1877-0509 © 2016 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.076
language. Neither do all NLP tools use such lists nor do all tools prefer not to use them. There have been even instances when even more than one type of list of stop-words has been used by researchers.

It is notable that Function-words are not synonymous with Stop-words though Function-words may include articles (‘the’, ‘a’, etc.), pronouns (‘he’, ‘him’, ‘she’, ‘her’, etc.), particles (‘however’, ‘then’, ‘if’, ‘thus’, etc.), etc. Hence, Function-words act as a subset of Stop-words. In the context of current research work it is also important to note that the paper uses the words ‘language’ and ‘script’ to mean ‘the human spoken natural language’ and its ‘writing system consisting of alphabets’, respectively. The alphabet set for a script may in turn consist of all or some or combinations of consonants, vowels, etc.

The author believes that the language does not become rich if it has a long list of stop-words. Rather, such a language will turn out to be a non-good choice from the perspective of NLP algorithms. It is so because a sentence, a query, etc. written in such a language will have so many words removed from it under the title of stop-words, that probably after the removal of such words the actual intended meaning of the sentence is not conveyed properly. Such a situation becomes a big challenge for NLP, AI and TM tasks including but not limited to query-processing, semantic web-development, text-analysis, text-retrieval, text-summarization, document classification and machine translation.

It is however important to be kept in mind that the language processing is an area in which development of text-processing algorithms or being able to comment on the text is relative and subject to the language and type of text including the domain of consideration.

2. Related Works

There are good numbers of research instances wherein the authors have worked on different international languages and language families as well as Indian languages in context of NLP. But there are fewer instances of research works where the authors have provided statistical analysis of more than two stop-words lists. Sadegahi and Vegas have shown coincidence between Persian and English stop-words. They have defined light stop-word as a unigram stop-word containing few letters. Hence they have provided a comparison of at-least two languages. Alajmi et al. present a statistical approach for extraction of Arabic stop-words list. They have used an approach consisting of calculation of mean probability, entropy and variance probability as well. They have used statistical analysis approach but for extraction of Arabic stop-word list and not comparative analysis of Arabic stop-word list with more than two other language stop-words lists. Similar to the research work for Arabic language, Zou et al. proposed an automatic aggregated methodology based on statistical and information models for extraction of stop-words for Chinese language. Their results showed that their list is quite general and approach is promising enough for implementation purpose. NLP works in languages other than Arabic and Chinese have also been done. For instance, Rakhelia and Saini have worked on Gujarati language while Saini and Desai have worked on Hindi language. There are also instances of research works involving multi-lingual documents.

There are also research instances wherein the authors have used stop-words lists for application purpose. Such an application to a specific domain like Twitter data is presented by Choy. He has used Term-Frequency (TF), Inverse Document Frequency (IDF) and Term Adjacency (TA) for developing a stop words list for the Twitter data source. He has proposed a new technique using combinatorial values as an alternative measure to effectively list out stop words for Twitter data. Another work on Twitter data is presented by Saif et al. They have investigated the effect of removal of stop-words on the effectiveness of Twitter sentiment classification methods. They have found that stop-words do have a great effect on the studied classification methods. Kaur and Saini have presented a Part-of-speech (POS) word class based categorization of Gurmukhi script Punjabi language stop-words.

Arun et al. have explored the use of Latent Dirichlet Allocation (LDA) on stop-words and showed that it usage is employable for suitably handling the authorship attributions. They have also proved that this approach using stop-words is also effective in correct identification of author genders using stylometric studies for textual contents.

The related literature contains many instances of NLP algorithm applications wherein the application of stop-words has not been done or has been done partially too. Identification and analysis of most frequently occurring significant proper nouns in 419 Nigerian scams, textual analysis of digits used for designing Yahoo-group identifiers and structural analysis of username segment in email addresses of MCA institutes of Gujarat state are examples of this.
The current work is based on available data but this data could not be considered scanty as owing to general intelligence concept, the language used most widely only will have the maximum probability of availability of data. This means that to say that the language for which much data is not available is the one whose usage is not much prominent in the world.

3. Methodology

In order to present a comparative analysis of stop-words lists of different international languages based on statistical interpretations, the first step was to collect data of stop-words. The data for this purpose was sourced from Doyle4 as well as research work of Kaur and Saini8. All the stop-words considered for current research work are based on the unigram model of representation of lexicons. A total of 42 different international languages were used for the research purpose. Out of these 42 languages, 21 languages were from Europe, 20 languages were from Asia and 1 language was from South America. The 21 languages from Europe were Basque, Bulgarian, Catalan, Czech, Danish, Dutch, English, Finnish, French, Galician, German, Greek, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Slovak, Spanish and Swedish. Similarly, the 20 languages from Asia were Arabic, Armenian, Bengali, Chinese, Hindi, Indonesian, Irish, Japanese, Korean, Kurdish, Latvian, Lithuanian, Marathi, Persian, Punjabi, Russian, Thai, Turkish, Ukrainian and Urdu. The only South American language under consideration was Brazilian.

It is noteworthy to mention that these languages were not chosen randomly. The availability of a list of stop-words for a language was the most important criterion for selection of a language for the current research work. A sample snap-shot of different languages chosen for current research work is presented in Fig. 1. Due to lack of space, only a sample set consisting of ‘first’ 25 stop-words of ‘first’ 19 languages only is presented here. The significance of word ‘first’ does not need to be emphasized as it is used here to mean that both the list of stop-words as well as the list of languages is sorted and only respective first ‘n’ values are presented here.

It is also noteworthy to mention that this paper has used the term Indonesian language to refer to the Indonesian language as well as its other commonly used term called Malay language. Similarly, this paper deploys use of Kurdish language for Sorani language, Persian language for Farsi language and Irish language for Gaeilge language, respectively.

In addition to dispersion of these languages among 3 continents, the languages could also be classified based on the script used to code that language in written format. Accordingly, there were 25 languages for which English script is required whereas there were 17 languages which used non-English scripting system. It is worth mentioning here that...
from linguistics perspective there is nothing like English or non-English script. English language uses Roman script and it is often called the Latin script. This paper, however, to more clearly highlight the usage of English language, uses the term Roman Script, English Script and Latin Script interchangeably. The only South American language under study used English scripting system. Also, most of European languages use English for scripting hence most of the non-English scripted languages turned out to be Asian languages. The Arabic script is the basis of various languages including Arabic, Kurdish, Persian and Urdu. It is noteworthy that Urdu is scripted using Persian alphabets but they in turn are derived from Arabic itself. Arabic hence is a root to scripts of many languages. Cyrillic is the script for Bulgarian, Russian and Ukranian languages. Armenian, Bengali, Chinese, Greek and Thai similarly provide scripting alphabets to languages named Armenian, Bengali, Chinese, Greek and Thai respectively. Remarkably, these languages and their scripting systems have same names. It is also noteworthy to found that Bulgarian though being a European language has scripting system common with Russian and Ukranian languages, which are Asian languages. Hence, language does not recognize political boundary of land-pieces of earth. Also, such common roots to basic scripts also occur owing to migrations, common civilizations and neighboring countries, though in different continents.

Moving further, Hindi and Marathi have common script called Devanagari. Japanese language uses Kani and Kana scripting system. Kana script further is composed of Hiragana and Katakana styles. It is noteworthy that writing experts consider the mentioned two kana styles as two different scripts, and not merely as two different styles. Korean, which is believed to be an isolate language, uses Hanja and Hangul script. Korean, in fact, is rarely found to use the Hanja script today. At most, it is seen only in a few scattered around Korean texts to highlight a particular Chinese-derived morpheme. Punjabi language uses two scripting styles, viz. Shahmukhi and Gurmukhi. The Shahmukhi style is influenced by Arabic script. This paper used Gurmukhi style of scripting Punjabi language for stop-words list.

4. Results and findings

Table 1 presents a listing of 42 different international languages along with their numbers of Stop-words. Figure 2 presents the details of Table 1 with more clarity in graphical format. Based on the analysis of these 42 languages, a total of 8387 stop-words were found. The maximum number of stop-words was found to be 747, present in Finnish language. Finnish is a European and English scripted language. The minimum number of stop-words was found to be 28 in Ukranian language. Ukranian is an Asian and Cyrillic scripted language.

The average number of stop words among 42 languages was found to be 199.69 (~200). These are the overall maximum, minimum and average values under study. These values were also found based on continent-wise and script-wise divisions, generating different sets of combinations. These statistical values corresponding to these combinations are presented in Table 2.

Brazilian, which is a South American and English scripted language, being the only language outside Asia and Europe was kept out for statistical analysis. Brazilian language was treated as an outlier for the current research work to prevent over-analysis and mis-interpretations based on single language represented by its domain through only one language from one continent.

Based on the statistical data presented in Table 2, it has been found that for Asia, the maximum and minimum numbers of stop-words for English scripted languages have been found to be 507 and 109, respectively for Lithuanian and Irish languages. The maximum and minimum numbers of stop-words for non-English scripted languages for Asia has been found to be 677 and 28 respectively for Korean and Ukranian languages. Similarly, for Europe, the maximum and minimum numbers of stop-words for English scripted languages have been found to be 747 and 35, respectively for Finnish and Hungarian languages. The maximum and minimum numbers of stop-words for non-English scripted languages for Europe has been found to be 259 and 79 respectively for Bulgarian and Greek languages.

Further, from Table 2, it can also be seen that the average number of stop-words in English scripted languages is more than the average number of stop-words in non-English scripted languages. This statement is true irrespective of continental divisions, i.e., 250 > 216 and 180 > 169: both hold true. Also, it is also observed that the average number of stop-words in Asian languages is more than the average number of stop-words in European languages. This statement also holds true for English as well as non-English scripted languages of both continents, i.e. (250 and 216) > (180 and 169). It is notable that this statement is true even though the overall highest number of stop-words were found in Finnish, a European language. The comparison presented in current paper should not be purely seen as
Table 1. International Languages and their Numbers of Stop-words.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Language</th>
<th>Stop-words</th>
<th>Sr. No.</th>
<th>Language</th>
<th>Stop-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arabic</td>
<td>102</td>
<td>22</td>
<td>Italian</td>
<td>134</td>
</tr>
<tr>
<td>2</td>
<td>Armenian</td>
<td>43</td>
<td>23</td>
<td>Japanese</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>Basque</td>
<td>98</td>
<td>24</td>
<td>Korean</td>
<td>677</td>
</tr>
<tr>
<td>4</td>
<td>Bengali</td>
<td>363</td>
<td>25</td>
<td>Kurdish</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>Brazilian</td>
<td>128</td>
<td>26</td>
<td>Latvian</td>
<td>163</td>
</tr>
<tr>
<td>6</td>
<td>Bulgarian</td>
<td>259</td>
<td>27</td>
<td>Lithuanian</td>
<td>507</td>
</tr>
<tr>
<td>7</td>
<td>Catalan</td>
<td>126</td>
<td>28</td>
<td>Marathi</td>
<td>99</td>
</tr>
<tr>
<td>8</td>
<td>Chinese</td>
<td>125</td>
<td>29</td>
<td>Norwegian</td>
<td>119</td>
</tr>
<tr>
<td>9</td>
<td>Czech</td>
<td>138</td>
<td>30</td>
<td>Persian</td>
<td>332</td>
</tr>
<tr>
<td>10</td>
<td>Danish</td>
<td>101</td>
<td>31</td>
<td>Polish</td>
<td>138</td>
</tr>
<tr>
<td>11</td>
<td>Dutch</td>
<td>48</td>
<td>32</td>
<td>Portuguese</td>
<td>147</td>
</tr>
<tr>
<td>12</td>
<td>English</td>
<td>174</td>
<td>33</td>
<td>Punjabi</td>
<td>184</td>
</tr>
<tr>
<td>13</td>
<td>Finnish</td>
<td>747</td>
<td>34</td>
<td>Romanian</td>
<td>281</td>
</tr>
<tr>
<td>14</td>
<td>French</td>
<td>116</td>
<td>35</td>
<td>Russian</td>
<td>421</td>
</tr>
<tr>
<td>15</td>
<td>Galician</td>
<td>160</td>
<td>36</td>
<td>Slovak</td>
<td>173</td>
</tr>
<tr>
<td>16</td>
<td>German</td>
<td>129</td>
<td>37</td>
<td>Spanish</td>
<td>178</td>
</tr>
<tr>
<td>17</td>
<td>Greek</td>
<td>79</td>
<td>38</td>
<td>Swedish</td>
<td>386</td>
</tr>
<tr>
<td>18</td>
<td>Hindi</td>
<td>97</td>
<td>39</td>
<td>Thai</td>
<td>115</td>
</tr>
<tr>
<td>19</td>
<td>Hungarian</td>
<td>35</td>
<td>40</td>
<td>Turkish</td>
<td>114</td>
</tr>
<tr>
<td>20</td>
<td>Indonesian</td>
<td>357</td>
<td>41</td>
<td>Ukranian</td>
<td>28</td>
</tr>
<tr>
<td>21</td>
<td>Irish</td>
<td>109</td>
<td>42</td>
<td>Urdu</td>
<td>550</td>
</tr>
</tbody>
</table>

Fig. 2. Graphical Representation of Languages and their Stop words.

highlighting on the English scripted and non-English scripted languages. The comparison is also noteworthy in terms of the dimension called continent or the geographical area wherein the language is used. It is notable that the two English-scripted languages with the largest number of stop words are Finnish and Lithuanian respectively for Europe and Asia. It is also worth mentioning here that these two languages have respectively rich agglutinative and inflectional morphology.

Rarely, in the related literature is the relationship between the type of script a language uses is related to the number of stop words a language has. The current paper was an attempt to present the same, that is the interplay of the three viz., stop-words, language and script. Obviously, if “can” and “could” are both in one’s stop words list for English,
Table 2. Continent and Script-wise Division-based Statistical Measures for Stop-words.

<table>
<thead>
<tr>
<th>Statistical Measure for Stop-words</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asian</td>
</tr>
<tr>
<td></td>
<td>English Scripted</td>
</tr>
<tr>
<td>Total</td>
<td>1250</td>
</tr>
<tr>
<td>Maximum</td>
<td>507</td>
</tr>
<tr>
<td>Minimum</td>
<td>109</td>
</tr>
<tr>
<td>Average</td>
<td>250.00 (=250)</td>
</tr>
</tbody>
</table>

then one has two words that are from one modal verb. If there is a richly inflected language under consideration, one might have 20 forms or more for the same modal verb. This leads us to the discussion of inflectional complexity of a language in perspective of its scripted stop words, which will be in the future scope of research work.

5. Conclusions

Based on the analysis of statistical measures for stop-words of lists of 42 different international languages, it has been concluded that there is a vast difference between the numbers of stop-words in a list of such words for a given international language. It is concluded that the languages may belong to different continents but still may have the common scripting system. The most important conclusion of this research work is that the average number of stop-words in any given language will be 200. The English scripted languages with maximum and minimum numbers of stop-words for Asia were Lithuanian and Irish languages respectively while their counterparts for Europe were Finnish and Hungarian languages respectively. Similarly, the non-English scripted languages with maximum and minimum numbers of stop-words for Asia were Korean and Ukrainian languages respectively while their counterparts for Europe were Bulgarian and Greek languages respectively. It is also concluded that the average number of stop-words in English scripted languages is more than the average number of stop-words in non-English scripted languages, irrespective of continental divisions. It is further concluded that the average number of stop-words in Asian languages is more than the average number of stop-words in European languages. Also, another significant conclusion of the current research work is that most of the European languages are English script based while most of the Asian languages are self-scripted.

The current research work, directly or in-directly, does not claim that a particular language is better or not better than the other. It just presents an un-biased and critical comparison based on the statistical measures derived from the analysis of stop-words lists of different international languages. This paper does not intend to advocate the preeminence or inadequacy of a language, script, country, continent, civilization or any community. The current paper presents first results just for academic research purpose and are best reported on the available data used for research purpose.

The results reported here could be used for predicting about stop-words list of a language for which such a list even does not exist yet. The results could also be used for development of various language dependent tools as well as various NLP algorithms.

References


