Adaptive Distribution of Vocabulary Frequencies: a novel estimation suitable for social media corpus

Rodrigo Augusto Igawa, Guilherme Sakaji Kido, José Luis Seixas Jr., Sylvio Barbon Jr.
Department of Computing
State University of Londrina
Londrina-PR, Brazil
igawa.rodrigo@gmail.com, guilhermekido@gmail.com, jlseixasjr@gmail.com, barbon@uel.br

Abstract—This paper aims to propose a mathematical model that evaluates the distribution of the vocabulary frequency terms in proportion to a probabilistic ideal. Once we are able to evaluate it, it becomes possible to use it in order to examine text denoising. We propose this new metric based on the classic Zipf’s law statistic method. The experimental set to test the classic Zipf’s law and our developed model is based on some books of the classic literature and some tweets sets of Twitter. Thus, our main result is that the model proposed in this work is more sensitive to the presence of text noises than Zipf’s law and is asymptotically quicker, suitable to corpus of social media networks.

Keywords—Information Retrieval, Text preprocessing, Zipf’s Law, Social Media Networks.

I. INTRODUCTION

Substantial part of text produced by online social medias is considered noisy as are provided unreliable terms for textual Information Retrieval. There are many examples such as: jargons, slangs, word contractions, misspelled abbreviations which ignore the rules of spelling or grammar, punctuation errors or omissions, phonetic spelling, misspelling for verbal effect or other intentional misspelling and recognition of out-of-dictionary named entities [1], [2]. These noises are regular features in web chat rooms, online forums or social media which are an important data source for Information Retrieval (IR). Although, online textual content might be noisy, they are very important because millions of opinions expressed about a certain topic are highly unlikely to be biased [1].

Through the last decade, some works [3]–[7] around the relevance of the data preprocessing (DPP) in text mining has been performed. These authors contributed about the importance of DPP in different approaches like: stopwords to text mining in a relational database, the matter of stopwords removal and the impact of combining preprocessing methods for text searches. In this paper, our goal is to demonstrate a novel mathematical model, Adaptive Distribution of Vocabulary Frequencies (ADVF), to evaluate the noise level reduction that improve the use of DPP mainly in social media networks. The experiments are based on stopwords filtering and the proposed model is performed using a linear function that create an ideal distribution of frequency based on classical Zipf’s law. By doing so, we are able to mathematically understand how efficient a DPP approach is to evaluate the noise level reduction. The work results are based on a experimentation of ADVF potential in a real world scenario.

This paper is organized as follows. Next section provides a brief literature review on the major areas of IR applications and problems in the past of last years. Section III deals with the explanation of the classic Zipf’s law and it’s application in social media. In Section IV we propose a new approach to evaluate the terms vocabulary distribution. Section V shows how the experiments were performed in this work. Section VI presents our results and discussions about the classic Zipf’s law application and the new ADVF proposed. Finally, Section VII provides the implications and limitations found and also explores the directions for future researches and discuss it.

II. RELATED WORK

IR has become an important tool used to help researchers to find relevant information in big amount of documents [8], [9], they related problems about finding the best document since there are many ways to write exactly the same thing. In business, [10] has proposed a method of quick searching in web pages in order to help enterprise information to be extracted as quickly as needed. In other languages, processing narration has also been useful to identify entities from prophetic narration texts [11]. At last, in web environments like forums, IR has been represented by processing texts produced in learning environment where the goal is to evaluate users’ content [12], [13].

Noise, the focus of the presented work, was reported by [1], [9]. In [9], an entity could be represented differently by using another abbreviation or word, this fact could result in problems of misunderstanding that dictionary-based, rule-based and machine learning approaches were not enough to deal completely. Considering social media like Twitter, we have even more problems binding noisy terms to knowledge discovery like use of jargons, word contractions, slangs, acronyms, emoticons and phrases that ignores the classic grammar rules [1].

In many cases, the solution is to choose a better set of DPP operations. The most commons are: tokenization, stopwords removal, stemming and lowercase conversion. A good combination is very important for text classification and can enhance the final results of the preprocessing [4]. In corpus built from social media, the noise problem is frequent and a measure of noisy terms, as well as the DPP quality, are not problems completely solved. Some other alternatives forms of DPP are presented by [14] and [15] that propose mathematical models as Wavelets and Fourier Domain Transform to deal with denoising. One approach from Neural Networks is shown in [16] which, despite denoising is not its main
goal, has achieved some noise reduction with the use of Backpropagation Neural Networks. A last alternative provided in the literature is a manual form of denoising, where the staff members from a web-based learning environment were responsible for preprocessing the data to be analysed by [17], [18].

In social media scenario, denoising has been dealt with some different approaches. For example, in order to perform a sentiment analysis from Twitter with real-time results as an ultimate objective, denoising was denoted by just picking the right subset of tweets given a specific subject and, therefore, more than half of the initially collected set of tweets were classified as noise [19]. As mining social media for event relatedness evaluation, noise had another perspective. This time, the noise was denoted as tweets that were not related to any event and were discarded by the density-based clustering algorithm which was the solution to the problem [20]. No approach was concerned about the asymptotic complexity cost. In this work, we want to introduce a novel estimation suitable for social media on denoising evaluation along with its asymptotic cost. We deal with noise as the terms not needed in a Space Vector (as explained in Section V) to perform an Information Retrieval.

III. ZIPF’S LAW ESTIMATION

In this section, we present the classic application of Zipf’s law in a corpus. In the first moment we describe its use on Alice in Wonderlands, a corpus that provides formal English language use, as it shown in Section III-A. Later, in Section III-B, we discuss its applicability over a social media corpus with very casual Portuguese obtained from tweets about last Christmas in Brazil.

A. An Overview on Zipf’s law

Zipf’s law is a classic measure of literature that studies the distribution of frequency terms in a corpus. It was developed by George Kingsley Zipf which some consider to be the father of lexical statistics [21]. There are many ways to use it as explained by [23]–[25]. In order to study term frequency distribution, we must first tokenize the text and count all the instances of distinct terms that occur in the corpus.

Once we have tokenized the entire corpus and counted how many times each token appears, we are able to know the number of the terms which there are in the corpus. This measure is called Corpus size (N) as seen in [21]. In addition, we are also able to know how many unique terms we have in our corpus and we call it vocabulary size (V) as called by [21] or lexicon (L) earlier named by [22].

The next step is to count how many times each term appears. Afterward, a rank is given to each term. The first rank is usually given to the term most frequent in a corpus, therefore, the last rank is given to the term less frequent [21]. Usually, the first positions are corresponding to prepositions, articles and pronouns which in text mining are usually part of a stopwords list [4].

Finally, it is possible to plot a graphic where a relation between ranks and frequencies is visible. When plotted in a log space, the x-axis as the ranked terms and the y-axis as frequencies, a pattern should be found in any text base [21], [23]. This pattern is shown in Figure 1.

However, the model that later would be known as the Zipf’s law is a proposal of frequency prediction of a term given its rank. The model is shown in Equation 1 with \( r(w) \) as the rank of a given word [21], [24].

By applying log on both sides of the Equation 1 with some basic log properties, we get Equation 2 where \( f(w) \) is the frequency from \( w \)-th term, \( C \) and \( \alpha \) are constants that are found by using the standard least squares method. The Zipf’s law in form of Equation 2 is the most used form and it is an attempt to predict how frequent that term should be given its own rank [21], [23]–[25].

\[
f(w) = \frac{C}{r(w)^\alpha} \quad (1)
\]

\[
log(f(w)) = log(C) - \alpha * log(r(w)) \quad (2)
\]

In order to exemplify an statistical application of the law described, we picked up a book from the classic literature: Alice in Wonderlands. By following the steps explained in this Section, the measures obtained from the book are: corpus size with 27,337 terms and vocabulary size with 2,570 terms. In Table I is shown the top 12 terms ranked in the book with their frequencies. Note that, mostly of the top terms ranked are prepositions, pronouns and articles (stopwords). In addition, it is possible to realize that stopwords are far from the Zipf’s law line as shown in Figure 1. Thus, as proposed by [7], in the Space Vector approach used in this work we considered them noise.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>1,643</td>
<td>7</td>
<td>of</td>
<td>545</td>
</tr>
<tr>
<td>2</td>
<td>and</td>
<td>872</td>
<td>8</td>
<td>said</td>
<td>514</td>
</tr>
<tr>
<td>3</td>
<td>to</td>
<td>729</td>
<td>9</td>
<td>you</td>
<td>462</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>632</td>
<td>10</td>
<td>you</td>
<td>411</td>
</tr>
<tr>
<td>5</td>
<td>it</td>
<td>595</td>
<td>11</td>
<td>alice</td>
<td>398</td>
</tr>
<tr>
<td>6</td>
<td>she</td>
<td>553</td>
<td>12</td>
<td>in</td>
<td>369</td>
</tr>
</tbody>
</table>

Fig. 1. The Zipf’s law Log-log plot of Alice in Wonderlands terms

TABLE I. TOP 12 RANKED TERMS FROM THE BOOK ALICE IN WONDERLANDS

283
TABLE II.  **TOP 12 RANKED TERMS FROM SET OF TWEETS DURING CHRISTMAS IN BRAZIL**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>de</td>
<td>4,152</td>
<td>7</td>
<td>a</td>
<td>2,024</td>
</tr>
<tr>
<td>2</td>
<td>natal</td>
<td>3,868</td>
<td>8</td>
<td>que</td>
<td>1,952</td>
</tr>
<tr>
<td>3</td>
<td>e</td>
<td>3,255</td>
<td>9</td>
<td>t</td>
<td>1,639</td>
</tr>
<tr>
<td>4</td>
<td>se</td>
<td>2,865</td>
<td>10</td>
<td>co</td>
<td>1,618</td>
</tr>
<tr>
<td>5</td>
<td>festa</td>
<td>2,336</td>
<td>11</td>
<td>topo</td>
<td>1,525</td>
</tr>
<tr>
<td>6</td>
<td>presente</td>
<td>2,072</td>
<td>12</td>
<td>eu</td>
<td>1,406</td>
</tr>
</tbody>
</table>

TABLE III. **PREPROCESSED TOP 12 RANKED TERMS FROM SET OF TWEETS DURING CHRISTMAS IN BRAZIL**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
<th>Rank</th>
<th>Term</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>natal</td>
<td>3,726</td>
<td>7</td>
<td>vai</td>
<td>742</td>
</tr>
<tr>
<td>2</td>
<td>festa</td>
<td>2,955</td>
<td>8</td>
<td>ano</td>
<td>545</td>
</tr>
<tr>
<td>3</td>
<td>presente</td>
<td>2,091</td>
<td>9</td>
<td>so</td>
<td>518</td>
</tr>
<tr>
<td>4</td>
<td>compra</td>
<td>1,024</td>
<td>10</td>
<td>ja</td>
<td>516</td>
</tr>
<tr>
<td>5</td>
<td>vou</td>
<td>1,000</td>
<td>11</td>
<td>quero</td>
<td>465</td>
</tr>
<tr>
<td>6</td>
<td>nao</td>
<td>938</td>
<td>12</td>
<td>mais</td>
<td>464</td>
</tr>
</tbody>
</table>

**B. Social Media Application**

The example shown in Section III-A was based on Alice in Wonderland which is a classic book. However, the purpose is to work with social media, in this paper, Twitter. Therefore, we are now presenting the exactly law exemplified before, but now, from a set of Tweets from the last Christmas in Brazil. The set is compounded by 8,596 tweets, the corpus size is a collection of 119,059 terms which 17, 451 (vocabulary) are unique.

The tokenization step is done the same way that is done with Alice in wonderlands, the entire set of tweets is tokenized until each term is separated from another, then the count of frequencies starts and the rank to each unique term is attached. The top 12 terms ranked are shown on Table II and a plot in log space of rank per frequency is shown in Figure 2(a).

Later, the experiment is reproduced on the same set of tweets. Though, this time a set of preprocessing methods is executed. Our set of preprocessing methods is composed by: lowercase conversion and stopwords removal. In this set of stopwords we added some more terms despite of pronouns, articles and prepositions. We also found important to remove links, hashtags and user citations as it was seen as a noise by [1].

By doing so, the corpus size was reduced from 119,059 to 108,608 terms. The vocabulary size was reduced from 17,451 to 12,963 unique terms. Therefore a new top 12 ranked terms table is generated and shown in Table III. A new plot in a log space of rank per frequency can be seen in Figure 2(b).

Note that a preprocessed set of tweets is closer to the prediction of Zipf’s law line than a non-preprocessed set. Zipf’s law is very sensitive on the edges that are usually terms that have been used more than it should and terms were so infrequently that they could have been avoided [21].

**IV. PROPOSED APPROACH**

It is possible to use Zipf’s law to analyse the presence of noise in a text produced by a social media. However, we propose an adaptive form of the law in order to evaluate the noise presence with a low computational complexity as described in this Section.

Firstly, the Zipf’s law is about plotting a line in a log space that represents a probabilistic distribution of terms. As described in Section III, the line to be plot depends on 2 constants: C and α. Some works tend to set α as -1 and C as the highest frequency term in corpus, but this is a simplification [21]. To obtain the most accurate experiments as possible, this simplification is not used and the constants are found as proposed traditionally using the least square method.

By observing the description of the Zipf’s law it is possible to realize that the line plotted is a line that best fits all the presented frequencies. However, this line could disguise the noise that is the attribute that this work aims to emphasize.

Thus, we propose another approach based on Zipf’s law: the Adaptive Distribution of Vocabulary Frequencies (ADVF). It is also a straight line, but not a line that best fits all the frequencies. The new line would start from the highest frequency as the simplified version of Zipf’s law, but the last point plotted by the new straight line would match the last element ranked frequency. Once this line do not best fits all the frequencies, noise presence should be highlighted. The ADVF line can be easily found as shown in Equation 3 and Equation 4. In Equation 3, n is the last member of a Space Vector, therefore, f(n) would be the last frequency found for the last ranked term and m is the slope to be found. In Equation 4, given a n’ ranked term, the ADVF(n’) represents its adaptive frequency distribution. Note that equations are always using
the parameter in its logarithmic form once the Zipf’s law usually works in log spaces.

By choosing the the last term frequency as the second point to be used in the ADVF line, we are able to emphasize the terms with lower frequencies. As later shown in Section VI, social media corpus usually provides a greater number of terms with low frequencies, which is a result from the very casual English described by [2] used in social medias.

$$m \cdot (\log(f(1)) - \log(f(n))) = (\log(r(1)) - \log(r(n))) \quad (3)$$

$$ADV(F(n') = m(\log(r(n'))) + (\log(f(1))) \quad (4)$$

To compare the noise existence expressed by both approaches, a set of tweets of several sizes are used. In addition, two books from Shakespeare are also included in the experimental set to observe the results found by approaches in texts from a classic literature.

V. EXPERIMENTAL SETTINGS

Twitter, the social media used in this work, is considered a micro blogging service. Unlike other social media, Twitter is known by the fact that its users post short texts (140 characters at maximum) using the Web or mobile devices [26]. These short texts, named tweets, are available publicly as default, and are immediately broadcasted to the users followers [27].

The Twitter developer team offer a set of streaming that give other developers low latency access to Twitter’s global stream of Tweet data. The tweets sets used as samples in this experiment were collected using this service. The data received from the Twitter API contained many fields, for example, the identification number of the message, the identification number of the user who authored the tweet, the short text field (the tweet itself) and some more meta data fields [27]. To this work, the most important field is the 140 characters of each tweet that are used to analyse terms frequency distribution. All the received data from the Twitter API is persisted with the use of the MySQL database.

The text preprocessing step consists basically in stopwords removal and lowercase conversion. Lexical analysis indicates that the most frequent terms are prepositions, articles and pronouns. Although very important to syntax analysis, they are not significant in discriminating the document contents. Such common terms are referred to as stopwords. Stopwords tend to diminish the impact of frequency difference among less common terms, and result in unproductive processing if left in the text [10]. Thus, for the propose of this work, we considered stopwords as text noise. The stopwords removal along with lowercase conversion and all plots needed were executed with the MathWorks software.

Then, for each corpus listed in the Table IV, the Zipf’s law and the ADVF lines are calculated. The same corpus is used twice to plot both lines. The line is plotted once for the corpus in its raw form, we named it non-preprocessed corpus. In the second time, it is preprocessed and, just after that, calculated. This way, we have the result of both Zipf’s law and ADVF over a text base before and after preprocessed. As a final step of the experiment, a structural equivalence analysis is performed for both Zipf’s law and ADVF with the observed distributions.

In order to find the closeness from observed word frequency distribution to the probabilistic one, the structural equivalence is selected due to represent how substitutable two Space Vectors are. The structural equivalence performed in this work is the first form presented in [28]. It is based on Euclidean Distance between two vectors. The Euclidean Distance is also used in other work of text mining as exemplified by [29]. The formula used in this work is shown in Equation 5, where the $V$ are two frequency vectors and $V_i$ is the observed frequencies vector. $V_j$ might be represented by the ADVF or Zipf’s law line, $m$ is m-th term which has a range from 1 to $n$ (Space Vector size).

$$EDis(V_i, V_j) = \sum_{m=1}^{n} (V_{im} - V_{jm})^{1/2} \quad (5)$$

The analysis and results of this experiment are shown in the Section VI.

VI. RESULTS AND DISCUSSION

The results are presented in Table IV that shows the reduction of each experimental set of tweets along with classical books. As this work uses Space Vectors to represent the existing terms in each corpus, the noise reduction measure corresponds to the vocabulary reduction measure in each experimental set.

The first notable result is shown in Figure 4. It is possible to realize that both experimental sets from classical literature books did not show a great amount of noise reduction, which confirmed the fact that social media provides texts with greater quantity of noisy terms due all the kinds of informal writings [2]. This result is shown in Figure 4 where the terms filtered from formal english books (sets 1, 2 and 3) corpus are quite lesser than in social media corpus.

A second perspective to be discussed are the two premises considered in the experiments for evaluate the proposed method: asymptotic computational complexity and noise sensitivity.

When determining the computational load of a model algorithm, an important approach is to determine the asymptotic complexity. Zipf’s law classical distribution have a polynomial (or algebraic) complexity function, $O(n^3)$. Considering a large amount of data a polynomial complexity can be impeditve to perform the processing based on that model. A linear complexity, $O(n)$, such as ADVF is desired for IR mass of data. The growth of both functions is shown in Figure 3, $T(n)$ is time-complexity function of the algorithms.

The main operation which implies in a polynomial complexity function is the least squares computing, that dominates
TABLE IV. Corpora EXPERIMENTAL SETS

<table>
<thead>
<tr>
<th>Set</th>
<th>Name (Type)</th>
<th>Corpus</th>
<th>Vocabulary</th>
<th>Vocabulary after Reduction in %</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hamlet (Book)</td>
<td>31,487</td>
<td>4,598</td>
<td>4,474</td>
<td>English</td>
</tr>
<tr>
<td>2</td>
<td>Rommel and Juliet (Book)</td>
<td>25,026</td>
<td>3,579</td>
<td>3,459</td>
<td>English</td>
</tr>
<tr>
<td>3</td>
<td>Alice (Book)</td>
<td>24,123</td>
<td>2,837</td>
<td>2,711</td>
<td>English, Spanish and Portuguese</td>
</tr>
<tr>
<td>4</td>
<td>BBB TV Show (Twitter)</td>
<td>122,245</td>
<td>9,031</td>
<td>17.52</td>
<td>Portuguese</td>
</tr>
<tr>
<td>5</td>
<td>Soccer Ria x Jap (Twitter)</td>
<td>102,206</td>
<td>10,939</td>
<td>29.03</td>
<td>Portuguese</td>
</tr>
<tr>
<td>6</td>
<td>Christmas (Twitter)</td>
<td>119,059</td>
<td>12,963</td>
<td>25.72</td>
<td>Portuguese</td>
</tr>
<tr>
<td>7</td>
<td>The Walking Dead (Twitter)</td>
<td>20,209</td>
<td>3532</td>
<td>24.84</td>
<td>English, Spanish and Portuguese</td>
</tr>
<tr>
<td>8</td>
<td>FIFA Conf. Cup’s (Twitter)</td>
<td>158,098</td>
<td>20,172</td>
<td>28.16</td>
<td>Portuguese</td>
</tr>
<tr>
<td>9</td>
<td>Game (Twitter)</td>
<td>31,358</td>
<td>4,289</td>
<td>29.90</td>
<td>English, Spanish and Portuguese</td>
</tr>
<tr>
<td>10</td>
<td>NBA (Twitter)</td>
<td>25,103</td>
<td>20,172</td>
<td>31.27</td>
<td>English, Spanish and Portuguese</td>
</tr>
<tr>
<td>11</td>
<td>Food (Twitter)</td>
<td>52,342</td>
<td>12,963</td>
<td>40.11</td>
<td>English, Spanish and Portuguese</td>
</tr>
<tr>
<td>12</td>
<td>Hugh Laurie Reel (Twitter)</td>
<td>220,457</td>
<td>22,670</td>
<td>42.76</td>
<td>English, Spanish and Portuguese</td>
</tr>
</tbody>
</table>

Fig. 3. Asymptotic complexity of Zipf’s law and ADVF

Fig. 4. Sensitivity of propose model evaluated in experimental sets

Fig. 5. The Zipf’s law along with ADVF plots from Christmas in Brazil

asymptotically the other computations. In ADVF the linear distribution is based on the frequency of first term, avoiding extra computing keeping a low complexity.

Furthermore, the sensitivity of the ADVF is another advantage regarding noise evaluation. All the experiments demonstrates the superior sensitivity of the ADVF compared to Zipf’s law approach. The sensitivity of both methods is proportional to the corpus size, as visible in Figure 4, but the ADVF obtained an average of 35.89% more sensitivity considering all experiments, books and microblog. Treating singly books and microblog, ADVF model reaches sensitivities of 44.56% and 33.00%.

Based on the experiments it is possible to assert that the noisy terms amount identified by stopwords filtering regards the type of corpus. The classical books get a reduction of 3.74% in terms, like sets 1, 2 and 3 in Figure 4, in other words, just this percentage was considered noise by stopwords filtering. However, the microblogs corpus reach 28.99% of reduction, expliciting the amount of noisy terms that must be treated in this kind of IR action. In Figure 4 is visible that terms filtered by stopwords approach from microblog corpus in experimental sets 4, 5, 6, 7, 8, 9, 10, 11 and 12. Note that, the noise is more present even small social media corpus like sets 7, 9 and 10 which provide an amount of total words similar to the books. The growth of sensitivity, mainly the difference between Zipf’s law and ADVF values, can be explicited in a noisy corpus, a common feature in social medias.

VII. CONCLUSION

As presented in Section II, there are many forms to identify noise over social media corpus. However a measure to evaluate
how much denoising was performed in a DPP step would be interesting to point the effectiveness of such process.

As presented in Section IV, Zipf’s law uses a straight line in log space that best fits all the distributed frequencies which could disguise the real amount of noise. Thus, this works proposed a new approach: the ADVF that by not plotting a line that best fits all the frequencies would not disguise the noise amount.

In Section VI, as expected the ADVF has shown a bigger sensitivity than Zipf’s law in relation to the presence of noise. The observed distribution is less equivalent to ADVF than it is in relation to Zipf’s law straight line. As result, the ADVF has shown more sensitivity from social media than Zipf’s law and, thus, is a better measure of DPP results.

In addition, Section VI also provides another evaluation of the ADVF perspective: the asymptotic complexity. This is very important in this work since a small set of text from social media corpus represent a huge amount of terms. As shown, ADVF has a linear complexity which is more desirable than the polynomial complexity from the Zipf’s law.

However, a gap that ADVF presents is the necessity of being used in social media as it is highly sensitive to the noise presence and corpus from classical books do not exhibit as much noise as, the studied media, Twitter. A future task would be to study the ADVF behavior in order to find the best size of corpus to indexing and weighing terms.

VIII. ACKNOWLEDGMENT

We are grateful to CNPQ that made this paper possible by sponsoring our work, process 479821/2013-5.

REFERENCES


