Influence of Unique Words on the Performance of Corpus-Based Keyword Detection Methods

O. S. Kushnir, V. V. Yaremkiv, I. Y. Dovhan, and A. I. Kashuba
Department of Optoelectronics and Information Technologies
Ivan Franko National University of Lviv
107 Tarnavsky Street, 79017 Lviv, Ukraine
o.s.kushnir@lnu.edu.ua, volodymyr.yaremkiv@lnu.edu.ua

Abstract—**We study the performance of corpus-based keyword detection methods, including TF-IDF, in a particular case when a text under investigation contains unique words, which are absent or rare in the other texts of corpus. The two points are subjects of our main attention, the quality of keyword list and propriety of the corresponding keyness scores, as well as criticality of the methods to small perturbations of the corpus. We conclude that a number of heuristically introduced TF-IDF-like measures compete quite successfully with TF-IDF in their performance but, on the other hand, they cannot cope with the problem of criticality of their scores inherent to the unique words**

Index Terms—**Keywords; corpus-based keyword detection methods; TF-IDF; unique words; criticality**

I. Introduction

Since keywords summarize in a concise manner the main semantics and contents of texts, automated extraction of these words represents a useful tool for the fields of indexing and categorization of textual documents and, more generally, in textual data mining and information retrieval. Roughly, keyword detection methods can be divided into 'domain-dependent' and 'domain-independent' groups, according to whether they involve a reference textual database (a collection of texts, or corpus) or not. The term 'domain-dependent' implies that a corpus can be referred to some domain or topic, so that the keywords extracted from a text under analysis reflect the meanings that distinguish a given text against the background of the domain described by a corpus. For instance, a word 'physical' is hardly a keyword in the case of corpus associated with pure physics, although it can quite happen that it is so with respect to a more general collection of texts, e.g. on natural or social sciences. In spite of this inconvenience as well as evident drawbacks linked to relatively low operation speed, the corpus-based keyword detection methods, e.g. a well-known TF-IDF (Text Frequency – Inverse Document Frequency) approach [1, 2], play a central role in modern web search engines. Moreover, in most cases they outperform standard domain-independent detection techniques that rely upon a single text under test and engage no corpora (see [3–7]). As a consequence, the results derived with the corpus-based methods can be used as an authoritative reference or benchmark, when comparing various (higher-speed) domain-independent methods and judging which of them is better. In this respect, the data of corpus-based methods can be regarded as a useful alternative to commonly used human-made keyword lists which, of course, might be subjective.

Despite a large amount of empirical and theoretical work on the domain-dependent methods for detecting keywords, we believe that the subject is still not concluded. In particular, this concerns a point of our present attention, so-called 'unique words'. We define them as the words present in a given text (to be compared with a corpus) but absent in all the texts of the corpus. Put another way, the number n_t of texts from the corpus where such a 'truly unique' word t occurs is equal to $n_t = 0$. We have found that the unique words represent a rather general phenomenon, being typical for a large majority of texts. Neologisms, words invented on purpose, uncommon and rarely used scientific or technical terms, and even typos are ready examples. It is also useful to expand the discussion to the case of so-called 'quasi-unique' words that occur very rarely in a corpus ($n_t << n$ though $n_t > 0$, with n being the overall number of texts in a corpus). Like the 'truly unique' words, the ranking of 'quasi-unique' words yielded by the domain-dependent methods can suffer from criticality. For a convenience, our term 'unique word' embraces the both classes of 'truly unique' and 'quasi-unique' words. Although these criticality problems are intuitively well understood by a wide information-retrieval community, the appropriate analysis has been chiefly reduced to rather schematic or purely qualitative arguments. To the best of our knowledge, the problem has still not been addressed in a direct quantitative manner. In the present work we study and compare the performance characteristics for a number of corpus-based keyword detection methods under the condition when the unique words are available in the text.

II. Materials and Methods

A. Corpus and Texts under Analysis

We have prepared a corpus of literary works taken from the free text collection "Project Gutenberg" [8]. There is $n = 4829$ texts in our corpus and its size amounts to 1.88 GB in UTF-8 coding. The total length L of all the texts in the units of word tokens is approximately equal to $L = 3.81\times10^8$, while the total vocabulary V in the units of word types is $V = 1.23\times10^6$. Then the average text length l_m in this corpus is nearly $l_m = 0.79\times10^5$. The main text we have analyzed is J. R. R. Tolkien's novel

"The Lord of the Rings" (abbreviated hereafter as LOTR), which has the length $l = 4.79\times10^5$ and the vocabulary $v = 1.43\times10^4$. The size of the file is equal to 2.31 MB. Another text under analysis is H. Harrison's novel "West of Eden" (abbreviated as WOE) with the parameters $l = 1.67\times10^5$ and $v = 8.00\times10^3$ (the file size 0.92 MB).

One can safely restrict oneself to considering, as keyword candidates, only those words of which absolute frequency F is higher than some threshold F_{th}. In case of our relatively long literary texts, we have loosely $F_{th} = 10$, rather loosely. Then the remaining 'above-threshold' vocabularies that consist of the word types with $F \geq F_{th}$ are equal to $v_{th} = 3.08\times10^3$ for LOTR and 1.49×10^3 for WOE. Supplementary filters can be built basing on the thresholds in relative-frequency percentage. The reasons of our choice of the texts under analysis are as follows. First, they contain a lot of unique words, some of which, though not all, represent important keywords. Second, in case of literary works, the list of keywords is rather trivial; quite naturally, it is a list of main characters and place-names. Here there is no room for controversy, as might be the case if, e.g., scientific texts are scrutinized (cf. discussions [3–5]). Therefore, the first point enables studying the influence of unique words on different keyword detection methods, whereas the second one facilitates a high-quality 'manual' check of their performance by anyone who knows these literary works well.

B. Keyword Detection Methods

In the present work, we analyze the following keyword detection methods.

1. Since a keyword is often defined as a term of which frequency in a given text is notably higher from the frequency commonly typical for a corpus, the simplest measure of a 'keyness' is given by

$$R = f/f_m, \tag{1}$$

where f denotes the relative frequency of a word t in the text under study ($f = F/l$ in terms of the absolute frequency F and the text length l) and f_m is the mean frequency of this word in the corpus.

2. A canonical heuristic TF-IDF metrics is defined via the relation (see [1, 2])

$$T = f \log(n/n_t), \tag{2}$$

where n and n_t are respectively the total number of texts in the corpus and the number of texts where the word t occurs. Further on we use a natural logarithm in (2), which is implemented via a standard Python function `math.log()`.

3. Trying to eliminate evident drawback of the measure (1), which is associated with a lack of statistical importance evaluation, one can improve formula (1) as

$$Z = (f - f_m)/\Delta f, \tag{3}$$

where Δf implies the mean-square deviation of f from its corpus-averaged value f_m. Formula (3) is nothing but a well-known Z-score which is commonly used when testing different statistical hypotheses [2].

In an attempt to find out and compare critical features of the methods (1)–(3) and, possibly, improve their performance, we have introduced, purely heuristically, a number of combined TF-IDF-style measures:

$$\text{TF-Z} = f\,(f - f_m)/\Delta f, \tag{4}$$

$$\text{Z-IDF} = [(f - f_m)/\Delta f]\,\log(n/n_t), \tag{5}$$

$$\text{TF-IDF-Z} = f\,[(f - f_m)/\Delta f]\,\log(n/n_t). \tag{6}$$

A particular point of our interest in (4) and (6) is whether a good 'keyness' measure of a term is simply proportional to its frequency f as postulated in [9], or more complex relations, with $\text{TF} \sim f^\beta$ ($\beta \neq 1$), can be used.

The mean frequency and standard deviation entering formulae (1) and (3)–(6) can be calculated with weighting of text length or with no weighting. Therefore, each of these methods has its unweighted and weighted version (R and R_w, Z and Z_w, and so on). The only exception, TF-IDF method, does not depend on weighting.

C. Description of Program

Our program for detecting keywords with the corpus-based methods consists of two subprograms. The first reads a corpus of texts and performs standard preprocessing, including transforming of letters to lower case and removal of non-letter characters. No stemming, parsing and dropping of stopwords has been used. Although these approaches would have improved much the performance of keyword detection methods, here we intend to investigate the resources of 'neat', i.e. unaided, methods. The following statistical information is indexed: (1) the lengths l_i and the vocabularies v_i of the texts involved in the corpus, (2) all the word types that happen at least once in at least one text, (3) the total absolute frequency ΣF of every word type in the corpus, (4) the relative frequencies f_m and $f_{m,w}$ and the standard deviations Δf and Δf_w for every word type in every text, and (5) the number n_t of text documents in which every word type occurs. The last stage of work of the first subprogram is saving the results to a disk as two files. The first one, `corpus_info`, summarizes the main characteristics of the corpus, while the second index file, `corpus_dict`, contains the basic statistical metrics of all the word types. The latter file has the size 175 MB in case of our 1.88 GB corpus. The computation time for the case of our corpus is about 2.0 h with Intel Core i3-M370, 2.40 GHz 8GB RAM, and Python 3.6.4. Note that this calculation needs to be done only once the corpus is built, whereas further index updates after new texts are added to the corpus can be performed fast in an incremental manner.

The main subprogram reads the updated index file and the text under study, and calculates the keyness parameters for all the word types occurring in the text according to the methods described above. The main subroutine is accomplished in ~ 1–10 s. After calculations, the data is stored on a disk in the form of a ranked list of words sorted according to descending keyness scores. Here the absolute-frequency threshold F_{th} represents a used-defined parameter. We use another relative-frequency filter: the keywords fall into the list whenever their relative frequencies f are higher than the thresholds f_{th} or $(f_w)_{th}$. To make possible comparisons of the keywords obtained for different texts, we normalize R, T, Z and the other statistical parameters, so that the sum of the corresponding parameters R_n, T_n, Z_n and others are equal to unity for all the words filtered using the two thresholds.

A special subroutine has been written to find the statistics of unique words in the corpus. Finally, in order to compare the resulting data obtained with different methods under different conditions, we have used a vector space model [2] and a standard cosine similarity of the vectors whose dimensionality is equal to the number of keywords remained for the analysis (e.g., $N = 100$ or 200), and the component values are equal to normalized keyness scores.

III. Results and Discussion

A. Some Statistical Characteristics of Corpus, and Unique Words

The top ten word types with the highest total frequencies ΣF in our corpus are listed in table I. Note that the differences between the weighted and unweighted relative frequencies and standard deviations are typically less than 5% and 20% for the top-frequency words, though they can be much larger for the medium- and low-frequency words. In particular, different rankings of the words *to* and *of*, as well as *he*, *in* and *i*, as obtained according to ΣF (or $f_{m\,w}$) and f_m parameters, are noteworthy. As a consequence, the weighted and unweighted measures (1) and (3) can differ significantly for some words. Statistical characteristics of the most important keywords found for the texts LOTR and WOE are also listed in table I.

Irrespective of the main problem under discussion, i.e. keyword search, it would be interesting to touch upon general statistical characteristics of such a large corpus as ours. Fig. 1 shows that the frequency–rank dependence for the overall corpus can be roughly described with the known Zipf law. Similarly to the results [10–12], a crossover between the regions of core and extended vocabularies can be seen at approximately $r_c = (1 \div 2) \cdot 10^4$. The exponents corresponding to these regimes are equal to $\alpha_1 = 1.17$ and $\alpha_2 = 1.86$, where the Zipf law is expressed as $\Sigma F \sim r^{-\alpha}$. The parameters derived by us agree roughly with those reported in the earlier studies [10] ($r_c = (5 \div 6) \cdot 10^3$, $\alpha_1 = 1.01$ and $\alpha_2 = 1.92$), [11] ($r_c = 8 \cdot 10^3$, $\alpha_1 = 1.00$ and $\alpha_2 = 1.77$) and [12] ($r_c = 10^4$, $\alpha_1 = 1.08$ and $\alpha_2 = 1.70$). Moreover, it turns out that the standard deviations Δf and Δf_w are linked through $\Delta f \sim f_m^{\gamma}$ with the mean frequencies f_m and $f_{m,w}$, where $\gamma \approx 0.77$ (see Fig. 2 and discussions [13, 14]). The fact $0.5 < \gamma < 1$ testifies that some long-range effects are present for the word frequencies, thus hindering 'asymptotic' estimations of the frequency f for the case of infinitely large corpora. In other words, the relative frequency f cannot be interpreted mathematically as a 'probability'.

TABLE I. Some statistical characteristics of top ten word types with the highest total absolute frequencies ΣF in our corpus ($n = 4829$), and keywords of LOTR and WOE discussed further on. n_t implies the number of texts in the corpus that include a word t. Keywords of WOE are marked with asterisk

Word	ΣF	f_m	Δf	n_t
the	21705842	0.0585	0.0133	4829
and	10739848	0.0298	0.0083	4829
to	9979717	0.0255	0.0042	4827
of	9449847	0.0259	0.0093	4829
a	8592598	0.0228	0.0041	4825
he	5813341	0.0146	0.0067	4793
in	5623089	0.0151	0.0031	4827
i	5454258	0.0150	0.0103	4791
was	5119428	0.0131	0.0050	4799
it	4523035	0.0118	0.0036	4803
…	…	…	…	…
sam	30287	$7.28 \cdot 10^{-5}$	0.0007	1046
pippin	170	$4.13 \cdot 10^{-7}$	$1.31 \cdot 10^{-5}$	72
gandalf	14	$2.78 \cdot 10^{-8}$	$5.43 \cdot 10^{-7}$	14
frodo	11	$1.93 \cdot 10^{-8}$	$5.06 \cdot 10^{-7}$	10
hobbits	6	$1.31 \cdot 10^{-8}$	$3.80 \cdot 10^{-7}$	6
tanu *	5	$5.41 \cdot 10^{-9}$	$3.76 \cdot 10^{-7}$	1
gollum	4	$7.32 \cdot 10^{-9}$	$3.41 \cdot 10^{-7}$	3
gimli	3	$4.22 \cdot 10^{-9}$	$2.94 \cdot 10^{-7}$	1
legolas	3	$4.04 \cdot 10^{-9}$	$2.13 \cdot 10^{-7}$	2
middleearth	3	$1.11 \cdot 10^{-8}$	$6.38 \cdot 10^{-7}$	3
aragorn	2	$2.82 \cdot 10^{-9}$	$1.96 \cdot 10^{-7}$	1
boromir	2	$3.14 \cdot 10^{-9}$	$1.55 \cdot 10^{-7}$	2
mordor	2	$5.17 \cdot 10^{-9}$	$2.54 \cdot 10^{-7}$	2
enge *	1	$2.15 \cdot 10^{-9}$	$1.49 \cdot 10^{-7}$	1
gondor	1	$1.30 \cdot 10^{-9}$	$9.07 \cdot 10^{-8}$	1
tirith	1	$1.26 \cdot 10^{-9}$	$8.75 \cdot 10^{-8}$	1
eistaa *	0	0	0	0
faramir	0	0	0	0
fargi *	0	0	0	0
herilak *	0	0	0	0
kerrick *	0	0	0	0
murgu *	0	0	0	0
saruman	0	0	0	0
stallan*	0	0	0	0
ustuzou *	0	0	0	0
vaintè *	0	0	0	0
yilanè*	0	0	0	0

Now we analyze the extent to which the unique words are spread in our corpus. Unlike the definition given in Section 1 and focused on contrasting a corpus versus a text under study, here we redefine for a while the unique words as those that appear in a single text of a corpus (i.e., $n_t = 1$). To be consistent and do not deal with misprints and text recognition errors, we

have dropped any unique words that occur less than F_{th} times in a text. Nevertheless, our analysis has testified that the unique words represent a widely-spread phenomenon. In particular, only 38.5% of the texts include no unique words or, more precisely, include only those unique words that occur $F < 10$ times. In total, the unique word types make up roughly 1.7% of the corpus vocabulary, or 17 words per 1000 vocabulary items.

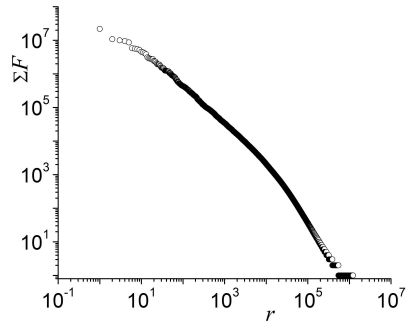

Figure 1. Dependence of total absolute frequency ΣF of words in our corpus on the word rank r introduced according to descending ΣF

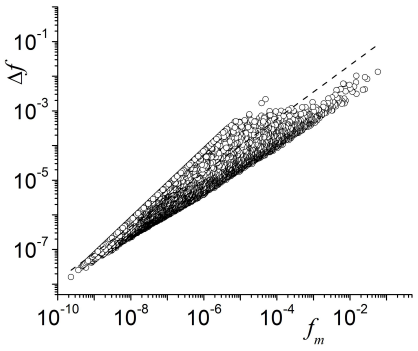

Figure 2. Dependence of frequency deviation on the mean frequency of words in our corpus. Dash line corresponds to linear fit with $\gamma = 0.77$

B. *Modifications of Keyword Detection Methods in the Presence of Unique Words*

Let a truly unique word t ($n_t = 0$) be available in the text under study. One cannot strictly compare the text with the corpus where this text is absent, since division by zero occurs in formulae (1)–(3) ($n_t = 0, f_m = 0$ and $\Delta f = 0$ in denominators – see also table II). There are two ways out, introducing either the overall text or its vocabulary into the corpus. Although there is no difference between the both cases for the TF-IDF method, the other measures have difficulties in the former case. Namely, we obtain for the unique words (the index "u")

$$R_u = n + 1 = const\ ,\ (R_w)_u = (L + l)/l = const\ ,$$

$$Z_u = \sqrt{n} = const\ ,\ (Z_w)_u = \sqrt{L/l} = const\ . \qquad (7)$$

Therefore the unique words will certainly dominate in the ranked list of keywords and, moreover, the scores of all of these words appear to be the same, irrespective of their frequency. This keyword degeneracy contradicts drastically the obvious claim that any keyness measure of a word should increase with increasing frequency of this word [9]. Eventually, this deteriorates the R and Z metrics. Note also that the unweighted and weighted cases in (7) agree perfectly with each other in the limiting case of equal-size text lengths ($l_1 = l_2 = ...$ and $L/l = n$, where the notation is explained in Section 2).

TABLE II. PERFORMANCE OF THE MAIN KEYWORD DETECTION METHODS UNDER THE CONDITION THAT THE TEXT UNDER STUDY INCLUDES THE UNIQUE WORDS: "+" OR "–" IMPLY THAT THE METHOD CAN OR CANNOT FUNCTION PROPERLY, WHILE "±" CORRESPONDS TO RESTRICTED FUNCTIONALITY

Choice	f/f_m method	TF-IDF method	Z-score method
1. Text absent in corpus	–	–	–
2. Text included in corpus	±	+	±
3. Vocabulary of text included in corpus	+	+	+

The simplest way to eliminate the above shortcoming of the R and Z metrics is to add only the vocabulary v of the text under study to the corpus. In other terms, we inject into the corpus a synthetic 'text', in which every word from the initial text occurs only once (i.e., its frequency amounts to $F = 1$, so that we have $l = v$). Instead of (7), one gets

$$R_u = (n+1)(v/l)F\ ,\ (R_w)_u = [(L+v)/l]F\ ,$$

$$Z_u = \frac{1}{\sqrt{n}}\left[(n+1)\frac{v}{l}F - 1\right],\ (Z_w)_u = \sqrt{\frac{v}{L}}\left(\frac{L+v}{l}F - 1\right), \quad (8)$$

where the length l now refers to the text under study only, while the total length of the corpus in case of its equal constituents amounts to $L = nv$ (i.e., each text in the corpus has the length v). In particular, $(f/f_m)_u$ and $(f/f_{m,w})_u$ are proportional to n, whereas Z_w and $(Z_w)_u$ to $\sqrt{n}$ at $n >> 1$, like in (7). A comparison of formulae (7) and (8) demonstrates that inclusion into the corpus of the vocabulary, instead of the whole text, repairs a desired behavior of the R and Z metrics ($R, Z \sim F$), at least asymptotically (i.e., for the large enough corpus). Although the TF-IDF works well for the both choices 2 and 3 described in table II, we have decided in favor of the choice 3, i.e. inclusion of the text vocabulary in the corpus, in order to compare all the methods under the same conditions.

C. *Comparison of Performance of Different Keyword Detection Methods. Similarity of the Methods*

The top ten ranked keywords obtained using the methods (1)–(3) and (6) are displayed in tables III–IX. Here the rank r corresponds to the keyness score not the frequency F. Notice also that the sum of normalized scores is not unit, because the normalization concerns all the words with $F \geq 10$ rather than the first ten words.

For the both texts, LOTR and WOE, the top keywords given by the T, Z, Z-IDF and TF-IDF-Z measures, in both their unweighted and weighted versions, reflect perfectly the

principal characters of the literary works and the main place-names, in their logical succession (for the sake of conciseness, similar results for Z-IDF and the weighted methods are not displayed in tables III–IX). For instance, the first words *frodo*, *gandalf* and *aragorn* in the lists correspond to the topmost protagonists of the novel LOTR. In the case of WOE, the top ranks belong to the main protagonists and antagonists, as well as the names of feuding races introduced in the novel. It is worth noting that the keyword lists include both the truly unique words of the texts under test (*eistaa*, *faramir*, *kerrick*, etc.) and the quasi-unique words found in some other texts (e.g., *frodo*, *gandalf* and *pippin* – see table I). Moreover, the top 200 or even more keywords from the lists obtained with the *T*, *Z*, Z-IDF and TF-IDF-Z scores in fact do not include a single word that can be considered as irrelevant. This demonstrates great resources of all of the above methods.

TABLE III. TOP 10 KEYWORDS FOUND USING NORMALIZED f/f_m SCORE FOR LOTR

r	Word	R_n	*F*
1	gondor	0.1598	433
2	aragorn	0.1341	708
3	middle-earth	0.0589	108
4	frodos	0.0542	129
5	tirith	0.0528	138
6	frodo	0.0524	1864
7	mordor	0.0497	256
8	gimli	0.0441	378
9	boromir	0.0432	267
10	legolas	0.0411	339

TABLE IV. TOP 10 KEYWORDS FOUND USING NORMALIZED TF-IDF SCORE FOR LOTR

r	Word	T_n	*F*
1	frodo	0.0672	1864
2	gandalf	0.0369	1080
3	aragorn	0.0327	708
4	hobbits	0.0219	567
5	gimli	0.0174	378
6	gollum	0.0168	399
7	pippin	0.0163	658
8	legolas	0.0148	339
9	faramir	0.0142	282
10	saruman	0.0132	262

TABLE V. TOP 10 KEYWORDS FOUND USING NORMALIZED TF-IDF SCORE FOR WOE

r	Word	T_n	*F*
1	kerrick	0.1123	884
2	vaintè	0.0703	553
3	herilak	0.0515	405
4	murgu	0.0398	313
5	yilanè	0.0386	304
6	ustuzou	0.0375	295
7	fargi	0.0367	289
8	stallan	0.0360	283
9	eistaa	0.0191	150
10	tanu	0.0179	153

On the other hand, both of the unweighted and weighted versions of *R* and TF-Z measures given by (1) and (4) manifest more or less serious drawbacks. The TF-Z score is simply inadequate. In spite of a quite correct top-ten keywords compatible with TF-IDF and the other metrics, the keyword list based upon TF-Z includes a couple of the most frequent stop-words on too high positions, presumably due to abundantly strong influence of their text frequencies. Indeed, under the condition $f >> f_m$, which holds true for the strongest keywords, we have TF-Z $\sim f^{\beta}$ with $\beta \sim 2$ in (4). We therefore conclude that the exponent β should not be as high as two, in order that a keyword detection method be successful (cf. with the effective exponent β close to ½, which has been introduced using quite another reasoning in the frame of domain-independent method [6]). On the other hand, TF-IDF-Z is also characterized by $\beta \sim 2$ at $f >> f_m$, though it demonstrates quite reliable data. Then a deeper reason for better performance of the TF-IDF-Z metrics can be associated with its better balance between word frequency and word uniqueness.

TABLE VI. TOP 10 KEYWORDS FOUND USING NORMALIZED Z-SCORE FOR LOTR

r	Word	Z_n	*F*
1	frodo	0.1106	1864
2	gandalf	0.0640	1080
3	aragorn	0.0422	708
4	hobbits	0.0337	567
5	gollum	0.0238	399
6	gimli	0.0225	378
7	legolas	0.0202	339
8	faramir	0.0168	282
9	gondor	0.0160	268
10	boromir	0.0159	267

TABLE VII. TOP 10 KEYWORDS FOUND USING NORMALIZED Z-SCORE FOR WOE

r	Word	Z_n	*F*
1	kerrick	0.1067	884
2	vaintè	0.0668	553
3	herilak	0.0489	405
4	murgu	0.0378	313
5	yilanè	0.0367	304
6	ustuzou	0.0356	295
7	fargi	0.0349	289
8	stallan	0.0342	283
9	tanu	0.0185	153
10	eistaa	0.0181	150

The *R* measure also reveals obvious weak points, though not so large-scale as with TF-Z. It tends to rank highest the unique words which are rarest in the corpus (see tables I and III), e.g. *gondor* ($n_t = 1$) and *aragorn* ($n_t = 1$). On the other hand, the keyness of the words that occur more often in the corpus becomes underestimated, in spite of their higher frequencies in the text (e.g., *frodo* with $n_t = 10$ or *gandalf* with $n_t = 14$ – see tables I and III). As a result, the central keyword *frodo* (7% and 11% importance, according to the normalized estimations of *T*- and *Z*-metrics – see tables IV and VI) is found beyond the top group of five, since it is mentioned more often in the other texts than, e.g., *gondor* and *aragorn*. The results of R_w-modification of this measure are qualitatively the same (e.g, *frodo* is ranked 9th).

TABLE VIII. TOP 10 KEYWORDS FOUND USING NORMALIZED TF-IDF-Z SCORE FOR LOTR

r	Word	$(\text{TF-IDF-Z})_n$	*F*
1	frodo	0.4678	1864
2	gandalf	0.1489	1080
3	aragorn	0.0868	708
4	hobbits	0.0466	567
5	gollum	0.0251	399
6	gimli	0.0247	378
7	legolas	0.0189	339
8	faramir	0.0150	282
9	pippin	0.0136	658
10	saruman	0.0130	262

TABLE IX. TOP 10 KEYWORDS FOUND USING NORMALIZED TF-IDF-Z SCORE FOR WOE

r	Word	$(\text{TF-IDF-Z})_n$	*F*
1	kerrick	0.4106	884
2	vaintè	0.1607	553
3	herilak	0.0862	405
4	murgu	0.0515	313
5	yilanè	0.0486	304
6	ustuzou	0.0457	295
7	fargi	0.0439	289
8	stallan	0.0421	283
9	eistaa	0.0118	150
10	tanu	0.0113	153

As a useful test for correct performance and quality of data obtained with different keyword detection methods, we compare the ranks of a standard stopword *and*. While this word is the next to last in the cases of TF-IDF, TF-IDF-Z and Z-IDF, its rank r_{and} is equal respectively to $r_{and} = 13$ (2125), 24 (2140), 1867 (2448) and 1828 (2450) with TF-Z, TF-Z_w, *R* and R_w, where the figures in brackets denote the total numbers of words that passed the frequency filters. This is why the approaches based upon TF-Z and, partly, upon a naive *R* score can be judged as insufficient from the viewpoints of performance and data reliability.

As stated before, the combined method Z-IDF shows the results similar to TF-IDF and *Z* (not shown in tables III–IX). Finally, of all the combined scores introduced in Section 2, one can distinguish TF-IDF-Z (see tables VIII and IX). Perhaps, it suggests the best keyword lists for the ranks as low as 250 for both LOTR and WOE. The only reservation about this method is too great relative importance of the foremost keywords. So, the cumulative contributions of the top-ten items into the sum of normalized keyness scores for all the words with $F \geq 10$ are equal to 25.1% and 46.0% for TF-IDF, 36.6% and 43.8% for Z-score, and 86.0% and 91.2% for TF-IDF-Z, where the first (second) figures correspond respectively to LOTR (WOE). It would be natural to cast some doubt on propriety of this score distribution, although the problem is open for discussion.

Perhaps, the largest problem of all corpus-based methods is unsuitably low ranks of the keywords with relatively large n_t's. For example, this is the case with the name of one of protagonists in LOTR, *sam*, which is expected to be within the top five. However, we have the ranks $r = 14$, 235, 301, 256, 308, 60 and 105, according to the *T*, *Z*, Z_w, Z-IDF, Z-IDF$_w$, TF-IDF-Z and TF-IDF-Z_w scores. This is because *sam* represents a common word in our corpus of literary works, with $\Sigma F \approx 3.03 \cdot 10^4$ and $n_t = 1046$ – see table I. Moreover, much greater problems would have arisen if the name of principal character in LOTR were, say, *John* or *And*, instead of *Frodo*. These names would surely have been crushed against the background of many important unique words present in this text. In other terms, any corpus-based metrics is exposed to the problem of (often unjustified) dominating of unique words over widely disseminated ones, and it seems impossible to overcome this intrinsic defect of the domain-dependent methods. This differs drastically from the domain-independent keyword detection approaches, which do not rely upon the frequency or uniqueness of words but instead consider the inhomogeneity of word distribution in a text [3] or the properties of word network in a text (see, e.g., [16, 17]). Nonetheless, it is noteworthy that the above problem is not so serious for the TF-IDF and, in part, TF-IDF-Z methods.

Since the outputs of different keyword detection methods cannot be fully comprehended while 'manually' comparing their top-ten keyword lists, we have undertaken their analysis using a standard cosine-based similarity measure for the 100 and 200 top keywords of LOTR. For conciseness we do not list the overall similarity matrix that contains all of the methods under different conditions. Some of the results obtained with $N = 100$ are displayed in table X, where the main attention is given to comparison with TF-IDF as a benchmark. Here all the results discussed above are described by a so-called case (i), when the extended corpus includes the vocabulary of LOTR. This corresponds to the choice 3 in table II. The alternative cases (ii) and (iii) that correspond to modified corpora will be explained in detail in the next Subsection.

TABLE X. SIMILARITY OF 100 TOP KEYWORD LISTS OBTAINED USING DIFFERENT NORMALIZED SCORES FOR LOTR IN THE CASES (i), (ii) AND (iii) (SEE THE TEXT)

Method (case (i))	Similarity with TF-IDF (case (i)), %	Similarity with case (ii), %	Similarity with case (iii), %
TF-IDF	100	98.1	97.2
Z	85.4	50.9	48.0
Z_w	83.2	32.9	29.1
Z-IDF	86.2	53.5	49.5
Z_w-IDF	82.7	28.0	29.6
TF-IDF-Z	62.3	70.1	73.5
TF-IDF-Z_w	62.7	65.2	69.0

As seen from table X, Z-IDF and, especially, Z-score are closer to TF-IDF, if compared with TF-IDF-Z. Of course, we cannot simply conclude about some advantages of the Z and Z-IDF measures on this ground, since the similarity represents a rather formal parameter. It does not take into account that any dissimilarity can imply deviations of the keyword list in either 'better' or 'worse' directions. For instance, a surely insufficient TF-Z measure turns out to have 95÷98% similarity with a quite successful TF-IDF-Z score. Therefore, the above figures should be treated with a great caution. In particular, TF-IDF-Z is dissimilar to TF-IDF primarily because of higher importance of the top keywords in the former method, rather than due to different keywords or even keyword sequence. Finally, the results for $N = 200$ top keywords are only slightly different from the case of $N = 100$, and the same is true of the results

obtained for WOE. Here the TF-IDF, Z and Z-IDF keyword lists are about 97% similar, while the TF-Z and TF-IDF-Z data reveals a nearly 70% similarity with TF-IDF.

D. A Case of Similar Texts Present in Corpus: Criticality of the Methods to Small n_t Changes

When we examine some text and compare it with a large corpus, we cannot a priori know whether this corpus already contains this text, or not. Moreover, the corpus can include some texts which are (more or less) similar to the text of our interest. Texts on similar topics, plagiarized ones, or the texts written by the same author are ready examples, of which vocabularies are close to that of our text. Since having the information of this kind checked up permanently would cost a heavy computational price, we are to submit to the fact that a single text or a couple of texts identical with (or similar to) our text can always be present in the corpus. Then we deal with small changes Δn_t of the initial number n_t of texts. Of course, such a small perturbation of corpus has to be regarded as insignificant for any words t with large enough initial n_t's, although this is not the case for the unique words. Since the latter are characterized by small initial numbers n_t, the relative changes $\Delta n_t / n_t$ for them can become large, and the same is true for the changes in their frequencies f_m. Therefore it would be natural to suppose that it is just the keyness scores of the unique words that can be affected the most by the above perturbation. As a result, a proper analysis of stability (or criticality) of different keyword detection methods must be performed for the unique words. Below we study this problem both empirically and analytically.

Suppose that we detect the keywords in our text, using the same choice 3 as before, i.e. we add its vocabulary to the corpus. Now let us consider three different cases: (i) our text (e.g., LOTR) if surely absent in the corpus, (ii) a further text identical to LOTR is already included in the corpus, and (iii) besides of LOTR as in case (ii), the corpus additionally involves several texts of the same author. They are as follows: "The Hobbit", "The Adventures of Tom Bombadil", "Tom Bombadil – Preface", "Sir Gawain and the Green Knight", "Farmer Giles of Ham", "Mythopoeia" and "Middle-Earth Glossary" by J. R. R. Tolkien.

The top ten keywords extracted with the TF-IDF, Z and TF-IDF-Z approaches in the cases (ii) and (iii) are presented in tables XI–XIII. They are subject to comparing with the case (i), which has been in fact displayed in our earlier tables IV, VI and VIII. It is evident from tables IV and XI that the TF-IDF score is the most stable if we pass from the case (i) to (ii) and (iii), while the TF-IDF-Z and Z-measures reveal a criticality with respect to the small changes made in the corpus. The effect is especially pronounced for the Z-measure (see table XII where both drastic changes in the scores and noticeable re-ranking of keywords occur, if compared with table VI). Notice also that the most principled changes happen during the transition (i) → (ii), while the quantitative differences between the cases (ii) and (iii) are less.

As expected, the greatest score variations are typical for the unique words with the least n_t. It is instructive to illustrate these effects for the words *gimli* ($n_t = 1$), *frodo* ($n_t = 10$) and *sam* ($n_t = 1046$). At the transition (i)→(ii) (i.e., after injecting the whole LOTR text into the corpus), the corresponding frequencies f_m and the standard deviations Δf change roughly as follows: 1) f_m: $4\cdot10^{-9} \to 2\cdot10^{-7}$ and Δf : $3\cdot10^{-7} \to 1\cdot10^{-5}$ (*gimli*), 2) f_m: $3\cdot10^{-8} \to 8\cdot10^{-7}$ and Δf : $6\cdot10^{-7} \to 6\cdot10^{-5}$ (*frodo*), and 3) f_m: $7\cdot10^{-5} \to 7\cdot10^{-7}$ and Δf : $7\cdot10^{-4} \to 7\cdot10^{-4}$ (*sam*). The Z-score changes displayed in table XII are their immediate consequences. In particular, the unique word *gimli* reveals a hyper-sensitivity to this small perturbation of the corpus. Note that all of the methods with reliable performance manage properly with such non-unique words as *sam*. Then the transition (i)→(ii)→(iii) yields in the rank changes 14→14→11, 235→235→222, 256→256→256 and 60→56→46 for TF-IDF, *Z*, Z-IDF and TF-IDF-*Z*, respectively.

TABLE XI. TOP 10 KEYWORDS FOUND USING NORMALIZED TF-IDF SCORE FOR LOTR IN THE CASES (ii) AND (iii)

r	Word	T_n	Word	T_n
1	frodo	0.0683	frodo	0.0682
2	gandalf	0.0377	gandalf	0.0378
3	aragorn	0.0319	aragorn	0.0305
4	hobbits	0.0222	hobbits	0.0213
5	gimli	0.0171	pippin	0.0172
6	pippin	0.0168	gimli	0.0168
7	gollum	0.0168	gollum	0.0163
8	legolas	0.0147	legolas	0.0146
9	faramir	0.0134	faramir	0.0130
10	saruman	0.0125	saruman	0.0121

TABLE XII. TOP 10 KEYWORDS FOUND USING NORMALIZED Z-SCORE FOR LOTR IN THE CASES (ii) AND (iii)

r	Word	Z_n	Word	Z_n
1	frodo	0.0143	gimli	0.0158
2	gandalf	0.0142	frodo	0.0157
3	aragorn	0.0140	legolas	0.0155
4	hobbits	0.0138	boromir	0.0144
5	gollum	0.0133	faramir	0.0142
6	gimli	0.0132	pippin	0.0141
7	legolas	0.0130	éomer	0.0140
8	faramir	0.0125	théoden	0.0140
9	gondor	0.0124	strider	0.0139
10	boromir	0.0124	aragorn	0.0136

TABLE XIII. TOP 10 KEYWORDS FOUND USING NORMALIZED TF-IDF-Z SCORE FOR LOTR IN THE CASES (II) AND (III)

r	Word	$(\text{TF-IDF-Z})_n$	Word	$(\text{TF-IDF-Z})_n$
1	frodo	0.1673	frodo	0.2017
2	gandalf	0.0916	gandalf	0.0908
3	aragorn	0.0766	aragorn	0.0779
4	hobbits	0.0525	gimli	0.0499
5	gimli	0.0387	pippin	0.0455
6	gollum	0.0383	legolas	0.0426
7	pippin	0.0338	faramir	0.0348
8	legolas	0.0328	gollum	0.0339
9	faramir	0.0289	boromir	0.0311
10	saruman	0.0263	éomer	0.0279

There is another nontrivial manifestation of instability for some of these methods. Namely, the rank r_{and} of the stopword

and (see Subsection 3.3) in the cases (ii) and (iii) remains the last but one for TF-IDF, Z-IDF and TF-IDF-Z, although we arrive at $r_{and} = 426$ or 586 for the Z or Z_w methods in the case (ii), and $r_{and} = 423$ or 585 for Z or Z_w in the case (iii). This result is rather poor, since the overall lists encounter respectively 2100 and 2500 word types. Therefore a notable intrinsic instability of the Z-score takes place under 'unprotected' conditions (ii) and (iii), which undermines the performance of the method.

Some extra aspects of stability of the keyword detection methods follow from the similarity among the cases (i), (ii) and (iii) (see table X). Here TF-IDF reveals ideal stability and remains beyond comparison (nearly 98% similarity of the results obtained in these cases), while TF-IDF-Z is the runner-up (72%). The other methods are exceedingly critical to this small reconstruction of the corpus. Like in the most of our comparisons, unweighted modifications of the methods perform better that their weighted counterparts, although this advantage becomes reduced for the methods with better stability. Issuing from comparison with the TF-IDF data for the particular cases (ii) and (iii) (not shown in table X), one concludes that TF-IDF-Z improves substantially its stability in these cases, unlike the situation described in Subsection 3.3 for the case (i). The same comparison reveals that, according to their similarities in the cases (ii) and (iii), the keyword detection methods can be tentatively divided into the two groups: Z and Z-IDF, and TF-IDF and TF-IDF-Z. In this relation the instability of Z-IDF with respect to small corpus perturbations (see table X) seems somewhat counterintuitive, since one might have expected the opposite due to availability of a 'stabilizing' IDF term (cf. also with the TF-IDF-Z method).

Finally, we emphasize that the performance of the methods (which has been mainly estimated in Subsection 3.3 in a rough manner, as a similarity with TF-IDF) and their stability represent rather independent, perhaps even 'orthogonal' characteristics. For instance, the TF-Z method with insufficient performance manifests, curiously enough, not the least stability at (i)→(ii)→(iii) transition, which can be found from its cross-comparisons with the other methods.

IV. Conclusions

The main point of the present study is the performance of corpus-based keyword detection methods under condition when a text under analysis includes so-called unique words, i.e. the words which are absent or very rare in the texts of a corpus. In order to understand better the work of the domain-independent methods, we have suggested a number of heuristic TF-IDF-like keyness measures, and compared them with the canonical TF-IDF measure under different conditions, using the literary texts LOTR and WOE as investigated objects, and a corpus that contains nearly 5000 literary works.

First we have demonstrated that, even under demanding frequency condition $F \geq 10$, the unique words still represent a typical phenomenon for our corpus. Then we have modified the domain-dependent keyword extraction methods such that they become able to work in the case of $n_t = 0$. The two complementary and partly independent points have been studied after that, adequacy of the keyword lists and the scores assigned to different keywords, as well as stable performance of the methods under conditions of perturbations of the reference corpus. A practical scheme for studying both the performance and the stability of the methods has been suggested, which is based on small changes in the corpus.

[1] K. S. Jones, "A statistical interpretation of term specificity," J. Document., vol. 28, pp. 11–21, 1972.

[2] C. D. Manning and H. Schütze, "Foundations of Statistical Natural Language Processing," Massachusetts: Massachusetts Institute of Technology, 1999.

[3] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñnoz, and A. M. Somoza, "Keyword detection in natural languages and DNA," Europhys. Lett., vol. 57, pp. 759–764, 2002.

[4] H. Zhou and G. W. Slater, "A metric to search for relevant words," Physica A, vol. 329, pp. 309–327, 2003.

[5] J. P. Herrera and P. A. Pury, "Statistical keyword detection in literary corpora," Eur. Phys. J. B, vol. 63, pp. 135–146, 2008.

[6] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado, and J. L. Oliver, "Level statistics of words: finding keywords in literary texts and symbolic sequences," Phys. Rev. E, vol. 79, pp. 035102(R), 2009.

[7] R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "Analysis of domain independent statistical keyword extraction methods for incremental clustering," Learning and Nonlinear Models – J. Brazil. Soc. Comp. Intelligence (SBIC), vol. 12, pp. 17–37, 2014.

[8] Guten Free Ebooks – Project Gutenberg. Retrieved from: https://www.gutenberg.org/

[9] H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of Research and Development, vol. 2, pp. 159–165, 1958.

[10] R. Ferrer i Cancho and R. V. Solé, "Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited," J. Quant. Linguist., vol. 8, pp. 165–173, 2001.

[11] M. Gerlach and E. G. Altmann, "Stochastic model for the vocabulary growth in natural languages," Phys. Rev. X, vol. 3, pp. 021006, 2013.

[12] V. V. Bochkarev, E. Yu. Lerner, and A. V. Shevlyakova, "Deviations in the Zipf and Heaps laws in natural languages," J. Phys.: Conf. Ser., vol. 490, pp. 012009, 2014.

[13] O. S. Kushnir, A. M. Bayovskyi, L. B. Ivanitskyi, and S. V. Rykhlyuk, "Fluctuations of the frequencies of letters and symbols in Ukrainian and Russian texts (in Ukrainian)," In: Proc. VII Ukrainian–Polish Conference "ELIT-2015", pp. 76–79. Lviv: Lviv University Publishing, 2015.

[14] O. S. Kushnir, O. S. Bryk, V. Y. Dzikovskyi, L. B. Ivanitskyi, I. M. Katerynchuk, and Y. P. Kis, "Statistical distribution and fluctuations of sentence lengths in Ukrainian, Russian and English corpora (in Ukrainian)," Bulletin of Lviv Polytechnic University, vol. 854, pp. 228–239, 2016.

[15] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," Contemp. Phys., vol. 46, pp. 323–351, 2005.

[16] Y. Matsuo, Y. Ohsawa, and V. Ishizuka, "KeyWorld: extracting keywords from document as a small world," In: K. P. Jantke and A. Shinohara (Eds.): Discovery Science. DS 2001. Lecture Notes in Computer Science, vol. 2226, pp. 271–281. Berlin, Heidelberg: Springer, 2001.

[17] G. K. Palshikar, "Keyword extraction from a single document using centrality measures," In: A. Ghosh, R. K. De, and S. K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 503–510, Berlin Heidelberg: Springer-Verlag, 2007.

[18] Y. Malevergne, V. Pisarenko, and D. Sornette, "Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities," Phys. Rev. E, vol. 83, pp. 036111, 2011.

[19] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," J. Document., vol. 60, pp. 503–520, 2004.

[20] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing and Management, vol. 39, pp. 45–65, 2003.