Unsupervised extraction of local and global keywords from a single text

Anonymous ACL submission

Abstract

We propose an unsupervised method to extract keywords from a single text. It is based on spatial distribution of words and the response of this distribution to a random permutation of words. The method allows inference of two types of keywords: local and global. Several classic literature texts demonstrate that such a classification of keywords is meaningful, and that this method significantly outperforms existing methods (such as YAKE and LUHN) in terms of keyword extraction. Additionally, it is language-independent, applies to short texts (e.g. scientific papers) and uncovers basic themes in texts. Yet another keyword extraction scheme is proposed, but it applies only to texts with many chapters. It is less efficient than the previous one, and is formally similar to metrics used to evaluate scientists (h-index).

1 Introduction

Keyword identification is important for information retrieval and NLP, but is also challenging, because this concept did not so far got a formal definition (Firoozeh et al., 2020; Hasan and Ng, 2014; Kaur and Gupta, 2010; Siddiqi and Sharan, 2015). There is a general understanding that a keyword is likely to be a non-polysemic noun that should relate to themes of the text, in contrast to text's rhemes. Poor results of evaluation metrics for keyword extraction prove that this task is not yet solved (Firoozeh et al., 2020; Hasan and Ng, 2014; Kaur and Gupta, 2010; Siddiqi and Sharan, 2015). There is even difficulty to generate ground truth keywords for documents (Firoozeh et al., 2020).

Several approaches for keyword extraction employ linguistic-based handcrafted rules (Firoozeh et al., 2020; Mihalcea and Tarau, 2004; Hulth, 2003). They lack language independence power and ability to rank keywords via their relevance. The mathematical approaches fall into two main categories: unsupervised and supervised. The latter includes methods like (Gollapalli and Yang, 2017; Witten and Nevill-Manning, 1999; Turney, 2003; Song and Hu, 2003), it is also worth to mention KeyBERT (key), which leverages pretrained BERT(Devlin et al., 2018) based embeddings for keyword extraction. Unsupervised approaches (Mihalcea and Tarau, 2004; Bougouin et al., 2013; Florescu and Caragea, 2017; Wan and Xiao, 2008; Jones, 2004; Robertson, 2004; Rose et al., 2010; Campos et al., 2018) include methods from statistics, information-theory and graph-based ranking. The first statistical approach to rank keywords based on the simple frequencies of words was proposed by Luhn (Luhn, 1958). He used Zipf's law for selecting frequent content words as keyword candidates (Luhn, 1958). The best known and widely used statistical approach is perhaps TF-IDF scoring function (Jones, 2004; Robertson, 2004; Firoozeh et al., 2020). It is based on the assumption that important words occur frequently in a given document, and appear rarely in the rest documents of a corpus. In graph-based methods (Mihalcea and Tarau, 2004; Bougouin et al., 2013; Florescu and Caragea, 2017; Wan and Xiao, 2008) text is represented as a graph where nodes are words and relations between words are expressed by edges. Better connected nodes (as determined by PageRank algorithm) relate to keywords (Brin and Page, 1998). These methods mainly differ by the principles used to generate edges between words (Bougouin et al., 2013). Graph-based methods need only document information, and hence are corpus independent in contrast to TF-IDF. Ref. (Ortuño et al., 2002) was one of the first attempts to use spatial distribution of words in detecting keywords. In (Ortuño et al., 2002), the variance of the spatial distribution is used for ranking keywords. Later works (Herrera and Pury, 2008; Carretero-Campos et al., 2013; Mehri and Darooneh, 2011; Mehri et al., 2015; Zhou and Slater, 2003) suggest several modifications which appears leading to improved results; e.g. Ref. (Zhou and Slater, 2003) proposes an alter-

native metric for keyword extraction.

Here, we also use the spatial distribution of words for keyword detection. Our corpus-independent method is based on comparing this distribution before and after a random permutation of words. In this way we capture two different types of keywords: global and local. Global keywords are spread through the text and the variance of their spatial distribution decreases after a random permutation of words. In contrast, local keywords are localized in certain parts of the text, so that the variance increases after a random permutation. Analyzing several classical texts, we saw that this structural difference between the keywords indeed closely relates to the content of the text; e.g. global and local keywords refer to (resp.) main and secondary characters of the text. Thus, global keywords give the general idea of the document, whereas local keywords focus our attention to some part of the text.

Our method provides significantly better efficiency of keyword extraction then several known methods including LUHN (Luhn, 1958) and YAKE (Campos et al., 2018). In contrast to LUHN and YAKE, it does have a well-working score for keywords which allows to uncover themes of the text. Our method applies to relatively short text (scientific papers) and is nearly language-independent, as verified using translations in three languages: English, Russian and French.

The rest of the paper is organized as follows. In section 2, we introduce the main method analyzed in this work and apply it to a few long texts. Section 3 studies shorter texts (scientific papers). Section 4 is devoted to another keyword extraction method that employs the fact that a long text is divided over sufficiently many chapters. We summarize in the last section.

2 Method

2.1 Lemmatization of texts

English texts were preprocessed using WordNetLemmatizer imported from nltk.stem (nlt). This library looks for lemmas of words from the WordNet Database. The lemmatization uses corpus for excluding stop words (functional words) and WordNet corpus to produce lemmas. WordNetLemmatizer identifies the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire text. We applied this lemmatization algorithm on nouns, adjectives, verbs and adverbs to get maximal clean up of the text. Any stemming procedure will be inappropriate for our purposes of extracting keywords, since stemming may mix different parts of speech.

For inflected languages (e.g. Russian), the lemmatization rules are more complex. For French and Russian texts we used lemmatizers LEFFF (fre) and pymystem3 (rus), respectively.

2.2 Spatial distribution of words

Let $w_{[1]}, ..., w_{[\ell]}$ denote all occurrences of a word w along the text. Let ζ_i denotes the number of words (different from w) between $w_{[i]}$ and $w_{[i+1]}$; i.e. $\zeta_i + 1 \geq 1$ is the number of space symbols between $w_{[i]}$ and $w_{[i+1]}$. Define the average period $t(w)$ of this word w, and the the spatial frequency $\tau(w)$ via (Yngve, 1956):

$$t(w) = \frac{1}{\ell - 1} \sum_{i=1}^{\ell-1} (\zeta_i + 1), \qquad (1)$$

$$\tau(w) \equiv 1/t(w). \qquad (2)$$

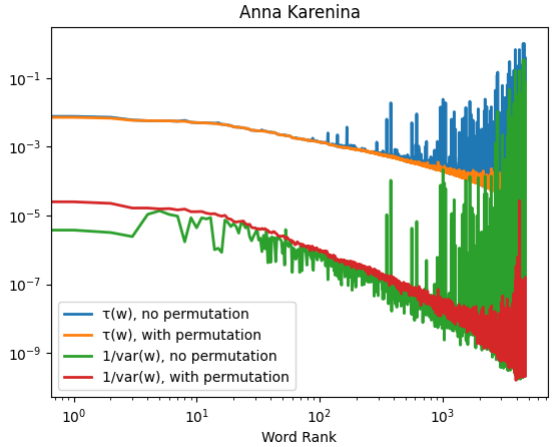
Eq. (1) is not defined for $\ell = 1$, i.e. for words that occur only once; hence such words are to be excluded from consideration. Note that $(\ell-1)(t(w)-1)$ equals to the number of words that differ from w and occur between $w_{[1]}$ and $w_{[\ell]}$. Hence $t(w)$ will stay intact under redistributing $w_{[2]}, ..., w_{[\ell-1]}$ for fixed $w_{[1]}$ and $w_{[\ell]}$. Hence a random permutation of all words in the text will leave $t(w)$ nearly intact for frequent words, and will increase it for not frequent words; see Figs. 1(a) and 1(b). (Random permutations were generated via Python's numpy library (per).) Here $f(w) = N_w/N$ is the (ordinary) frequency of w, where N_w is the number of times w appeared in the text, while N is the full number of words in the text. Appendix A explains an interesting relation $\tau(w) > f(w)$ that holds for the majority of words.

Given the average (1), let us define also the variance of the spatial period for word w (Herrera and Pury, 2008; Carretero-Campos et al., 2013; Mehri and Darooneh, 2011; Mehri et al., 2015; Zhou and Slater, 2003; Ortuño et al., 2002; Yngve, 1956; Carpena et al., 2009; Montemurro and Zanette, 2010):

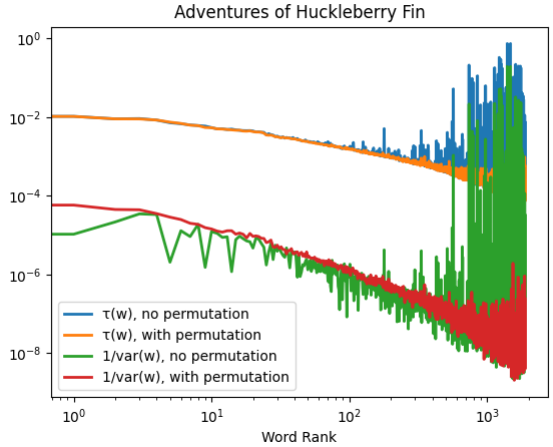
$$\mathrm{var}(w) = -t^2(w) + \frac{1}{\ell - 1} \sum_{i=1}^{\ell-1} (\zeta_i + 1)^2. \quad (3)$$

This quantity is already not invariant with respect to word permutations. Using (3), we define

$$A(w) = \frac{\mathrm{var}_{\mathrm{perm}}(w)}{\mathrm{var}(w)}, \qquad (4)$$



(a)



(b)

Figure 1: For two texts – *Anna Karenina* by L. Tolstoy (Tolstoy, 2013) (a) and *Adventures of Huckleberry Finn* by M. Twain (b) – we presented space frequency $\tau(w) = 1/t(w)$ and the inverse space variance $1/\mathrm{var}(w)$ *versus* word rank for all distinct words w of each text; cf. (1, 3). Ranking of distinct words was done via their frequencies, i.e. the most frequent word got rank 1 *etc*. We also show $\tau(w) = 1/t(w)$ and $1/\mathrm{var}(w)$ after a random permutation of the words. It is seen that the random permutation leaves $\tau(w) = 1/t(w)$ unaltered for frequent words. In contrast, $1/\mathrm{var}(w)$ is seriously changed by the random permutation.

where $\mathrm{var}_{perm}(w)$ is calculated via (3) but after a random permutation of all words of the text.

When checking the values of $A(w)$ for all distinct words of several texts, we concluded that sufficiently small and sufficiently large values of $A(w)$,

$$A(w) \leq \frac{1}{5}, \tag{5}$$

$$A(w) \geq 5, \tag{6}$$

can be employed for deducing certain keywords of the text. Eq. (5) uncovers global keywords of the text, i.e. keywords that go through the whole text. Taking a smaller value $\frac{1}{5} \leq A(w) \leq \frac{1}{3}$ in (5) leads to selecting a group of lower frequency global keywords. This effect is shown in Figs. 1(a) and 1(b) for two classic texts: *Anna Karenina* by L. Tolstoy (Tolstoy, 2013) and *Adventures of Huckleberry Finn* by M. Twain (Twain, 2003). When the words are arranged with respect to decreasing frequency, global keywords appear as local minima of $1/\mathrm{var}(w)$; cf. (3). These local minima do not survive a random permutation leading to a small value of $A(w)$ in (5); see Figs. 1(a) and 1(b). Likewise, (6) refers to local keywords, i.e. those that appear in specific places of the text. In Figs. 1(a) and 1(b) they are seen as local maxima of $1/\mathrm{var}(w)$. In contrast to local minima, maxima are located in the domain of infrequent words. Local maxima also disappear after a random permutation. Hence $A(w)$ in (5) assumes a larger value.

These relations of (5) and (6) with (resp.) global and local keywords make intuitive sense. As we checked in detail, spaces between global keywords assume a broad range of values. This distribution becomes more uniform after the random permutation, hence the variance decreases; cf. (5). Local keywords refer to infrequent words and are localized in a limited range of a text. Hence a random permutation obviously increases the dispersion, as implied by (6).

As our method relies on random permutations, our results are formally dependent on the realization of these permutations. Such a dependence is weak: we noted that only a few keywords change from one realization to another. However, we cannot avoid random permutations. In particular, we cannot rely on theoretical models of a random text; see e.g. (Herrera and Pury, 2008; Mehri and Darooneh, 2011). In a long text, the distribution for spaces ζ_i [c.f. (1)] after a random permutation is asymptotically geometrical. But for the majority of keywords this asymptotic is not reached, since their frequency is not big.

2.3 Keywords extracted from *Anna Karenina*

The evaluation of extracted keywords was done within our expert knowledge of classic Russian literature and specifically works by Tolstoy (Gustafson, 2014). We separated these keywords into 9 thematic groups: proper names of major char-

Table 1: Words of *Anna Karenina* extracted via our method. For global keywords strong and weak cases mean (resp.) that the words w were chosen according to $A(w) \leq \frac{1}{5}$ and $\frac{1}{5} \leq A(w) \leq \frac{1}{3}$; cf. (5). Local keywords were chosen according to $A(w) \geq 5$; see (6). Keyword classes are denoted by upper indices. $^{(1)}$ proper names of major characters; $^{(2)}$ proper names of secondary characters; $^{(3)}$ animals; $^{(4)}$ trains and railway; $^{(5)}$ hunting; $^{(6)}$ rural life and agriculture; $^{(7)}$ local government (zemstvo); $^{(8)}$ nobility life and habits; $^{(9)}$ religion. The last group $^{(10)}$ denotes words that were identified as keywords, but did not belong to any of the above groups. For each of 3 cases more frequent words appear first. It is seen that more frequent words are more likely to be keywords.

Global keywords strong cases	Global keywords weak cases	Local keywords
levin$^{(1)}$, anna$^{(1)}$, vronsky$^{(1)}$, kitty$^{(1)}$, alexey$^{(1)}$, stepan$^{(1)}$, alexandrovitch$^{(1)}$, arkadyevitch$^{(1)}$, dolly$^{(1)}$, sergey$^{(1)}$, ivanovitch$^{(1)}$, peasant, darya$^{(1)}$, alexandrovna$^{(1)}$, varenka$^{(1)}$, lidia$^{(1)}$, death, ivanovna$^{(1)}$, laborer$^{(6)}$, mow$^{(6)}$, district$^{(7)}$, stahl$^{(1)}$, bailiff$^{(5)}$, gun$^{(5)}$, snipe$^{(5)}$, plough$^{(6)}$, rain, lesson$^{(10)}$, lord$^{(9)}$, acre$^{(6)}$, platform$^{(4)}$, natalia$^{(1)}$, built, rich, overlook, river, crime$^{(10)}$, rail$^{(6)}$, relate, throb, contrast, puzzle, cheat$^{(10)}$, oppress, irrational$^{(10)}$	love, princess$^{(8)}$, brother, carriage$^{(4)}$, horse$^{(8)}$, prince$^{(8)}$, doctor$^{(8)}$, countess$^{(8)}$, madame$^{(8)}$, sviazhsky$^{(1)}$, land$^{(6)}$, seryozha$^{(1)}$, konstantin$^{(1)}$, picture, oblonsky$^{(1)}$, nikolay$^{(1)}$, agafea$^{(2)}$, katavasov$^{(2)}$, grass$^{(6)}$, yashvin$^{(1)}$, shoot$^{(5)}$, mihalovna$^{(2)}$, officer$^{(8)}$, box, marshal$^{(7)}$, mare$^{(6)}$, priest$^{(9)}$, tree$^{(6)}$, forest$^{(6)}$, laska$^{(3)}$, law$^{(10)}$, landowner$^{(6)}$, realize, scythe$^{(6)}$, telegram$^{(8)}$, meadow$^{(6)}$, bedroom$^{(8)}$, argument, sledge, nobleman$^{(8)}$, paint, article$^{(8)}$, professor$^{(8)}$, scream, sky, trap, birch$^{(6)}$, cow$^{(6)}$, debt$^{(10)}$, rent, punish, sow$^{(6)}$, annushka$^{(2)}$, lightly, sportsman$^{(8)}$, myakaya$^{(2)}$, invalid, smart, parent, vividly, maman$^{(8)}$, institution$^{(7)}$, stable, distance, salary$^{(10)}$, educate, firm, skirt, mahotin$^{(2)}$, reconciliation, yellow, plump, childrens, tatar$^{(2)}$, outer, steward$^{(8)}$, cousin, loathsome, sharp, splash, armchair$^{(8)}$, understands, coarse, quicken, grace, delicious, director$^{(8)}$, unseen, selfpossession, cheese, rate, physically, timidity, tucked, reassure, sunday, compartment, frost, minister$^{(8)}$, won, king, repent, clock, wage, shock, uncertain, deliver, cream, silently, monday, captain$^{(8)}$, shaft$^{(6)}$, matrona$^{(8)}$, strictly, original	vassenka$^{(2)}$, golenishtchev$^{(2)}$, election$^{(7)}$, skate$^{(10)}$, varvara$^{(2)}$, pyotr$^{(2)}$, lizaveta$^{(2)}$, landau$^{(2)}$, petrovna$^{(2)}$, gladiator$^{(3)}$, metrov$^{(2)}$, tit$^{(2)}$, vote$^{(7)}$, froufrou$^{(3)}$, ryabinin$^{(2)}$, volunteer$^{(8)}$, nevyedovsky$^{(2)}$, duel$^{(8)}$, scandal$^{(8)}$, tribe$^{(10)}$, snetkov$^{(2)}$, lukitch$^{(2)}$, mower$^{(6)}$, deacon$^{(9)}$, native, korsunsky$^{(2)}$, hospital, remote, mazurka$^{(8)}$, pilate$^{(10)}$, sappho$^{(10)}$, villa$^{(8)}$, rival, reed$^{(6)}$, bridegroom$^{(8)}$, krak$^{(3)}$, merkalova$^{(2)}$, vorkuev$^{(2)}$, photograph$^{(8)}$, yegor$^{(2)}$, mitya$^{(2)}$, kapitonitch$^{(2)}$, architect$^{(8)}$, intensely, elect$^{(7)}$, golenishtchevs$^{(2)}$, pa$^{(8)}$, birthday, trousseau$^{(8)}$, transition, chalk, potato$^{(6)}$, kritsky$^{(2)}$, ergushovo$^{(6)}$, katya$^{(2)}$, weep, sympathetic, repair, mais$^{(8)}$, seryozhas$^{(2)}$, ballroom$^{(8)}$, classical, vozdvizhenskoe$^{(6)}$, technique, bedchamber$^{(8)}$, opium$^{(8)}$, penetrate, tchirikov$^{(2)}$, rider, palazzo$^{(8)}$, crown$^{(8)}$, remove, miracle, intolerable, turk$^{(2)}$, ballot$^{(7)}$, custom, nevsky$^{(8)}$, adultery$^{(8)}$, ditch, musical

acters; proper names of secondary characters; animal names; trains and railway; hunting; rural life and agriculture; local governance; nobility life and habits; religion; see Table 1. Recall that this classic novel with more than 800 pages features more than a dozen major characters and many lesser characters. The names of these characters are certainly keywords, because they inform us about character's gender ('anna' *vs.* 'vronsky'), age ('alexandrovitch' *vs.* 'seryozha') and the social strata; e.g. 'tit' *vs.* 'levin'. Proper nouns provide additional information due to name symbolism employed by Tolstoy; e.g. 'anna'='grace'; 'alexey'='reflector'; 'levin'='leo' is the *alter ego* of Tolstoy (Gustafson, 2014).

All the main character names came out from our method as strong global keywords holding condition $A(w) \leq \frac{1}{5}$ in (5): 'levin', 'anna', 'vronsky', 'kitty', 'alexey', 'stepan', 'dolly', 'sergey'; see Table 1 for details. Many pertinent lesser characters came out as local keywords, as determined via condition (6); e.g. 'vassenka', 'golenishtchev', 'varvara'; see Table 1. Important characters that are not the main actors came out as weak global keywords, e.g. 'seryozha', 'yashvin', 'sviazhsky'.

The novel is also known for its animal characters that play an important role in Tolstoy's symbolism (Gustafson, 2014). Our method extracted as local keywords the four main animal characters: 'froufrou', 'gladiator' 'laska', 'krak'. Trains are a motif throughout the novel (they symbolize the modernization of Russia), with several major plot points taking place either on passenger trains or at stations in Russia (Tolstoy, 2013; Gustafson, 2014). Our method extracted among the global keywords 'carriage', 'platform' and 'rail'. Hunting scenes are important in the novel depicting the life of Russian nobility. Accordingly, our method uncovered keywords related to that activity: 'snipe', 'gun', 'shoot'. Two major social themes considered in the novel are local democratic governance (Zemstvo) and the agricultural life of by then mostly rural Russia. For the first we extracted keywords: 'district', 'bailiff', 'election' *etc*. And for the second: 'mow', 'lord', 'acre', *etc*. A large set of keywords are provided by Russian nobility's living and manners, including their titles, professions and habits; see Table 1. Religion and Christian faith is an important subject of the novel. In this context, we noted keyword 'Lord', 'priest', 'deacon'; see Table 1. Finally, a few words stayed out of these thematic groups but was identified as keywords: 'lesson', 'crime', 'cheat', 'salary', 'irrational', 'law', 'skate', 'tribe'.

2.4 Comparison with known methods of keyword extraction and language dependence

Using *Anna Karenina* (Tolstoy, 2013), we compared our approach with two well known methods that also apply to a single text (i.e. do not require corpus): LUHN (Luhn, 1958) and YAKE (Campos et al., 2018); see also (yak) that discusses advantages of YAKE with respect to several other methods.

– 282 words were extracted via each method. Then keywords were identified using our general expertise on classic Russian literature. Table shows that for three languages (English, Russian, French) our method is better in terms of the percentage of extracted keywords. The relatively poor performance of YAKE and LUHN can be explained via the fact they focus on relatively short content words that are not likely to be keywords. We quantified this by calculating the mean number of letter in each set of 282 words. For our method, LUHN and YAKE the mean is (resp.) 6.95, 5.43 and 5.5; cf. the fact that the average number of letters in English content word is 6.47 (for stop word it is 3.13) (Miller et al., 1958).

– The three methods have scores for words. In LUHN and our method the score coincides with the word frequency. However, for LUHN and YAKE the score did not correlate with the feature of being keyword. For our method it certainly did, i.e. by selecting only high-score words we can significantly enlarge the percentage of keywords compared to what is seen in Table 2. These two facts (low density of keywords plus no correlation with their score) make impossible to extract thematic groups of keywords via LUHN and YAKE; cf. the discussion after (6).

– Another comparison criteria between the three methods is the amount of nouns in words that were not identified as keywords. This criterion is a proxy for the difficulty of identifying keywords, which are known to be mostly nouns. Our method again fares better than both LUHN and YAKE; see Table 2.

Table 2 also addresses the language dependence of the three methods that was studied in three versions (English, Russian and French) of *Anna Karenina*. It is seen that our method performs compa-

rably for English and Russian, which are morphologically quite distinct languages. For French the performance is worse, but overall still comparable with English and Russian. Altogether, our method is language-independent.

3 Shorter texts: scientific papers

Our main example is a well-known paper written by Jaynes (Jaynes, 1957) in the cross-link of statistical physics (that studies features of many-particle systems in terms of entropy, energy and temperature) and probabilistic inference, which deals with random events, (subjective) probability events, estimation *etc*. These two different fields became mutually beneficial after Ref. (Jaynes, 1957) proposed the maximum-entropy method (Jaynes, 1982). Hence we expect two different sets of keywords.

It turns out that a relatively short length of Ref. (Jaynes, 1957) prevents the direct applicability of (5, 6). Instead, we followed the logic of Figs. 1(a) and 1(b): we ranked all distinct words of Ref. (Jaynes, 1957) with their frequencies, and then looked within this sequence for local minimas of $A(w)$; cf. (4). In a very few cases, where the local maxima was quasi-degenerate, i.e. two nearby words have close values of $A(w)$, we took the word that also provided a local maxima for $A_4(w)$ that is defined analogously to $A(w)$ in (4), but with the four-order variance $\mathrm{var}_4(w) = \frac{1}{\ell-1}\sum_{i=1}^{\ell-1}(\zeta_i+1-t(w))^4$ instead of the usual variance in (3). Words from the first column of Table 3 came out in this way (we mention only the first such 15 words, and the number in brackets is the frequency rank for each word). It is seen that not much keywords related to statistical physics came out. Looking at local maxima of $A(w)$ among the ranked words produced the the second column of Table 3. This set provides more non-keywords than in the first column. Still the majority are keywords, and some of them are highly-relevant, e.g. 'maximum-entropy'.

The method is limited (as compared e.g. to the analysis of *Anna Karenina*), since Ref. (Jaynes, 1957) is a relatively short text. Hence we tried the following extension of the method: we repeated the text two times, then applied a random permutation to the whole (twice longer) text and implemented (4). A new set of keywords came out via selecting local minimas of $A(w)$; see the third column of Table 3. It is seen that most keywords now relate to statistical physics. Combining the three columns of Table 3 together we get a set of keywords that does reflect the interdisciplinary character of (Jaynes, 1957). A peculiar point of scientific papers is that the first 5-10 most probable words do likely contain keywords. However, many keywords are not among the most-probable words. Our method was able to find them, as seen in Table 3. We should mention that some obvious keywords of (Jaynes, 1957) were not detected via our method.

4 Keyword extraction and distribution of words over chapters

Long texts are frequently divided into sufficiently many chapters. It is an interesting question whether this fact can be employed as an independent criterion for extracting keywords. To search for such criteria, let us introduce the following basic quantities. Given a word w and chapters $c = 1, .., N_{\mathrm{chap}}$ we define $m_w(c) \geq 0$ as the number of times w appeared in chapter c. Likewise, let $V_w(s)$ be the number of chapters, where w appeared $s \geq 0$ times; i.e. $\sum_{s\geq s_0} V_w(s)$ is the number of chapters, where w appears at least s_0 times. We have $\sum_{c=1}^{N_{\mathrm{chap}}} m_w(c) = N_w$, and

$$\sum\nolimits_{s\geq 0} sV_w(s) = N_w, \tag{7}$$

where N_w is the number of times w appears in the text. Hence, $m_w(c)/N_w$ is the probability that taking w randomly will end up in chapter number c. $sV_w(s)/N_w$ is the probability that taking w randomly will end up in a chapter, where w appear s times.

It appears that quantities deduced from $m_w(c)/N_w$ do not lead to useful predictions concerning keywords. In particular, this concerns the entropy $-\sum_{c=1}^{N_{\mathrm{chap}}} \frac{m_w(c)}{N_w} \ln \frac{m_w(c)}{N_w}$ and correlation function $\sum_{c_1,c_2=1}^{N_{\mathrm{chap}}} |c_1 - c_2| m_w(c_1) m_w(c_2)$ together with its generalizations. In contrast, the following mean

$$\sum\nolimits_{s\geq 0} \frac{s^2 V_w(s)}{N_w}, \tag{8}$$

related to $sV_w(s)/N_w$ predicts sufficiently many global keywords; see Table 4. Similar results are found upon using the entropy $-\sum_{s\geq 0} \frac{sV_w(s)}{N_w} \ln \frac{sV_w(s)}{N_w}$ instead of (8). Eq. (8) is calculated for each word and then words with largest value of (8) are selected. For *Anna Karenina*, at least the first 35-36 words selected in this

Table 2: Comparison of 3 different keyword extraction methods for English, Russian and French version of *Anna Karenina*. Here "nouns" means the percentage of nouns in non-keywords. For all cases our method fares better than LUHN and YAKE, whose performances are comparable.

Method	English keywords	English nouns	Russian keywords	Russian nouns	French keywords	French nouns
LUHN	15.6 %	54 %	14.1%	51.1%	19.2%	62.3 %
YAKE	15.6 %	55 %	14.8%	49.2%	18%	60 %
Our method	55.6 %	82 %	55%	86.2%	50.7%	77.3%

Table 3: Keywords of Ref. (Jaynes, 1957) extracted via various means. We shadowed non-keywords and underlined keywords related to statistical physics. Other words are keywords related to probabilistic inference. Square brackets indicate the rank of the word (ranked according to the frequency).

Local minima of $A(w)$ defned via (4)	Local maxima of $A(w)$	Local minima of $A(w)$ for the text repeated two times
probability [1], distribution [4], function [7], prediction [12], temperature [14], fact [24], subjective [26], argument [29], event [34], uncertainty [36], mathematical [42], form [47], method [50], equal [54], expectation [58],	statistical [1], theory [5], problem [9], case [11], maximum-entropy [13], inference [15], type [20], value [24], macroscopic [27], point [32], knowledge [40], photon [44], objective [48], average [53], question [57], total [62], maximum [66]	probability [1], entropy [6], energy [8], prediction [12], temperature [14], estimate [18], condition [20], reason [25], argument [29], event [32], noise [36], total [56], heat [62], definite [78], particle [94]

Table 4: First column: 36 words from *Anna Karenina* that have the highest score of YAKE (Campos et al., 2018). Keywords are indicated by the number of their group; see Table 1. Among 36 words there are 25 non-keywords. Keywords refer mostly to group $^{(1)}$.
Second column: 36 words of *Anna Karenina* extracted via looking at distribution of words over chapters, i.e. at the largest value of (8). Only 2 words out of 36 are not keywords. Several keyword groups are represented.

36 words having largest score of YAKE	36 words having largest values of (8)
levin$^{(1)}$, anna$^{(1)}$, vronsky$^{(1)}$, alexey$^{(1)}$, kitty$^{(1)}$, stepan$^{(1)}$, hand, alexandrovitch$^{(1)}$, smile, thought, arkadyevitch$^{(1)}$, time, love, face, eye, felt, man, feel, talk, life, answer, day, wife, begin, long, knew, turn, child, sergey$^{(1)}$, husband, work, princess$^{(8)}$, room, ivanovitch$^{(1)}$, people, woman	levin$^{(1)}$, alexey$^{(1)}$, alexandrovitch$^{(1)}$, varenka$^{(2)}$, vronsky$^{(1)}$, kitty$^{(1)}$, doctor$^{(8)}$, stepan$^{(1)}$, scythe$^{(6)}$, anna$^{(1)}$, arkadyevitch$^{(1)}$, marsh$^{(6)}$, countess$^{(8)}$, katavasov$^{(2)}$, priest$^{(9)}$, darya$^{(1)}$, veslovsky$^{(2)}$, alexandrovna$^{(1)}$, seryozha$^{(1)}$, mare$^{(6)}$, sviazhsky$^{(2)}$, mihailov$^{(2)}$, brother, dolly$^{(1)}$, grass$^{(6)}$, sergey$^{(1)}$, princess$^{(8)}$, mow$^{(6)}$, marshal$^{(7)}$, konstantin$^{(2)}$, ivanovitch$^{(2)}$, peasant$^{(6)}$, lidia$^{(1)}$, sick, petritsky$^{(2)}$

way are keywords. Minor exclusions are seen in Table 4, which also shows that this method is much better than YAKE both in quantity and quality of keyword extraction. The advantage of this chapter-based method is that it does not depend on random permutations. The drawbacks are seen above: it depends on the existence of sufficiently many chapters (hence it certainly does not apply to texts with a few or no chapters), and it addresses only some of keywords.

$V_w(s)$ effectively appears in scientometry: the word w, chapters of the text, and $V_w(s)$ can be mapped to (resp.) a scientist, papers he/she produced, and the number of citations each paper got (Sidiropoulos et al., 2007). Using this analogy, one can define for a word w its h-index h_w: w appears h_w times in at most h_w chapters (Sidiropoulos et al., 2007). A bigger h_w means that w appears more in a larger number of chapters. However, when it comes to uncovering keywords, h_w is less useful than (8).

5 Conclusion

We proposed a method for extracting keywords from a single text. The method employs spatial inhomogeneties in word distribution and motivates the introduction of two types of keywords, local and global. For long texts our analysis confirms that such a separation is semantically meaningful. The method was illustrated on several classic literature texts and scientific papers. Our main examples are Refs. (Tolstoy, 2013) and (Jaynes, 1957). In both situations we relied on expert evaluation of keywords and were able to extract thematic groups of keywords. The semantic difference between local and global keywords is blurred for short texts.

The method outperforms several existing methods for keyword extraction, such as LUHN (Luhn, 1958) and YAKE (Campos et al., 2018). The advantage of our method is not occasional, since we confirm that it generally extracts more nouns and longer content words than YAKE and LUHN. There is generally a correlation between both of these features and being a keyword. Our method is also language-independent, to the extent we were able to check with several translations of the same text. It shares this advantage with LUHN and YAKE.

We also worked out a method of keyword extraction that uses the fact that a text has sufficiently many chapters. This method is working better than LUHN and YAKE, but it is inferior to the previous one. However, it does have interesting similarities with metrics that are proposed to evaluate the productivity of scientists. We believe this method does have a potential for further development.

Our future work will be adding some functionality for n-grams analyses, so that we can extract from a text not only single words but also phrases of length 2 and bigger. Yet another feature we are going to implement is to modify the spatial mean and variance of the word [see (1, 4)] such that they reflect the local frequency of the word.

Acknowledgements

This work was supported by SCS of Armenia, grants No. 21AG-1C038. We acknowledge discussions with A. Khachatryan, K. Avetisyan and Ts. Ghukasyan. We thank S. Tamazyan for helping us with French texts.

References

French lefff lemmatizer.

Keybert.

Natural language toolkit (nltk).

numpy.random.permutation.

A python wrapper of the yandex mystem 3.1 morphological analyzer.

Yet another keyword extractor (yake).

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pages 806–810. Springer.

Pedro Carpena, Pedro Bernaola-Galván, Michael Hackenberg, AV Coronado, and JL Oliver. 2009. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E*, 79(3):035102.

C Carretero-Campos, P Bernaola-Galván, AV Coronado, and P Carpena. 2013. Improving statistical keyword

detection in short texts: Entropic and clustering approaches. *Physica A: Statistical Mechanics and its Applications*, 392(6):1481–1492.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nazanin Firoozeh, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. 2020. Keyword extraction: Issues and methods. *Natural Language Engineering*, 26(3):259–291.

Corina Florescu and Cornelia Caragea. 2017. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1105–1115.

Li-X.-L. Gollapalli, S. D. and P. Yang. 2017. Incorporating expert knowledge into keyphrase extraction. In *In Thirty-First AAAI Conference on Artificial Intelligence*, pages 3180–3187.

R.F. Gustafson. 2014. *Leo Tolstoy: Resident and Stranger*. Princeton University Press, Princeton.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273.

Juan P Herrera and Pedro A Pury. 2008. Statistical keyword detection in literary corpora. *The European Physical Journal B*, 63(1):135–146.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223.

Edwin T Jaynes. 1957. Information theory and statistical mechanics. *Physical review*, 106(4):620.

Edwin T Jaynes. 1982. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jasmeen Kaur and Vishal Gupta. 2010. Effective approaches for extraction of keywords. *International Journal of Computer Science Issues (IJCSI)*, 7(6):144.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Ali Mehri and Amir H Darooneh. 2011. The role of entropy in word ranking. *Physica A: Statistical Mechanics and its Applications*, 390(18-19):3157–3163.

Ali Mehri, Maryam Jamaati, and Hassan Mehri. 2015. Word ranking in a single document by jensen–shannon divergence. *Physics Letters A*, 379(28-29):1627–1632.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Georges A Miller, Edwin Broomell Newman, and Elizabeth A Friedman. 1958. Length-frequency statistics for written english. *Information and control*, 1(4):370–389.

Marcelo A Montemurro and Damián H Zanette. 2010. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153.

Miguel Ortuño, Pedro Carpena, Pedro Bernaola-Galván, Enrique Munoz, and Andrés M Somoza. 2002. Keyword detection in natural languages and dna. *EPL (Europhysics Letters)*, 57(5):759.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20.

Sifatullah Siddiqi and Aditi Sharan. 2015. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, 109(2).

Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos. 2007. Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280.

I.-Y.; Song, M.; Song and X. Hu. 2003. Kpspotter: a flexible information gain-based keyphrase extraction system. In *In Proceedings of WIDM2003*.

Leo Tolstoy. 2013. *Anna karenina*. Mermaids Classic. Translated by C. Garnett.

P. D Turney. 2003. Coherent keyphrase extraction via web mining. In *In Proceedings of IJCAI-0*, pages 434–439.

Mark Twain. 2003. *Adventures of Huckleberry Finn*. University of California Press, Berkely.

Xiaojun Wan and Jianguo Xiao. 2008. Collabrank: towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969–976.

Paynter-G. W. Frank E. Gutwin C. Witten, I. H. and C. G. Nevill-Manning. 1999. Kea: Practical automatic keyphrase extraction. In *In Proceedings of the Fourth ACM Conference on Digital Libraries*, pages 254–255.

Victor Yngve. 1956. Gap analysis and syntax. *IRE Transactions on information theory*, 2(3):106–112.

Hongding Zhou and Gary W Slater. 2003. A metric to search for relevant words. *Physica A: Statistical Mechanics and its Applications*, 329(1-2):309–327.

A Spatial frequency versus ordinary frequency

Here we discuss two features of space-frequency $\tau(w)$ of a word w [see (1)], and the ordinary frequency $f(w)$.

1. If a word w is distributed homogeneously, then $\tau(w)$ defined via (1) is expressed via the ordinary frequency $f(w)$. If in addition, this is a sufficiently frequent word, then $\tau(w) \approx f(w) = \ell/N$, where we assume that $N \gg 1$ and $\ell \gg 1$. Indeed, for the homogeneous distribution of w within the text all ζ_i are equal: $\zeta_i = \zeta$, where ζ is defined from placing the word w among N words (placing $Nf(w)$ times with equal intervals). Hence $Nf(w)+(Nf(w)+1)\zeta = N$ and $\tau(w) = \frac{1}{\zeta+1} = \frac{\frac{1}{N}+f(w)}{\frac{1}{N}+1}$. Whenever $f(w) \gg \frac{1}{N}$ (and naturally $1 \gg \frac{1}{N}$) we get $\tau(w) = f(w)$, i.e. the space frequency coincides with the ordinary one. It is seen that the largest value $\tau(w) = 1$ is achieved for $\zeta_i = 0$ when all appearances of the word w come after each other without any other word in between. The smallest value of $\tau(w) = \frac{1}{N-1}$ is achieved for $\zeta_1 = N - 2$ with just two appearances of w that come as the first and last words of the text.

2. In all texts we studied we noted the following relation

$$\tau(w) > f(w), \tag{9}$$

that holds ~ 80 % of text words w. This set includes frequent words. We validated the following explanation for (9). After (1) we indicated that $\tau(w)$ stays invariant with respect to a certain class of permutations of words in the text. Hence, aiming to calculate $\tau(w)$ for a given frequent word w we can employ the Bernoulli process of text generation, assuming that each word is generated independently from others, and equals w not (w) with probability $f(w)$ $(1 - f(w))$. For spatial intervals s between the occurrences of w the Bernoulli process produces the geometric distribution $p(s) = (1 - f)^s f$. Now the mean of this distribution is $f\sum_{s=0}^{\infty} s(1 - f)^s = \frac{(1-f)}{f}$, whose inverse $\tau(w) \simeq f(w)/(1 - f(w))$ holds (9).