

Assignment 1. REPORT

1. The program

1.2 The code

Choice of language

The assignment is coded in Perl.

Major reasons for this choice were:

- Perl is often used for NLP¹
- Desire to learn a new language (I have never programmed in Perl before and having an opportunity to write a small test program in this language was attractive. Examples in Brill's tutorial gave a good starting point for the coding.)

Data Structures

The program uses an array of hashes as the major data structure. Hash keys are word and frequency. There are two arrays: one stores unsorted data; second array stores the list of words sorted by frequency. Index in the second array corresponds to the rank of the word in the frequency list. Another array is used for intermediate storage of the frequencies of frequencies.

Running the program

The script runs under UNIX/LINUX. File name is `program_a1`. Please do `chmod 755 program_a1` and then `program_a1` at the prompt.

Since multiple tests with different sources have been conducted for this assignment it is necessary to replace input file name (line 16) to replicate them. The list of file names is given below in part 2.2 (description of the corpus) as well as in the program itself right before the line where filenames have to be entered.

Please note that while the program runs unexpectedly slow of inputs bigger than 100K (400K novel takes about 50 min).

Output was sent to three kinds of files: complete list of sorted words went to file `out_*_full` (where * has to be replaced by the reference to the input file, e.g. `out_novel_full`), and a list of words by 100 at every 1000 (files called `out_*_ranges`). Third file `freq_of_freq_*` contains the frequencies of frequencies.

1.2 Assumptions

The program was written under the following assumptions:

A word is "a string of contiguous alphanumeric characters with white spaces on either side; may include hyphens and apostrophes, but no other punctuation marks" (Kučera & Francis, 1967). In the report below this kind of word will be referred to as word-form as opposed to lexemes.

Consequences of this approach:

¹ There is an entire book just for that: M. Hammond: *Programming for Linguists: Perl for Language Researchers*, Blackwell, 2003.

1. Punctuation has been deleted. The following graphemes were considered as punctuation: ! ? . (as well as .. and ...) ; : --; “ , () [] *also ‘ when it was combined with a blank (in some text single quotes have the same function as double quotes). \$ % + = and similar signs were not considered punctuation and were kept.
2. No lemmatization has been performed, each variation of the same lexeme was considered as a separate entity (e.g., compound and compounds; does and doesn't, etc.). This approach may have some impact on the results of the processing (e.g., in the article on chemistry *cell* and *cells* are both very frequent and have consequent ranks of 8 and 9, have they been combined into one word this word would move up to rank 5, same holds for key terminology in all scientific articles, e.g. school/schools in education, user/users in HCI would rank in top 5 words if various forms have been combined). Another setback of such knowledge poor approach would be the increased amount of low-frequency words that will include not only *hapax legomena* but all sorts of graphic, orthographic and morphologic variations of more frequent words (e.g., p'fessor). In the same time the amount of effort that proper lemmatization would require is not justified by the impact of “noise” that is created by this approach. In general, Zipf's law is expected to hold not only for lexemes but for word forms as well.
3. Numbers and mathematical signs (e.g., = +) were considered words because (a) it was consistent with the definition of word given above (b) in scientific articles numbers play an important role.

Other pre-formatting issues:

The entire text has been converted into lowercase to facilitate processing.

Since some texts were in Windows format end of line signs had to be removed

Possessive 's and 's that replaces is in contractions were removed from text because they leave the form of the word unchanged (unlike *n't*, for example in *won't=will not*), in the same time keeping them does not add any relevant information.

2. The experiments and hypotheses

2.1 The hypotheses

Zipf's law² claims that “the frequency of all tokens belonging to to given type w is roughly inversely proportional to rank $r(w)$ of the type, $f(w) \approx const / r(w)$. Rank $r(w)$ is defined as the ordinal number of w on the list of all empirical types sorted in descending order according to $f(w)$ ”³.

Since Zipf's law does not hold very well for words with highest and lowest frequency there have been numerous attempts to improve it (the best known is the Mandelbrot's formula).

Given that Zipf's law is a result of observations and does not have any plausible theoretical explanation experimenting with different kinds of texts and observing the results seems to be the best strategy. The following hypotheses will be tested:

- It is usually considered that Zipf's law holds for different text genres. Texts from different genres – fiction, scientific articles, poetry, and newspapers will be analyzed to prove/disprove that;

² G. Zipf. *The psycho-biology of language: an introduction to dynamic philology*. Houghton Mifflin Company, Boston, MA, 1935

³ There exist multiple renderings of the Zipf's law. This wording is taken from <http://www.ipipan.waw.pl/~ldebowski/czasopisma/glottometrics2002.pdf>

- Size of the text can have an impact on the frequency/rank distribution, especially at the “ends” of the frequency spectrum (most/least frequent words). It would be interesting to see if there are observable differences between texts of small and medium size. The hypothesis is that medium-size texts are the closest to the Zipf’s formula.
- Scientific articles from different disciplines can significantly differ in style, vocabulary, composition, etc. For example, a text in the “soft” discipline would normally use more diversified lexicon and less repetitive terminology, while in “hard” domains the precision of descriptions is more important and therefore repetitive use of the terminology specific for the area and topic would lead to more “thick” upper end (more words with high frequency) and fewer words with frequency 1. Texts from different domains should be analyzed to check this hypothesis.
- Poetry is a genre where author’s style plays the most important role. Do these particularities influence the frequency distribution of the words in the texts?

The set of hypotheses defined the choice of the texts for analysis.

2.2 The corpus

All texts used in the experiments were selected from available Internet sources based on the following criteria:

1. The collection should include texts from different genres:
 1. Fiction
 2. Newspaper articles (a collection from a single newspaper issue)
 3. Research papers
 4. Poetry
2. Texts of different genres should be of comparable size (due to some considerations such as typical size of a scientific article, speed of processing, and storage limitations 35-50K was selected for most texts).
3. To test hypothesis related to the text size within two genres – fiction and scientific articles – texts of different size were considered. To ensure that author/topic related particularities do not influence the results work by the same author was analyzed in both cases.
4. Articles of approximately the same size belonging to different disciplines: software engineering, education, and chemistry were taken.
5. Two same-size collections of poems of two different authors were compiled. I tried to select authors with considerably different individual styles.

All texts were within genre selected randomly based uniquely on their availability and size in order to avoid any bias in the experiments.

Table 1 (next page) gives a complete list of texts used for the experiments:

File	Size	Source	Genre/topic
burns.txt	42K	Collection of poems by Robert Burns from Poetry Archive @ http://www.emule.com/poetry/?page=author_list	Poetry
wilde.txt	48K	Collection of poems by Oscar Wilde from Poetry Archive	Poetry
novel	410K	A novel "Sara, a Princess" by Fannie E.Newberry @ http://ibiblio.org/gutenberg/etext04/srprn10.txt	Fiction Large text
ch1_2	42K	First two chapters of the novel above	Fiction Medium-size text
Chem4..txt	42K	Doris Marko & al. Maillard Reaction Products Modulating the Growth of Human Tumor Cells in Vitro. In Chem. Res. Toxicol., 16 (1), 48 -55, 2003. @ http://0-pubs.acs.org/mercury.concordia.ca/cgi-bin/article.cgi/crtoec/2003/16/i01/html/tx025531a.html	Scientific article Chemistry
educ.txt	36K	Harris, Richard; Mercier, Michael. A test for geographers: the geography of educational achievement in Toronto and Hamilton, 1997 In <i>Canadian Geographer</i> v.44(3) Fall'00 pg 210-227.	Scientific article Education
gazette.txt	38K	A collection of editorials and columns from Montreal Gazette, February 03, 2003 Gazette's web-site: http://www.staging.canada.com/montreal/montreal_gazette/	Newspaper articles
Nielsen.txt Output: out_N1_*	7.5K	J.Nielsen Why You Only Need to Test With 5 Users. @ http://www.useit.com/alertbox/	Scientific article HCI small
Nielsen_94.txt Output: Out_N2_*	53K	J.Nielsen Guerrilla HCI: Using Discount Usability Engineering to Penetrate the Intimidation Barrier @ http://www.useit.com/jakob/	Scientific article HCI Medium-size
Nielsen_93.txt Output: Out_N3_*	120K	J.Nielsen Noncommand User Interfaces 1993. revised version of a paper that appeared in the <i>Communications of the ACM</i> 36 , 4 (April 1993), 83-99 @ http://www.useit.com/jakob/	Scientific article HCI Large
Nielsen_all.txt	180K	Three articles by J.Nielsen combined	Scientific articles HCI collection

Some manual pre-formatting was applied:

1. Novel.txt : header before the novel that starts with the e-book information and the table of contents were removed
2. Scientific articles: pictures were removed, tables were converted into text
3. Poetry : in both files author's name after each poem title was removed

3. The results

The experiments showed that most of the texts follow Zipf's law for largest part of the words, but have different patterns for most and least frequent words.

Some observations and conclusions:

- i. In general, it is obvious that most lexica of most tests roughly follow the Zipf's law, especially in the middle of the frequency spectrum.
- ii. Size of the text can influence the shape of the frequency-rank dependency – the graphs gets closer to the predicted values when the text size increases. Worst results were observed for very small text (7.5K), after 50K results become significantly closer to expected, and after 100K texts seem to follow Zipf's law quite closely.
- iii. In scientific articles key terminology pertinent to the topic is among most frequent words. This is different from all other genres where articles, conjunction, auxiliary verbs and prepositions top the list. All research publications of the same size showed the same tendency independently of their topic. The hypothesis about difference in frequency/rank distribution between “soft” and “hard” disciplines did not hold.
- iv. The behavior of the frequencies in the collection of newspaper articles was not very consistent with the expected. Most likely this is the result of the heterogeneous character of this collection – text belong to different columnists and treat different subjects. The analysis of the list of words shows that non-auxiliary words start to appear only from rank 33, while in other, more homogenous texts of the same size they appear somewhere between ranks 5 and 10. It shows that at least for medium size texts the heterogeneity is a factor.
- v. Both samples of poetry have similar frequency-rank distribution except that work by Robert Burns has a longer trail of words of frequency 1 and 2. To some extent it might be related to the fact that Burn used significant amount of dialect variations of words including rendering phonetic particularities in writing.
- vi. Analysis of the frequencies of frequencies (see attached page for some examples) shows that these values have similar distribution for most of texts, but the data is too scattered to make more important conclusions about this distribution.