Bigrams frequency counts using C++ threads

Filippo Mameli
filippo.mameli@stud.unifi.it

Abstract

Un bigramma (o digramma) è un gruppo di due lettere o parole consecutive. L'analisi della frequenza dei bigrammi è comunemente usata nella crittoanalisi o nell'apprendimento automatico. In questo documento si presenta un programma scritto in C++ che conta le occorrenze di bigrammi in un testo. Si mostrano inoltre le differenze di prestazioni tra il programma nella versione parallela e quella sequenziale. I principali strumenti utilizzati sono la classe std::Thread e la libreria Boost.

Permessi di distribuzione

L'autore di questa relazione permette che questo documento possa essere distribuito a tutti gli studenti UNIFI dei corsi futuri.

1. Introduzione

Un bigramma è una sequenza di due lettere o parole addiacenti. I bigrammi sono utilizzati in vari ambiti. L'analisi statistica del testo con i bigrammi è usata ad esempio in crittoanalisi, in apprendimento automatico o nel riconoscimento vocale. Nella relazione si descrive un programma parallelo scritto in C++ che conta le frequenze dei bigrammi di lettere in un testo. Nella sezione [bla] si mostrano inoltre le differenze prestazionali tra il programma parallelo e la sua controparte sequenziale, mostrando lo Speedup e i tempi di esecuzione tra differenti documenti di testo. Nel programma si utilizzano principalmente la classe della libreria standard Thread e la libreria Boost.

2. Algoritmo di base

Il programma deve parsare ogni singola parola e aggiornare il contatore dei bigrammi trovati. La struttura dati per contare le occorrenze è l'Hashtable. Questa è stata scelta perché l'operazione più usata è la ricerca del bigramma trovato e il consecutivo incremento del contatore. In un Hashtable l'operazione di ricerca ha complessità di tempo pari a $O(1)$ nel caso medio e per questa qualità, la struttura è adatta per l'aggiornamento delle occorrenze. In Algorithm 1 vi è descritto l'argoritmo in speudocodice.

```
Data: File di testo
Result: Frequenza dei bigrammi contati
lettura del file di testo;
while fine del file non raggiunto do
    for per ogni parola in file do
        for per ogni bigramma nella parola do
            Hashtable[bigramma]++;
        end
    end
end
```

Algorithm 1: Algoritmo di base

2.1. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler.

2.2. Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn't refer to it in the text doesn't mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: http://www.cvpr.org/doc/mermin.pdf.

2.3. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TeXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [?], and subsequently developed by Alpher and Fotheringham-Smythe [?], and Alpher *et al.* [?]."

This is incorrect: "... subsequently developed by Alpher *et al.* [?] ..." because reference [?] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al.*

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [?, ?, ?] to [?, ?, ?].

3. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

3.1. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high.

3.2. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please

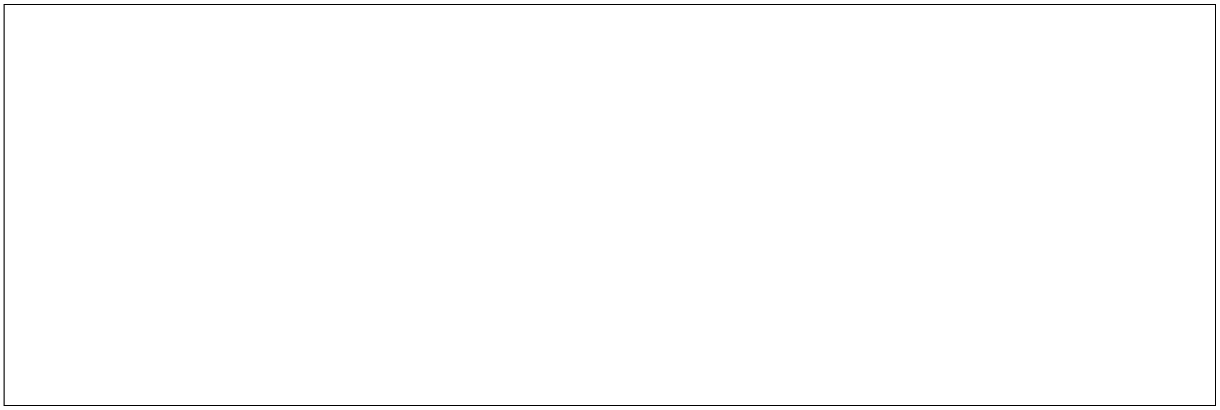

Figura 1. Example of a short caption, which should be centered.

do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures **??** and 1. Short captions should be centred.

Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

3.3. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Tabella 1. Results. Ours is better.

3.4. References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [**?**]. Where appropriate, include the name(s) of editors of referenced books.

3.5. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

3.6. Color

Color is valuable, and will be visible to readers of the electronic copy. However ensure that, when printed on a monochrome printer, no important information is lost by the conversion to grayscale.

4. Appendix

If your course project is part of a larger project from another class or research lab, please fill in this section and clearly spell out the following items:

1. Explicitly explain what the computer vision components are in this course project;
2. Explicitly list out all of your own contributions in this project in terms of:
   (a) ideas
   (b) formulations of algorithms
   (c) software and coding
   (d) designs of experiments
   (e) analysis of experiments
3. Verify and confirm that you (and your partner currently taking CS231A) are the sole author(s) of the writeup. Please provide papers, theses, or other documents related to this project so that we can compare with your own writeup.