A Sample *Proceedings of the VLDB Endowment* Paper in LaTeX Format[*]

Ben Trovato[†]
Institute for Clarity in Documentation
1932 Wallamaloo Lane
Wallamaloo, New Zealand
trovato@corporation.com

G.K.M. Tobin[‡]
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld[§]
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

Lawrence P. Leipuner
Brookhaven Laboratories
Brookhaven National Lab
P.O. Box 5000
lleipuner@researchlabs.org

Sean Fogarty
NASA Ames Research Center
Moffett Field
California 94035
fogartys@amesres.org

Charles Palmer
Palmer Research Laboratories
8600 Datapoint Drive
San Antonio, Texas 78229
cpalmer@prl.com

ABSTRACT

The abstract for your paper for the PVLDB Journal submission. The template and the example document are based on the ACM SIG Proceedings templates. This file is part of a package for preparing the submissions for review. These files are in the camera-ready format, but they do not contain the full copyright note. Note that after the notification of acceptance, there will be an updated style file for the camera-ready submission containing the copyright note.

1. INTRODUCTION

The *proceedings* are the records of a conference. ACM, as well as PVLDB, seeks to give these conference by-products a uniform, high-quality appearance. To do this, ACM / PVLDB has some rigid requirements for the format of the proceedings documents: there is a specified format (balanced double columns), a specified set of fonts (Arial or Helvetica and Times Roman) in certain specified sizes (for instance, 9 point for body copy), a specified live area (18 × 23.5 cm [7" × 9.25"]) centered on the page, specified size of margins (2.54cm [1"] top and bottom and 1.9cm [.75"] left and right; specified column width (8.45cm [3.33"]) and gutter size (.083cm [.33"]).

The good news is, with only a handful of manual settings[1], the LaTeX document class file handles all of this for you.

The remainder of this document is concerned with showing, in the context of an "actual" document, the LaTeX commands specifically available for denoting the structure of a proceedings paper, rather than with giving rigorous descriptions or explanations of such commands.

2. THE *BODY* OF THE PAPER

Typically, the body of a paper is organized into a hierarchical structure, with numbered or unnumbered headings for sections, subsections, sub-subsections, and even smaller sections. The command `\section` that precedes this paragraph is part of such a hierarchy.[2] LaTeX handles the numbering and placement of these headings for you, when you use the appropriate heading commands around the titles of the headings. If you want a sub-subsection or smaller part to be unnumbered in your output, simply append an asterisk to the command name. Examples of both numbered and unnumbered headings will appear throughout the balance of this sample document.

Because the entire article is contained in the **document** environment, you can indicate the start of a new paragraph with a blank line in your input file; that is why this sentence forms a separate paragraph.

3. PROBLEM

The problem we are dealing with is the fuzzy join problem. EXPAND

[*] for use with vldb.cls

[†] Dr. Trovato insisted his name be first.

[‡] The secretary disavows any knowledge of this author's actions.

[§] This author is the one who did all the really hard work.

[1] Two of these, the `\numberofauthors` and `\alignauthor` commands, you have already used; another, `\balancecolumns`, will be used in your very last run of LaTeX to ensure balanced column heights on the last page.

[2] This is the second footnote. It starts a series of three footnotes that add nothing informational, but just give an idea of how footnotes work and look. It is a wordy one, just so you see how a longish one plays out.

Table 1: Table 1

Name	Match
Douglas Adams	Douglas Noel Adams
Andreas Capellanus	Andrea Cappellano
John Adams Whipple	John A. Whipple

4. ALGORITHM

We started our process with a simple rule based method for matching and another one for blocking. We then replaced individual peices with a machine learning method and compared the preformance. In the final product we have an elegant completly machine learning based method.

4.1 Rule-based Matching

In the rule-based method, we block based on items that have one word in common. This is the simplest reasonable blocking strategy. Table 2 illustrates what the blocking of the three names from Table 1 would look like. Out of NUMBER1 pairs, NUMBER2 ended up in the same bucket as all their correct matches. NUMBER3 had at least one correct match in the same bucket. To perform the actual matching, we used three simple rules see Table 3:

1. If a word is unique to two items, we match them. (‘John Adams Whipple’ matches ‘John A. Whipple’ since ‘Whipple’ only apears in those two names)
2. If, when all spaces are removed, one of the items is a substring of the other, we match them. (‘the flower company’ matches ‘theflowercompany.com’ since ‘theflowercompany’ is a subset of ‘theflowercompany.com’)
3. If we treat each name as a set of words, and one set is a subset of the other set, we match them. (‘Douglas Adams’ matches ‘Douglas Noel Adams’ since $\{$‘Douglas’, ‘Adams’$\} \subseteq \{$‘Douglas’, ‘Noel’, ‘Adams’$\}$)

Even using these three rules in combination resulted in too many false negatives. Therefore we decided to treat all matches resulting from any of the three rules as valid. If we treat it as a success if there is a correct match in any of the first three slots, we get an F_1 of FSCORE1. Since the issue is mostly false negatives, and therefore even some of the first three slots were often empty, so even if we accept any of the top 1000 as a success, we get FSCORE2, not much better.

4.2 Siamese Netwok

The first part we replaced was the matcher. To compare the two matchers we created a rule-based matcher using the rules from the rule-based system. To replace it we created a siamese network (figure 1) INSERT FIGURE 1. A Siamese network is two deep neural networks that share weights. One entity is fed into each of the two networks. The output of the two networks is fed into a final layer which determines the distance between the two outputs and accepts or rejects. Siamese networks have been shown to work well for image matching.[4] Before feeding the entities into the networks, we used Kazuma charecter embeddings to encode each entity as a vector. We trained the network on our NUMBER1*0.95 pairs and withheld NUMBER1*0.05 for testing. We chose an equal number of negitive pairs to train the model on. We chose negitive pairs that had at least one word in common with the positive ones. We found that the fscore on the training data was FSCORE3 and FSCORE4 on the test date. This was better than the rule-based matcher which had an f-score of FSCORE5.

We next approached the blocking problem with machine learning. Since the hidden layer of the siamese network is optimized for matching, it should output an embedding of the entity which has its essential qualities with regaurds to matching (figure 2). We then used an aproximate nearest neighbors algorithm to find the nearest neighbors. EXPLAIN MORE Using these techniques we can find the most similar items without needing to block at all. The problem we ran into is only PERCENT1 of the correct matches were in the 10 nearest neighbors. The problem seems to be that since we trained the model using negative pairs that have at least one word in common, it did not learn that names that are completly different should be mapped seperatly (figure 3). We could try to fix this by feeding the model better negative pairs, but we found a more elegant solution. Instead of trying to aproximate the correct location using a siamese network trained for matching, we can use a triplet loss function and teach it to maximize the distance to the closest false pair and minimize the distance to the true matches. To find these closest entities we use our charecter embedings. This allowed us to get an f-score of FSCORE6 on the training set and FSCORE7 on the test set reading the 10 closest entities. Notice that at this point we can do away with the matching algorithm alltogether. We simply use the closest entiies as our matches.

5. RELATED WORKS

There have been a number of attemts to solve the fuzzy join problem. Many of the attemts use varios string matching algorithms for example [9]. The main issue with using string matching comparisons is that they do not work across a wide variaty of enties ASK DR. SRINIVAS. To our knowledge no serios machine learning aproach has been used for this problem. The problem of blocking has also been dealt with using MapReduce [8]. This approach requires alot of computing power and, since they use a string matching approach in the end, is hard to generalize accross many types of entities.

5.1 Type Changes and *Special* Characters

We have already seen several typeface changes in this sample. You can indicate italicized words or phrases in your text with the command `\textit`; emboldening with the command `\textbf` and typewriter-style (for instance, for computer code) with `\texttt`. But remember, you do not have to indicate typestyle changes when such changes are part of the *structural* elements of your article; for instance, the heading of this subsection will be in a sans serif[3] typeface, but that is handled by the document class file. Take care with the use of[4] the curly braces in typeface changes; they mark the beginning and end of the text that is to be in the different typeface.

[3] A third footnote, here. Let’s make this a rather short one to see how it looks.

[4] A fourth, and last, footnote.

Table 2: Table 2

Bucket Name	Contents
A	{John A. Whipple}
Adams	{Douglas Adams, Douglas Noel Adams, John Adams Whipple}
Andreas	{Andreas Capellanus}
Andrea	{Andrea Cappellano}
Cappellano	{Andrea Cappellano}
Capellanus	{Andreas Capellanus}
Douglas	{Douglas Adams, Douglas Noel Adams}
John	{John Adams Whipple, John A. Whipple}
Noel	{Douglas Noel Adams}
Whipple	{John Adams Whipple, John A. Whipple}

You can use whatever symbols, accented characters, or non-English characters you need anywhere in your document; you can find a complete list of what is available in the *LaTeX User's Guide*[6].

5.2 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

5.2.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin. . .\end` construction or with the short form `$. . .$`. You can use any of the symbols and structures, from α to ω, available in LaTeX[6]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\rightarrow \infty}x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

5.2.2 Display Equations

A numbered display equation – one set off by vertical space from the text and centered horizontally – is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\rightarrow \infty}x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty}x_i=\int_{0}^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

Table 3: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ^2_1	1 in 40,000	Unexplained usage

5.3 Citations

Citations to articles [1, 3, 2, 5], conference proceedings [3] or books [7, 6] listed in the Bibliography section of your article will occur throughout the text of your article. You should use BibTeX to automatically produce this bibliography; you simply need to insert one of several citation commands with a key of the item cited in the proper location in the `.tex` file [6]. The key is a short reference you invent to uniquely identify each work; in this sample document, the key is the first author's surname and a word from the title. This identifying key is included with each item in the `.bib` file for your article.

The details of the construction of the `.bib` file are beyond the scope of this sample document, but more information can be found in the *Author's Guide*, and exhaustive details in the *LaTeX User's Guide*[6].

This article shows only the plainest form of the citation command, using `\cite`. This is what is stipulated in the SIGS style specifications. No other citation format is endorsed.

5.4 Tables

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material is found in the *LaTeX User's Guide*.

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed dvi output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this

Figure 1: A sample black and white graphic (.pdf format).

Figure 2: A sample black and white graphic (.pdf format) that has been resized with the `includegraphics` command.

sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed dvi output of this document.

5.5 Figures

Like tables, figures cannot be split across pages; the best placement for them is typically the top or the bottom of the page nearest their initial cite. To ensure this proper "floating" placement of figures, use the environment **figure** to enclose the figure and its caption.

This sample document contains examples of **.pdf** files to be displayable with LaTeX. More details on each of these is found in the *Author's Guide.*

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption.

Note that only **.pdf** files were used; if you want to include **.ps** or **.eps** formats, you can use the `\epsfig` or `\psfig` commands as appropriate for the different file types.

5.6 Theorem-like Constructs

Other common constructs that may occur in your article are the forms for logical constructs like theorems, axioms, corollaries and proofs. There are two forms, one produced by the command `\newtheorem` and the other by the command `\newdef`; perhaps the clearest and easiest way to distinguish them is to compare the two in the output of this sample document:

This uses the **theorem** environment, created by the `\newtheorem` command:

Theorem 1. *Let f be continuous on $[a,b]$. If G is an antiderivative for f on $[a,b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

The other uses the **definition** environment, created by the `\newdef` command:

Definition 1. If z is irrational, then by e^z we mean the unique number which has logarithm z:

$$\log e^z = z$$

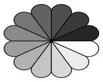

Figure 3: A sample black and white graphic (.pdf format) that has been resized with the `includegraphics` command.

Two lists of constructs that use one of these forms is given in the *Author's Guidelines.*

and don't forget to end the environment with figure*, not figure!

There is one other similar construct environment, which is already set up for you; i.e. you must *not* use a `\newdef` command to create it: the **proof** environment. Here is a example of its use:

Proof. Suppose on the contrary there exists a real number L such that

$$\lim_{x\to\infty}\frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c}\left[gx \cdot \frac{f(x)}{g(x)}\right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c}\frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □

Complete rules about using these environments and using the two different creation commands are in the *Author's Guide*; please consult it for more detailed instructions. If you need to use another construct, not listed therein, which you want to have the same formatting as the Theorem or the Definition[7] shown above, use the `\newtheorem` or the `\newdef` command, respectively, to create it.

A *Caveat* for the TeX Expert

Because you have just been given permission to use the `\newdef` command to create a new form, you might think you can use TeX's `\def` to create a new command: *Please refrain from doing this!* Remember that your LaTeX source code is primarily intended to create camera-ready copy, but may be converted to other forms – e.g. HTML. If you inadvertently omit some or all of the `\def`s recompilation will be, to say the least, problematic.

6. CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

7. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you.

Table 4: Some Typical Commands

Command	A Number	Comments
`\alignauthor`	100	Author alignment
`\numberofauthors`	200	Author enumeration
`\table`	300	For tables
`\table*`	400	For wider tables

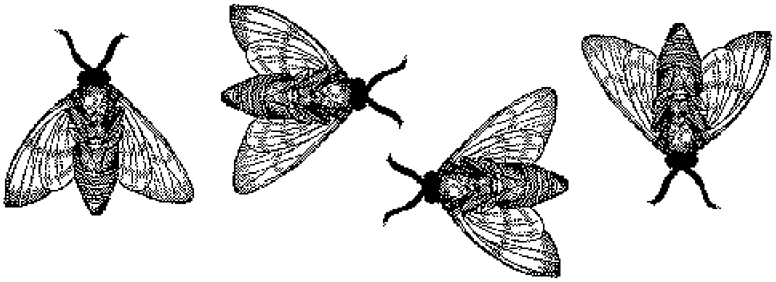

Figure 4: A sample black and white graphic (.pdf format) that needs to span two columns of text.

In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

8. ADDITIONAL AUTHORS

Additional authors: John Smith (The Thørväld Group, email: `jsmith@affiliation.org`) and Julius P. Kumquat (The Kumquat Consortium, email: `jpkumquat@consortium.net`)

9. REFERENCES

[1] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[2] J. Braams. Babel, a multilingual style-option system for use with latex's standard document styles. *TUGboat*, 12(2):291–301, June 1991.

[3] M. Clark. Post congress tristesse. In *TeX90 Conference Proceedings*, pages 84–89. TeX Users Group, March 1991.

[4] R. Hadsell, S. Chopra, and Y. Lecun. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR06).*

[5] M. Herlihy. A methodology for implementing highly concurrent data objects. *ACM Trans. Program. Lang. Syst.*, 15(5):745–770, November 1993.

[6] L. Lamport. *LaTeX User's Guide and Document Reference Manual.* Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.

[7] S. Salas and E. Hille. *Calculus: One and Several Variable.* John Wiley and Sons, New York, 1978.

[8] R. Vernica, M. J. Carey, and C. Li. Efficient parallel set-similarity joins using mapreduce. *Proceedings of the 2010 international conference on Management of data - SIGMOD 10*, 2010.

[9] J. Wang, G. Li, and J. Fe. Fast-join: An efficient method for fuzzy token matching based string similarity join. *2011 IEEE 27th International Conference on Data Engineering*, 2011.

9.1 References

Generated by bibtex from your .bib file. Run latex, then bibtex, then latex twice (to resolve references) to create the .bbl file. Insert that .bbl file into the .tex source file and comment out the command `\thebibliography`.

APPENDIX

A. PVLDB FORMAT HAS 8+4 PAGES

The PVLDB paper length is limited to 8 pages. You are permitted a 4 page appendix beyond these 8 pages. However, reviewers (as well as any readers) are not required to read this appendix, and the paper should be self-contained, complete and understandable within the 8 pages. Typically, it is appropriate to place proofs, algorithm pseudocode, data set descriptions, etc. in the appendix.

Any references to the appendix from the main paper should only be in the nature of ”for additional detail see..”. In particular, there should be nothing in the appendix that is necessary for a reader to understand the paper. This 8+4 page rule applies to both submissions and camera-ready.

B. FINAL THOUGHTS ON GOOD LAYOUT

Please restrain yourself from squeezing too much information into the first eight pages; you can use the appendix for optional proofs or details of your evaluation which are not absolutely necessary to the core understanding of your paper. This way, you can use readable font sizes in the figures and graphs, as well as avoid tempering with the correct border values, and the spacing (and format) of both text and captions of the PVLDB format (e.g. captions are bold).

At the end, please check for an overall pleasant layout, e.g. by ensuring a readable and logical positioning of any floating figures and tables. Please also check for any line overflows, which are only allowed in extraordinary circumstances (such as wide formulas or URLs where a line wrap would be counterintuitive).

Use the `balance` package together with a `\balance` command at the end of your document to ensure that the last page has balanced (i.e. same length) columns.