Projects on the Move

Who says statistics have to be a nightmare? Also in this issue: After the Deadline takes care of language issues.

By Carsten Schnober and Heike Jurzik

Even hardened nerds are often over-challenged by the less than intuitive field of statistics. Besides the theory, you need to know how to use the software that converts all the theory into a practical application.

The R environment [1] can be a big help on the software side. This free implementation of the S statistics programming language was launched in 1992. Initially, you will not be able to do much with R and its spartan command-line tools until you learn the language.

Although SPSS [2] is a commercial alternative, even students are asked to pay a three-digit licensing fee, and the quality isn't always on par with R.

Sofa Statistics for All

Free statistics software with an intuitive web interface is a niche that the Sofa (Statistics Open For All) [3] project fills. One of the project's aims is to give statisticians an easy-to-use tool. Like many other mathematical disciplines, statistics is an ancillary science that is used in many areas to interpret empiric surveys. For example, you can use a series of results to calculate the probability of dice throws. Social scientists can also forecast whether a survey is significant or will just output random values. Many statistical tests have different strengths and weaknesses that only specialists can assess.

Statistical tests are generally regarded as essential because human intuition is prone to misjudgment. The Monte-Carlo fallacy is just one example that leads to an incorrect assumption of coherencies and the resulting shift of probabilities between two unrelated events – for example, between the result of the next throw of the dice and the previous results with the same dice.

To cope with these and even more complex problems, Sofa (Figure 1) supports standards such as ANOVA (analysis of variance), Pearson's chi-square test, *t*-tests, and other methods. The statistics tool also provides basic values like means, medians, standard deviations, sums, maximums, minimums, and more.

Quick Change Artist

As mentioned previously, Sofa's strength is not just its functionality; R and others have all the features listed in the previous section. But Sofa also impresses with its easy accessibility and its import and export options.

Sofa can read data from an SQL database (MySQL, SQLite, and PostgreSQL), from Open Document spreadsheets (as used by OpenOffice and others) and from MS Access. The program presents the results in HTML format and relies on JavaScript to highlight specific rows of data in a targeted way when the user mouses over them, giving users the ability to build statistics directly into a website. Alternatively, Sofa stores its results directly in spreadsheet formats used by OpenOffice Calc and Microsoft Excel.

Users can point and click to define the style and content of their Sofa reports (Figure 2). Also, the software can automate output, via Python to create, for example, a new program in a blog when new data occur.

Functionality and Usability

Sofa relies on the free Dojo toolkit [4] to render the JavaScript-enhanced HTML output. Sofa itself is implemented in Java and thus runs on other platforms besides Linux. The homepage includes prebuilt packages for Ubuntu, Windows, and Mac OS X, as well as the source code.

The roadmap for future versions of the program includes the ability to export to the Oracle database format and other diagram types. Additionally, a plugin architecture is planned to facilitate the integration of further statistics functions.

Sofa developers also continue to pursue their original goal – that of making statistical methods more accessible to newcomers. To this end, the tool will include tips on typical applications and graphical examples in the future. They also plan to expand localization efforts: Right now, Sofa is only available in English and Galician, but more languages are planned. If you are interested in contributing to a translation or promoting the development of Sofa in any other way, you can contact the developers via the project homepage. Grant Paton-Simpson encourages users to report issues and vote through the Freshmeat and SourceForge open source portals.

From Theory to Practice

Automated processing of natural languages is one example of a practical application of sophisticated statistical methods. Whereas the early applications in this field, such as machine translation and text summarization and correction, quickly reached their limits, a new approach has dominated recent research: statistical models. Whereas statisticians used to apply complex rules manually in the language of their choice, encountering many new exceptions and supplements in the process, the current approach is for programs to learn from existing texts on the basis of statistical methods.

The advantage of statistical language models of this kind is that they need virtually no manual preparation. All they require is a large volume of text from which to learn. The disadvantage is that, with no upper limit for the volume of text, a system of this kind will need a fairly large amount of space, a couple of hundred megabytes at least, both on disk and in RAM, which is a big issue, especially on mobile devices.

Therefore, statistics-based grammar and spelling checkers are less well suited to use on the desktop. Who wants to swap their lean client fast enough for word processing for a fat one, just to run a single function?

The alternative is a spell checker that compares words with a dictionary and reports an error if it encounters something unknown. The mistakes it finds are fairly simple, and it will fail to discover errors in sentence structure, stylistic bugs, grammatical errors, and incorrect usage of words. Some commercial word processors do provide ruleset-based grammar checkers, but manual definitions will never be able to cover all the rules that govern human language.

Deadline for Typos

After the Deadline [5] offers an innovative solution to users with broadband In-

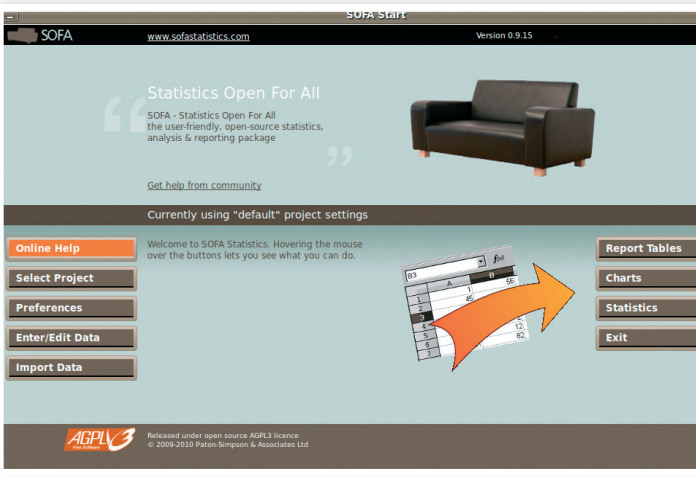

Figure 1: Sofa offers an intuitive user interface that pleases professional statisticians and newcomers alike.

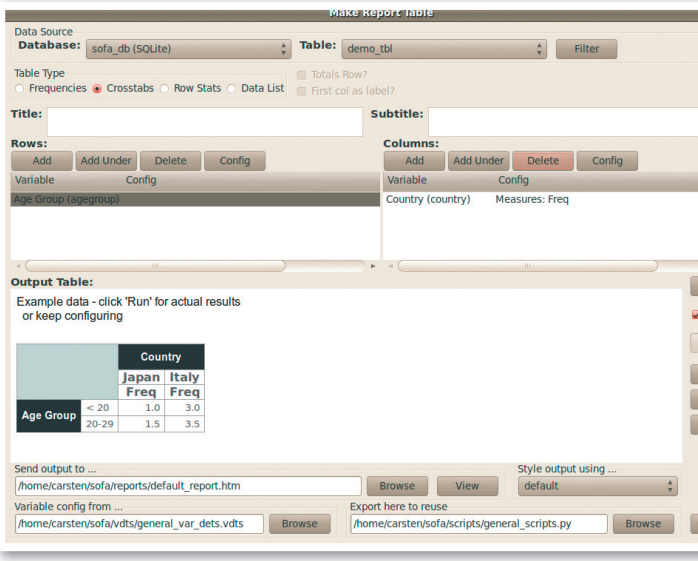

Figure 2: Sofa can create reports in a variety of file formats by simply selecting the data and choosing a presentation format.

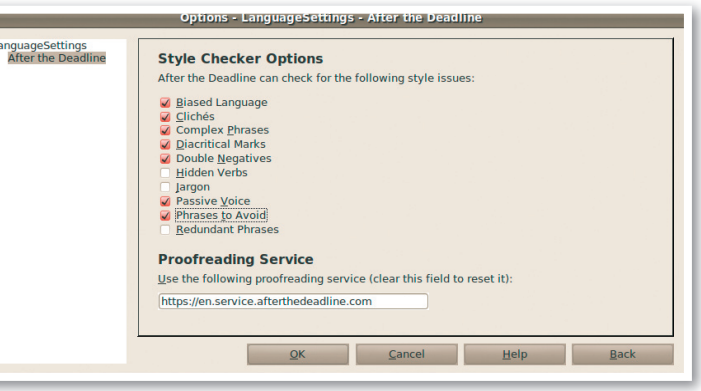

Figure 3: The OpenOffice plugin for After the Deadline supports stylistic corrections in addition to spell checking.

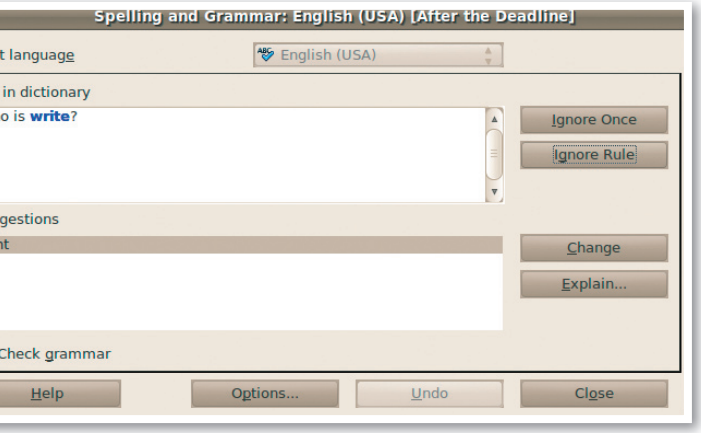

Figure 4: Known words can still be incorrect depending on the context: After the Deadline identifies mistakes that occur because of different words with identical pronunciation.

ternet connections. The free web service accepts texts and returns proofreading proposals. The name comes from the *New York Times* column that investigates stylistic blunders in day-to-day English.

After the Deadline is a statistical language model that weighs in at 1GB and resides in RAM server side, thus reducing the client requirements to something minimal. Users simply need a program to access the online service, and there are plugins to handle this for the OpenOffice word processor, Firefox, Google Chrome browsers, and the WordPress blogging software.

Also, extensions for jQuery and TinyMCE let users build the spell-checker into forms on websites. With open standards and source code, there is nothing to prevent more interfaces being developed in future.

Figure 3 shows the setup options for the OpenOffice extension. After the Deadline offers a long list of optional style checks that supplement the tool's core functionality. The tests will warn you, for example, if you use overly complicated sentence structure, nominalization (*Hidden Verbs*) and passive sentences.

Error Evaluation

The language model that After the Deadline uses is based on bigrams – that is, sequences of two words. Their relative frequency in the training material represents the probability of their occurrence in new text. The developers evaluated the texts of Wikipedia, the Gutenberg project, and various weblogs to support this. The model is supplemented by a trigram-based language model that is used to search for easily confused words, such as homonyns (Figure 4).

The probability values derived by this approach are first used to provide the most likely candidate to replace an unknown word. Besides considering the probability of the word occurring in the given context (i.e., with the preceding and following words), After the Deadline looks at other factors: the number of changes required to create another word from the word with the typo, and the match between the first letter and the context-dependent frequency of a word. A neuronal network helps assess the correction suggestions.

After the Deadline does resort to manual rules for any further checks; to this end, it integrates LanguageTool [6], which was developed in 2003 and is available as an independent OpenOffice plugin. Additionally, After the Deadline can structure a body of text in a meaningful way and generate suggestions for paragraph borders.

Unlimited Options

If you are interested in checking out the program's capabilities, you can test After the Deadline online [7]. The plugins referred to will also cooperate with one of the project's servers, although use is restricted to the English language and must be non-commercial.

Users who are interested in deploying the style and language checker for commercial purposes, or for multiple languages, will find an installation how-to for the server component on the project's website.

The GPL'd software requires Sun's Java version 1.6.0 and 1.5GB of RAM for the low-memory variant or 4GB of RAM for the full version. The 1.6.0 version is useful for short documents and a restricted amount of client access. Besides the software itself, a number of language models are also freely available. After the Deadline speaks not only English, but also Dutch, French, German, Indonesian, Italian, Polish, Portuguese, Spanish, and Russian.

Future versions of the software will be capable of finding omitted words. Also, plans are afoot to let the server automatically extend its language models at regular intervals.

The roadmap also envisages improvements to the plugins and a function that checks text for errors during input. If you want to help, or you have your own ideas, the project website provides details of how to contribute. ■■■

INFO

[1] R Project: *http://www.r-project.org*
[2] SPSS: *http://www.spss.com*
[3] Sofa project: *http://www.sofastatistics.com*
[4] Dojo toolkit: *http://www.dojotoolkit.org*
[5] After the Deadline: *http://afterthedeadline.com*
[6] LanguageTool: *http://www.languagetool.org*
[7] After the Deadline online spell checker: *http://www.polishmywriting.com*