The KrdWrd Add-on

July 8, 2008

1 Introduction

"The availability of large text corpora has changed the scientific approach to language in linguistics and cognitive science" [M&S]. Today, the by far richest source for authentic natural language data is the World Wide Web, and making it useful as a data source for scientific research is imperative.

Web pages, however, can not be used for computational linguistic processing without filtering: They contain code for processing by the Web browser, there are menus, headers footers, form fields, teasers, out-links, spam-text – all of which needs to be stripped.

The dimension of this task calls for an automated solution, the broadness of the problem for machine learning based approaches. Part of the KrdWrd project deals with the development of appropriate methods, but they require hand-annotated pages for training.

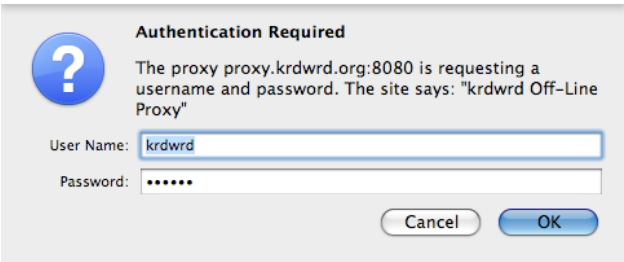
The *KrdWrd Add-on* aims at making this kind of tagging of Web pages possible. For users, we provide accurate Web page presentation and annotation utilities in a typical browsing environment, while preserving the original document and all the additional information contained therein.

2 Getting Started

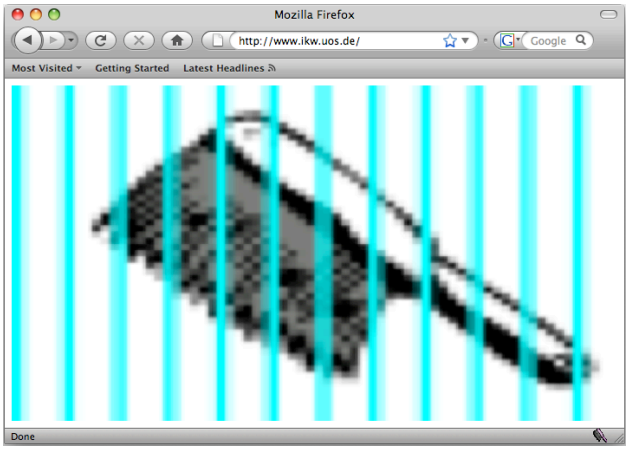
In this section, we will give you information about how to use the tool. If you have not installed it yet, go to krdwrd.org and get it. Of course, you will need Firefox, too.

- Since the add-on depends on a special proxy server to connect to the Internet - you can only grab and submit Web pages from the KrdWrd corpora - it may be a good idea to create a separate profile just for working with the add-on. If you want to create a profile but have no idea how to do that, have a look here.

- When grabbing a page for the first time, or selecting a corpus for the first time you will be asked to authenticate for the krdwrd Off-Line Proxy. The username and password in the dialog box are already filled in and it is save to leave the "Use Password Manager to remember this password" checked.

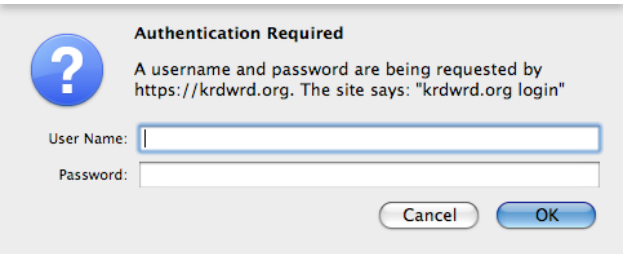


- The proxy server will deny all requests that are not part of the normal add-on operation. If you ever see something like



it is most likely because you tried to surf the Web with the wrong Firefox profile.

- You will be asked for authentication a second time. This authentication is for the KrdWrd Web site and requires your *RZ Account*[1]



- When you request a page from the corpus for the first time, Firefox will popup a security warning. The warning says *"you have requested an encrypted page that contains some unencrypted data"*. The warning is issued because the corpus page are issued unencrypted. Your login credentials are never send to the server unencrypted, there is no reason not to ignore this warning.

2.1 First Steps

- **How to Use the Mouse**
  When moving the mouse over a Web page, you will notice that certain areas are highlighted in pink. These are the blocks of text that you can tag. Sometimes the pink areas are fairly small (single words or lines of text), sometimes they are pretty large (whole paragraphs, or even whole pages). Thus it makes sense to move the mouse around a little before you actually start tagging because sometimes you want to tag big areas as, say, 'bad', and it saves you a lot of time if you do not have to tag every single line or paragraph. As a rule of thumb, it often makes sense to tag *everything* in red ('bad'), from top to bottom, and only then start tagging smaller pieces in yellow or green ('uncertain' or 'good', respectively) (see also **Examples, described on page ??, Tips & Tricks, page ??**).

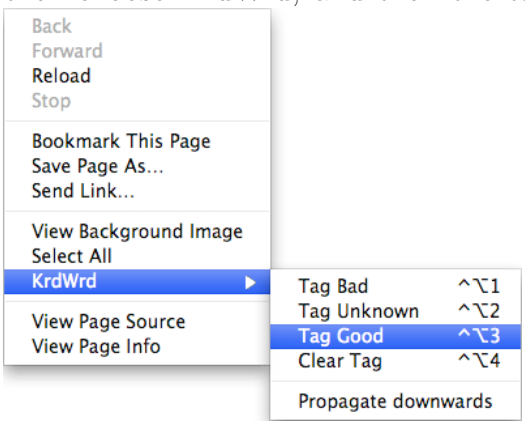
[1] This is the same login as for *Stud.IP* and *WebMail*; in case you want to "Use Password Manger" please also "Use a master password" to protect your sensitive information.

- **How to Choose the Tag**

This section deals with assigning tags. If you want information on how to choose the *right* tag to assign, go to the Annotation Guidlines on page **??**.

For tagging a pink-highlighted section as 'good', 'bad', or 'uncertain', you have two options: You can use (1) keyboard shortcuts (hotkeys) or you can use (2) the context menu (rightclick).

1. Keyboard Shortcuts
   - *bad*: ctrl+alt+1
   - *uncertain*: ctrl+alt+2
   - *good*: ctrl+alt+3
   - *clear annotation*: ctrl+alt+4
2. Context Menu
   - Rightclick when you are over the section you want to tag, then choose KrdWrd, and then the tag you want to assign.



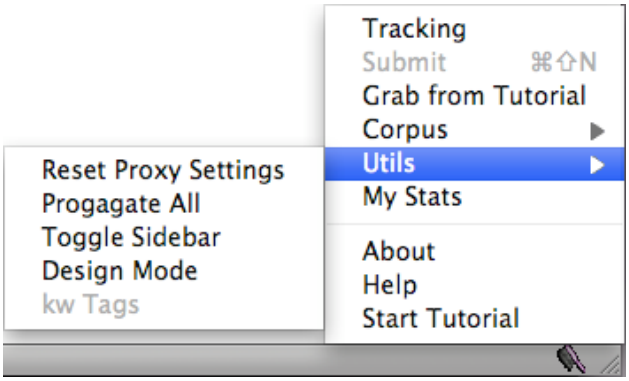
   - Using the context menu is not recommended, however. It is much more time-consuming to navigate the menu than to use the keyboard shortcuts.

– Note also that if the mouse cursor leaves the menu area, a possibly different part of the page will be highlighted for tagging (namely the part that is now 'under' your mouse).

- **Cookies and Certificates**
  When some of the pages are being loaded, your Web browser will ask you whether you want to accept cookies (of course, depending on your browser settings: If you use a separate profile for the KrdWrd add-on, just allow all cookies, see Tips & Tricks, page **??**). Actually, you *do not* have to accept any cookies; however, nothing bad will happen if you do accept them.

2.2 The Statusbar Menu

This section describes the status bar menu, depicted below.



- **Tracking**
  Here you can turn on or off whether sections of the Web page you are currently viewing are highlighted in pink. Usually there is no need to disable tracking. However, in some rare cases, this might help you to get a better view on a page before tagging it.

- **Submit**
  When you are done tagging a page, i.e. when everything on the page is green, red, or yellow, you can submit the page with this menu option - and the next page will load automatically. For your convenience, you

can also use the keyboard shortcut *options+shift+N*.

- **Grab Page**
  Clicking here loads a new, un-annotated page. Once you annotated the whole corpus, you will be redirected to your personal statistics page.

- **Corpus**
  Here you can select one of the (predefined) available corpora. But you should stick to the Canola corpus for now.

- **Utils**
  The options in this menu make your life easier when tagging pages.

  - **Propagate**
    Here you can explicitly propagate a given tag down to all sibling nodes. This is helpful when you have a large portion that should be tagged red but all its siblings should be tagged green. You can then tag the parent node green, propagate, and re-tag the parent node as red. This way you do not need to tag all the siblings separately. (Try this on the Examples, described on page **??** and check the Tips & Tricks, page **??**).
  - **Toggle Sidebar**
    Clicking here opens the sidebar. In the sidebar you can see all of the text in the current page and how it is tagged. A given tag is usually propagated down to lower nodes in the DOM tree automatically, but sometimes it may be unclear (i.e. not directly visible in the page) how a particular portion of text is tagged. In the sidebar you can easily see whether it is tagged red, green, or yellow.
  - **Design Mode**
    This is a debugging feature and you must not use it while tagging pages.

- **My Stats**
  This menu option will send you to your KrdWrd account. There you can see how many pages you have already tagged, and you can view, re-submit and delete your tagged pages.

3 How to Tag Pages

3.1 Annotation Guidelines

In the previous section, we described how to use the tool and how to assign tags. In the following, we give you guidelines regarding *which* tag should be assigned to a particular kind of text.

- Everything that is *boilerplate* is tagged **red**. *Boilerplate* is ...

  1. all navigation information,
  2. copyright information,
  3. hyperlinks that are not part of the text,
  4. all kinds of headers and footers.

  → Generally speaking, boilerplate is everything that can be used interchangeably with any other Web page or could be left out without changing the general content of the page.

- The following types of text are also tagged **red**:

  1. incomplete sentences, or text in telegraphic style,
  2. text containing 'non-words' such as file names,
  3. off-site advertisements (i.e. advertisement from an external page),
  4. text in any other language than English,
  5. lists and enumerations of any kind.

- All captions are tagged **yellow**. And also everything that does not belong in the red or green category is tagged **yellow**.

- All text that is left is tagged **green**, i.e. ...

1. text made up of complete sentences, **even if it is in a list or enumeration**,
2. text that makes use of 'normal' words.
3. text that is written in English.

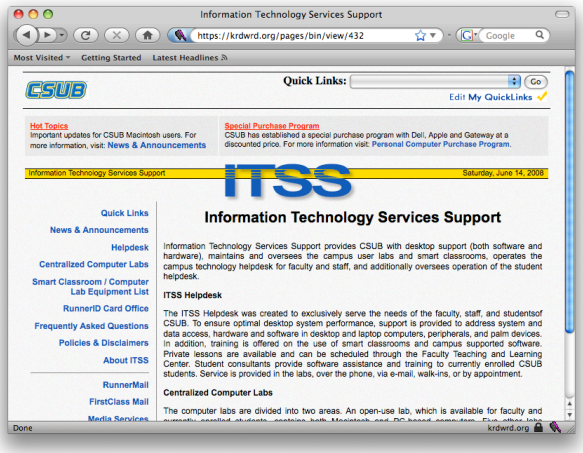
Simple, isn't it? You will notice that on some pages you can only highlight very large areas, on others the choices are less restricted. If you tag an element, the tag assigned is propagated to all elements that are contained in this area. However, if you are not sure whether a specific element is entailed, just tag it too to be on the safe side (remember the *sidebar option* mentioned in the previous section!).

In a previous section, we said that as a rule of thumb, it often makes sense to tag *everything* in red ('bad'), from top to bottom, and only then to start tagging smaller pieces in yellow or green ('uncertain' or 'good', respectively). The easiest way to tag a whole page red is to tag the outermost rim of the page and tag that as 'bad'. Due to the tag propagation, the whole page is now tagged as 'bad'. If you want to make sure that this is so, check the sidebar (see above).

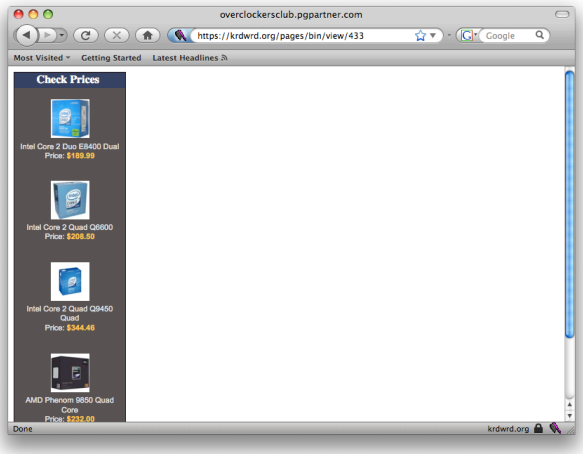
This may all be a bit confusing now. But fear not, in the next sections you will have the possibility to check whether you understood everything.

3.2 Examples - Easy

- Example 1
  This is a fairly standard Web page. Advertisements and boilerplate should be easy to spot and easy to tag.



- Example 2
  This should be easy, too.

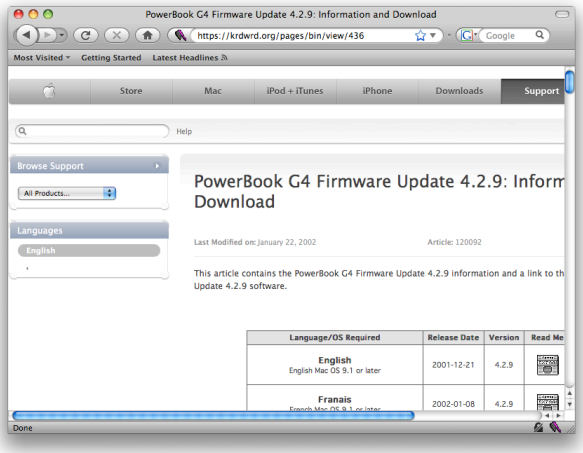


- Example 3
  Similar to *Example 1* but you will have to invest a little more time, since the layout is not as clean. Is there something that is not 'good' in the text portion? How should you treat the headlines? What about the headlines' subtitles?

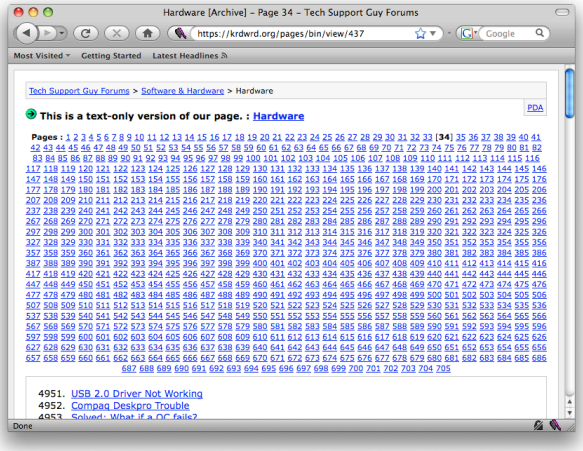
- Example 4
  Somehow similar to *Example 2*. Why is even the text portion not 'good'?

3.3 Examples - Medium

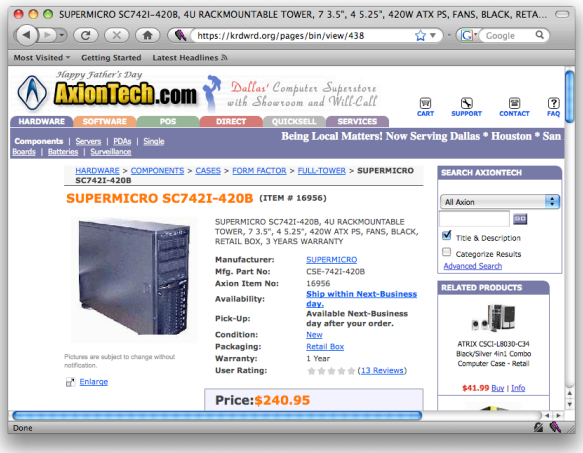
- Example 5
  Remember wich language you should tag (and that all text in another language is bad). How should you tag the enumerations?



- Example 6
  This one is all about enumerations.

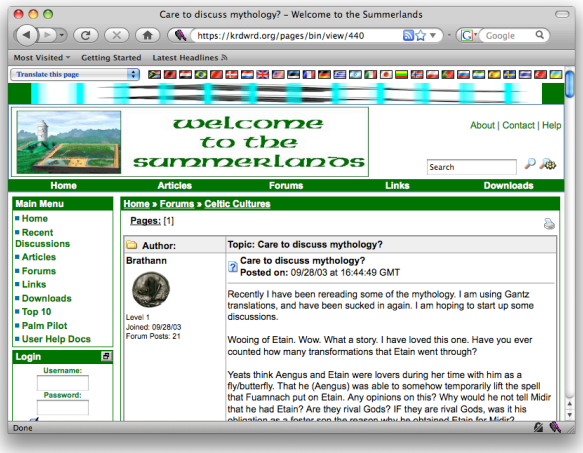


- Example 7
  Once you have decided how much of the text is junk, this is fairly easy. Propagate is you friend.



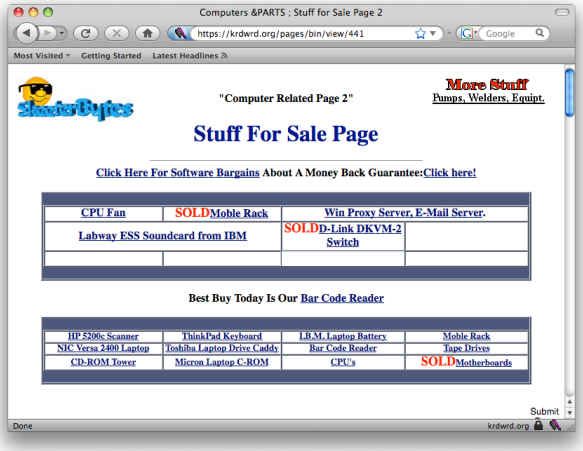
- Example 8
  This can be easy with the right strategy. One of the rare pages where it is easier if you don't start with tagging everything red first.

- Example 9
  Sometimes there are no technical difficulties.

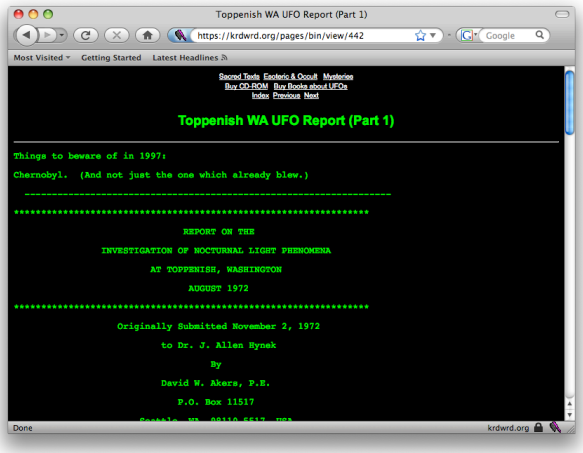


3.4 Examples - Hard

- Example 10
  By now this should be easy for you.



- Example 12
  This one is a bit like example 9 but with technical difficulties. You might rather want to go to your dentist. This is really as bad as it can be. If you can do this all other pages are a piece of cake.



4 Tips & Tricks

4.1 Keyboard Shortcuts

The default shortcuts for tags are

- *bad*: ctrl+alt+1
- *uncertain*: ctrl+alt+2
- *good*: ctrl+alt+3
- *clear annotation*: ctrl+alt+4

Depending on the size of your keyboard and your hands, this may really hurt after some pages. But you can change the shortcuts. All you need is the keyconfig add-on.

- Install keyconfig.
- Bring up the keyconfig menu by pressing *Ctrl+Shift+F12*. (Mac users press *Command+Shift+F12*).
- The commands you want to change are named **Tag Bad**, **Tag Good** and **Tag Unknown**.

- Close your Firefox window and reopen it; otherwise, the newly set shortcuts will not work.

4.2 How and When To Use Propagate

There are two main uses for the propagate utility. Either there are many good text portions embeded in bad text portions (or vice versa). Or there are many small chunks of text cluttered around the page.

With propagate you can often get around tagging each chunk individually.

- Remember to check each text portion's tag to be correct.

- It is important that text is tagged right, you don't have to care about the background color (and you really shouldn't).

- Propagate will tag text and text only. And you really should not care about the color of the background where the text is written on.

- Most pages in Examples - Medium, page **??** are significantly faster to tag when using the propagate utility.

4.3 How to *Undo*

Currently the add-on has no readily available *Undo* function (which might come handy in cases where you *propagated* the wrong tag). However, the *My Stats* page lets you *Delete* certain, committed pages.