KrdWrd Homework

due Friday, 18.07.2008

The main objective of this assignment is to give you first hand experience on a competitive manual tagging task on pages from the World Wide Web, where you will use an online tool to tag the pages. It is *competitive* because each of your individual results will compete with others' results on identical pages and it is *manual* because you will actually have to do the task.

Pages from the Web because the Web is an unprecedented and virtually inexhaustible source of authentic natural language data and offers the NLP community an opportunity to train statistical models on much larger amounts of data than was previously possible. However, after crawling content from the Web the subsequent steps, namely, language identification, tokenising, lemmatising, part-of-speech tagging, indexing, etc. suffer from 'large and messy' training corpora and interesting regularities may easily be lost among the countless duplicates, index and directory pages, Web spam, open or disguised advertising, and boilerplate. Therefore, thorough preprocessing and cleaning of Web corpora is crucial in order to obtain reliable frequency data.

The preprocessing can be achieved in many different ways, e.g. a naïve approach might use finite-state tools with hand-crafted rules to remove unwanted content from Web pages. The KrdWrd project is heading for a quite different approach:

1. Use the visual presentation of Web pages
2. Have an initial training set of Web pages annotated – the Gold Standard
3. Extract the visual, structural, and linguistic information to train a model using machine learning algorithms
4. Use the model to automatically clean Web pages while building a corpus

The KrdWrd project homepage is available at `http://krdwrd.org` and this is also the place to get started.

Exercise 1. (2 points)

Your task is to complete the online tutorial of the KrdWrd system, i.e. you have to launch Firefox[1], install the KrdWrd Add-on, get a copy of the manual, and go through the online tutorial.

1. Use Firefox to visit `http://krdwrd.org` and follow the instructions, i.e. install the necessary certificate
2. Go to `https://krdwrd.org/trac/wiki/AddOn` and follow the installation steps for the add-on
3. Read through the manual – make sure to cover at least *Introduction* and *Getting Started*
4. Start the tutorial by selecting *Start Tutorial* from the KrdWrd broom status bar menu on the lower right of your browser window
5. Read through the page – thoroughly – and finish the tutorial

[1] If you have not installed Firefox, yet, visit `http://www.mozilla.com` and download your copy.

Exercise 2. (8+10 points)

Your task is to tag pages from the Canola corpus: 15 well tagged pages will be worth 8 credits[2]; well tagged additional 10 pages will be worth 10 *extra* credits.

1. Select the *Canola* corpus from the *Corpus* menu and start tagging
2. Visit the *My Stats* page from time to time to see how many pages you have already tagged...

Notes:

- You can always interrupt tagging and continue at a later time: just select the Canola corpus and continue.
- In case where something goes wrong go to `https://krdwrd.org` and use the search feature to look for a solution.
- Iff you have not found a solution write a mail to `krdwrd@krdwrd.org` with a detailed problem description, i.e.:
  - What is the problem?
  - What were the steps that led to the problem?
  - Include the last lines of information from the *Help/About Mozilla Firefox* menu, i.e. the ones that start with `Mozilla/...` **and** include the first line of information from the *About* menu item in the add-on, i.e. the line `krdwrd version 0.x.y`.
  - * We know that well-written bug reports are an extra effort and encourage them. Every unique substantial bug report leding to a fix in the add-on is worth a maximum of three extra credits[3].
- In case you have no other hardware to use you can use the computers in B10 aka. 31/412; however, make sure to substitute every occurrence of Firefox with Iceweasel, i.e. `s/Firefox/Iceweasel/g`, in all documentation.

[2] Together with Task 1 this corresponds to 100% of this assignment.
[3] However, you cannot exceed the 10 *extra credits* hard limit for this assignment.