Description of the Annotation files

Each Web file (".htm") is accompanied by the corresponding text file (".txt") that contains the annotated data by the human expert. Table 1 shows an annotation text file

```
file:C:/Program Files/samplePages/adeli1-3780FHT.htm
1670
1165 1170 {32 MB} {ram} {ram} {} {product_1} {0}
1145 1151 {2,1 GB} {hdcapacity} {hdcapacity} {} {product_1} {0}
1117 1120 {TFT} {screentype} {screentype} {} {product_1} {0}
1105 1116 {12,1"} {screensize} {screensize} {} {product_1} {0}
1070 1077 {Pentium} {processorname} {processorname} {} {product_1} {0}
1078 1081 {166} {processorspeed} {processorspeed} {} {product_1} {0}
1275 1281 {3 years} {warranty} {warranty} {} {product_1} {0}
```

Table 1: Sample of an annotation text file

The **first line** is the string "file:" followed by the full path of the Web file. The full path is not important. The important issue is that the root name of the annotation file must be identical to the root name of the Web file. E.g. The annotation file for the Web file "adeli1-3780FHT.htm" must be called "adeli1-3780FHT.txt".

The **second line** is the character-size of the page. Also this information is of no particular interest.

Each line, from the **third-one** until the end of the file, contains information that corresponds to a different annotation. The two integers correspond to the (start, end) annotation offsets. The string in the first pair of curly brackets contains the annotation text, or may be empty, since the text may contain new-line characters, and thus resulting in problematic annotation files. The next two pairs of curly braces contain the name of the extraction field (e.g. manufacturer or ram). The next two pairs of curly braces are of no particular interest. The last pair of curly braces contains the confidence score attached to the corresponding predicted annotation by an information extraction system. Since we deal with hand-filled annotation files, this value is set to zero or may be empty.