The Sample

This is a simple test document for extracting some level of **layout and formatting** information in addition to the text from the document. The intent is to extract this information from any document not just this sample (*if it only works on the sample, that would not be too useful*)

Lists Are important as well

There are bullet lists AND multi-level lists in many types of documents (like a CV)

- Bullet lists should be tagged at the start and end with <li>list member</li>
- This is just like a list item in html, except no start / end list..
- **This is another member of this list**

Multi-Level Lists

These lists may be more difficult for extraction (I have no idea)

1. The first item in the list
2. The Second item in the list
    a. And then there was a child
    b. And the Child had a sibling
3. And the third item followed
    a. And it had a child

The last thing to be aware of is the Table!

For these extracting just the data will be fine, though I would like to know if it is possible tag the columns/rows (not in this first project…)

This is	Just	a					Table	Sample
Nothing	of the	table	aside	From	The	text	needs	to
be	extracted							

The other stuff you may find

Header information, footer information, images, etc, get the text or ignore (ignore is better)