ZANRAN
PDF PROCESSING

Discussion paper

PDFs, cognitive computing and layout analysis

Introduction

In this paper we we will be looking at the role of layout analysis. Where it is useful, where not – particularly with cognitive computing in mind.

Why are PDFs difficult to handle by cognitive computing?

Documents can have complex structures made up of successive chapters or sections over many pages and having regions with various different types of information: paragraphs, words, letters, graphics, images, tables, logos, equations, columns...

A human eye can immediately make sense of a page and automatically organise the information into letters, words, paragraphs, tables… A person doesn't need to be told where a word starts and ends. The brain re-creates the layout without the need of additional information such as word delimiters & boundaries.

Computers on the other hand are not good at this task and need to be told where the words, paragraphs, tables and other boundaries are. This information is usually defined in tags – for instance the tag <p> in html is used to delimit paragraphs.

Unfortunately many documents - including PDFs - have virtually no structural information.

PDF files have no underlying tags to delimit words, paragraphs or tables. A PDF at a low level can be seen as an assembly of letters and pixels with no concept of words, paragraphs, tables, diagrams, maps, etc. The PDF format was designed for printing - with no need for structural information or content labeling.

This lack of structure explains why it isn't possible to copy and paste a table into excel, and why one cannot directly edit and change a PDF file.

What is meant by 'cognitive computing'?

It's a broad term, and generally understood to be part of artificial intelligence (AI) that deals with 'understanding' written or spoken language. Some examples of cognitive computing might help

sentiment analysis, e.g.
- trying to assess attitudes towards a consumer product from thousands of Tweets

article summarisation – creating a paragraph that gives an overview of a long document

drug discovery - 'mining' lifescience reports for new connections

speech recognition – e.g. giving instructions via Apple's Siri or Microsoft's Cortana

In all these, a computer is doing a complex language task.

What's made cognitive computing happen now?

The ongoing revolution in the field of Artificial Intelligence was made possible by the combination of two factors :

1) the access to large corpora of text and data, and

2) the ability of computers to learn without being explicitly programmed (Machine Learning).

A scanned document (as an image or as a PDF) has even less layout information than a native PDF – it doesn't even differentiate letter pixels from image pixels. An OCR system is needed to recognise a group of pixels as representing a specific letter.

The lack of structure can be fixed by processing PDFs through complex algorithms that attempt to identify the various regions, the flow of text, etc.

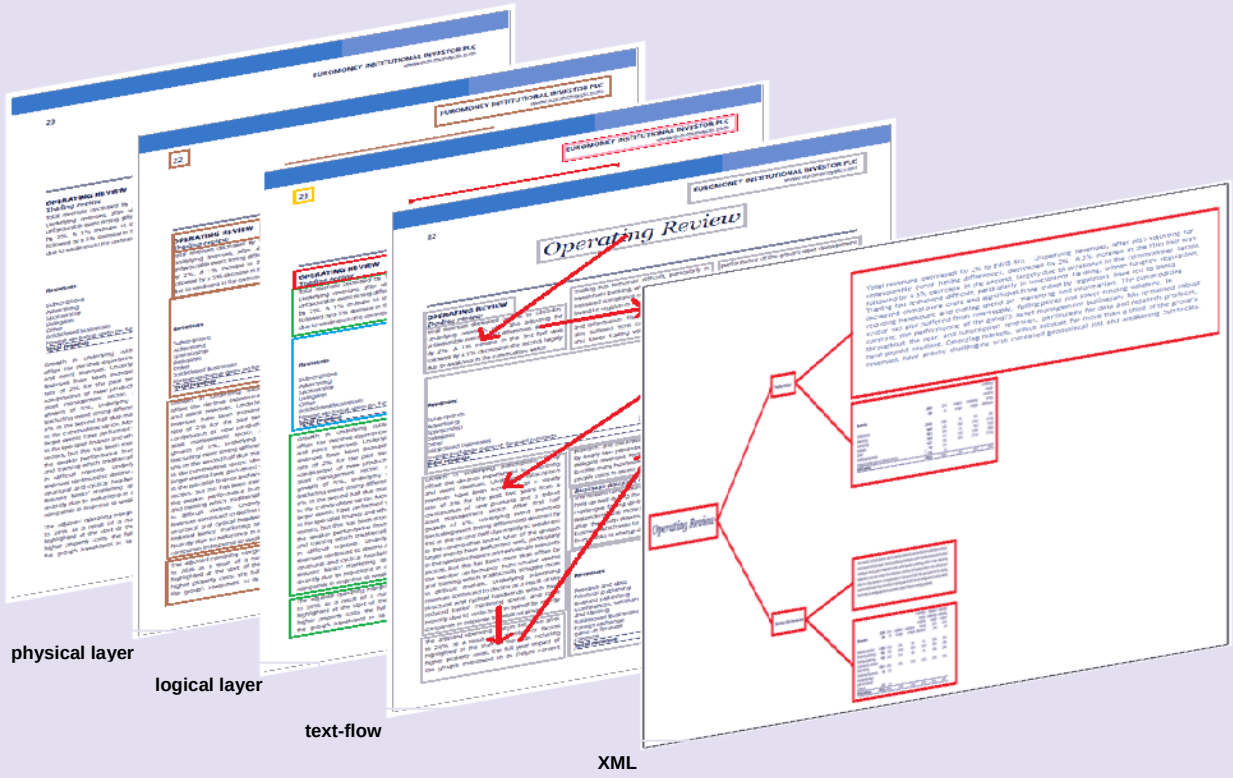
Layout analysis

The identification of all the regions on a page is done through a process known as 'document layout analysis'.

Document layout analysis first divides up the document into regions with homogeneous content and secondly assigns a meaning to the regions. The 'dividing up' step is called 'physical layout analysis', and is used to identify the geometric page structure. It's essentially a visual process – not reliant on the meaning of the content.

Taking the regions and labeling them (as titles, captions, footnotes, sections etc.) is the scope of 'logical layout analysis' and is essentially semantic – it relies on analysis of language. Document layout analysis is the combination of geometric and logical labelling.

On top of these two processes one can add a text-flow analysis layer and a tree-like representation layer of the document content. All this information can easily be kept in a single XML file.



In the figure above, the physical layout analysis is shown as the second layer and the semantic labelling as the third layer. The fourth layer shows the text-flow. And the fifth shows the representation as XML.

So when is this useful?

Examples and applications

We'd like to give several examples to give an indication of when layout analysis makes content extraction and cognitive computing easier or more viable. We'll start by briefly mentioning some cases where it is NOT useful.

Layout analysis NOT necessary

The key here is either to have text that's simple, or where it doesn't matter greatly if you get it wrong.

Example 1: extracting a standard piece of data from a document – for example, extracting the names of Chief Executives from Annual Reports

You can find the names automatically by using a well-designed regular expression that will pick out the capitalised words that come before 'Chief Executive', 'CEO', etc. You don't need to know the layout of the pages.

Example 2: enterprise search

If you are building a search engine to find internal documents, emails, etc, you need to index the text of the documents. It doesn't matter greatly if the text order is slightly wrong or if some spurrious details are included (page numbers, footnotes, graph labels, etc). So you don't need to know the layout of the pages.

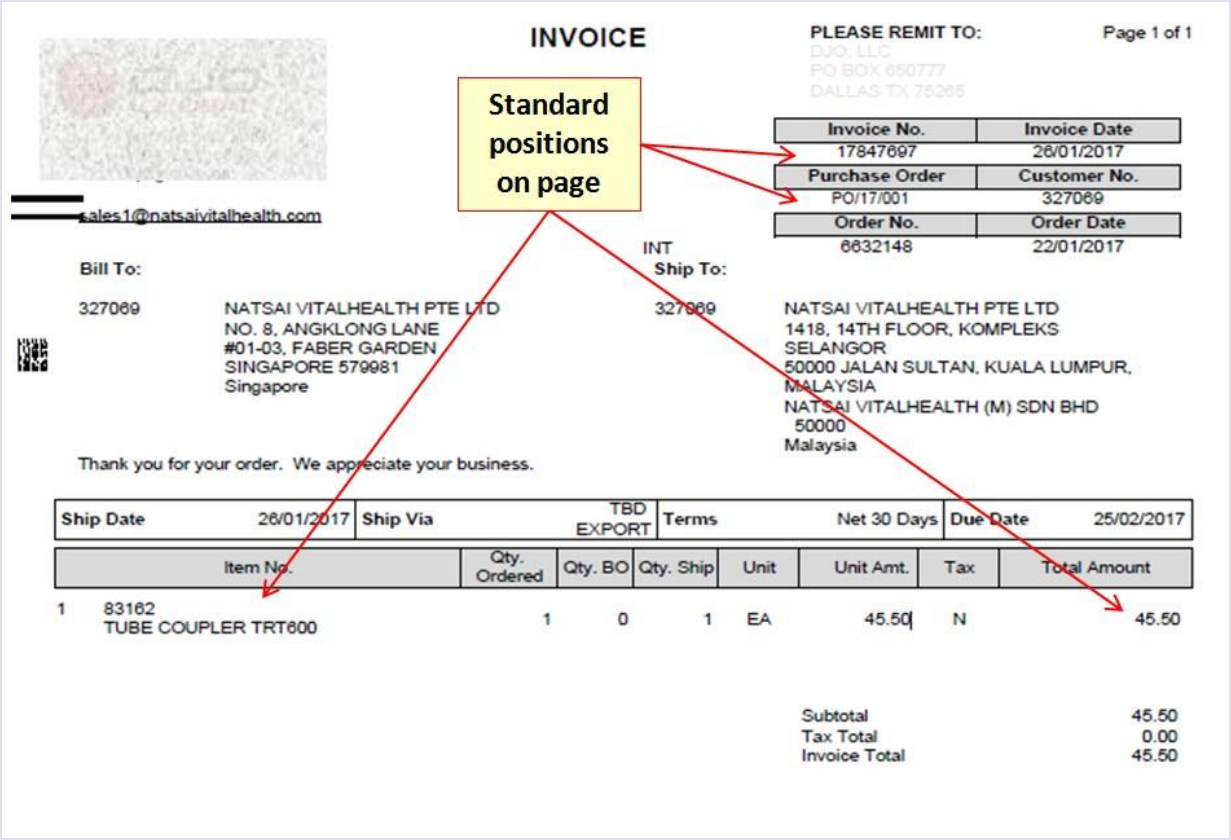
Example 3: sentiment analysis of simple text

If, for example, your source text has simple layout – like millions of Tweets or emails – then your analysis can focus on the text. You don't need to know the layout of the pages.

Layout analysis NOT necessary - *continued*

Example 4: ingesting invoices

Many large accounting systems take in invoices as PDFs (native or scanned). The content is extracted into the accounts package because the invoices from each supplier will look similar every time. The data can be extracted using a template which is built for each invoice type.

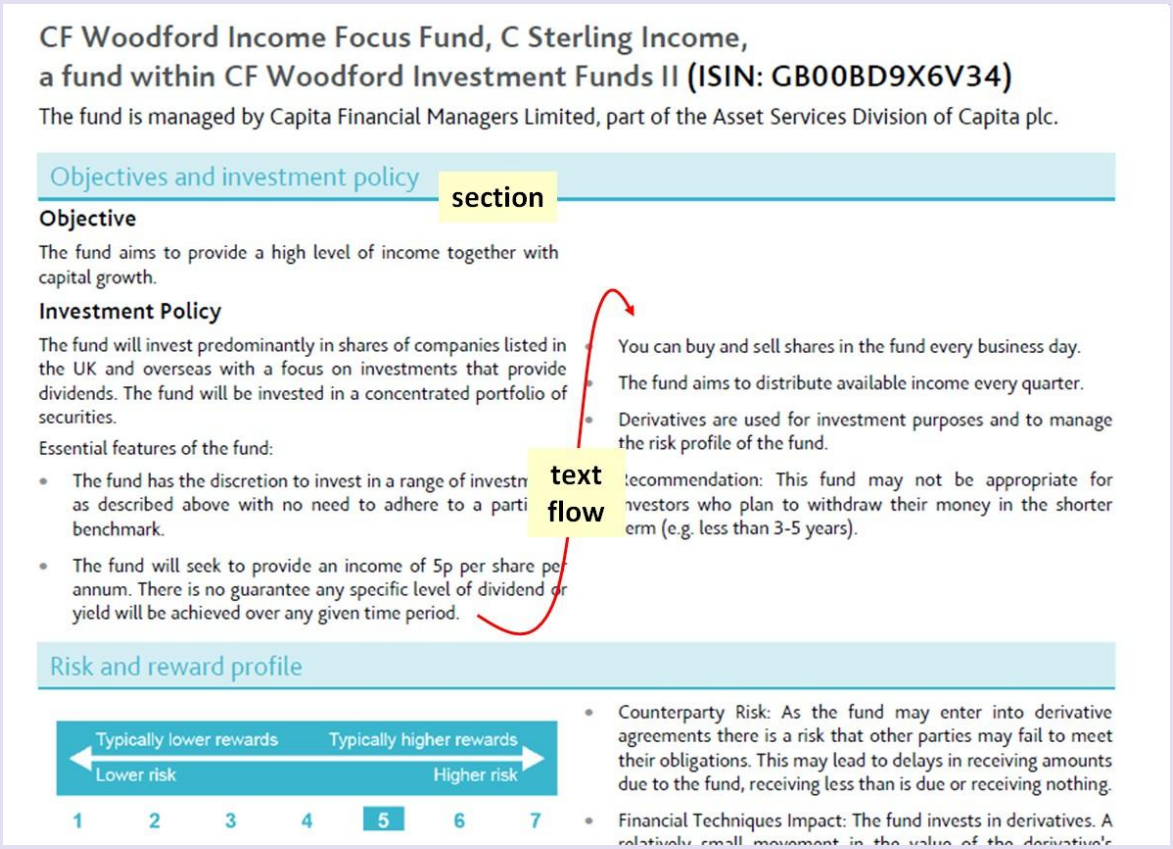


Although at first sight it appears that this would be a good application for layout analysis and cognitive computing, in practice the problems are addressed diffferently – because of the repeat nature of the documents. It doesn't need either layout analysis or AI.

Layout and text-flow analysis necessary

Example 5: extracting specific text from a document

5). First, here is an example where, to extract the text from each section of the document, the computer needs to understand both 'sections' and 'columns' as well as the text-flow. In other words you need layout analysis.



Layout and flow analysis necessary, *continued*

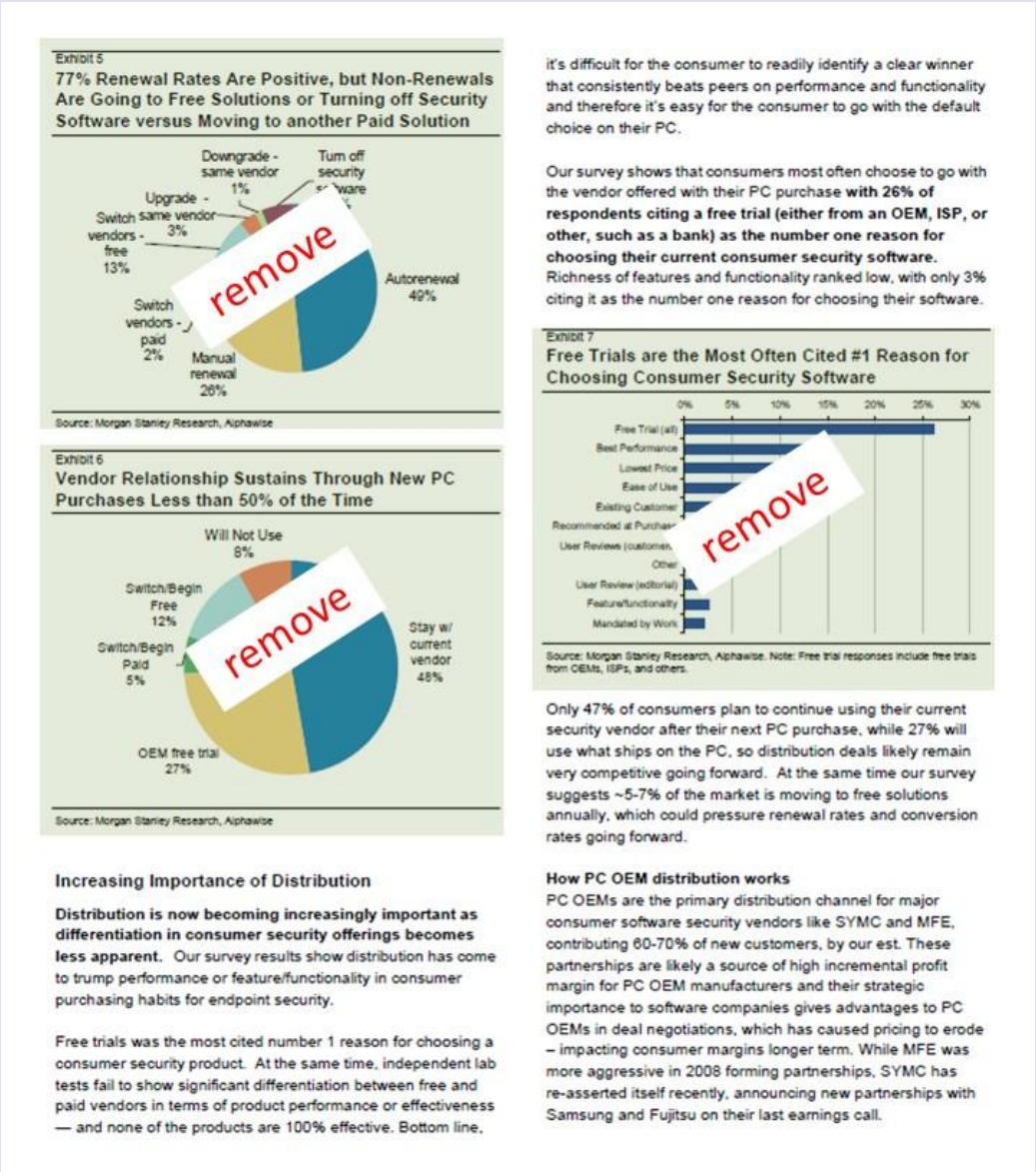
Example 6: textual analysis of a document

A cognitive analysis of a document will be much more effective if it can be done on a clean version of the document where all the text in charts & tables, footnotes, page headers, captions… has been removed.

This is necessary to avoid:

- introducing unwanted text such as axis labels, captions, tables headers, etc. in the natural flow of a document.
- interfering with cognitive algorithms; for instance, a search for a paragraph section with the phrase 'profit & loss' risks being pulled towards table regions that are likely to include some headers with the same words
- interfering with the statistical distribution of words; most algorithms rely on the word counts for indexing, categorization, searching purposes. The 'pollution' of the word distribution by unwanted regions will affect the efficiency of these algorithms

The example below shows how the charts, the table and their captions and notes (artificially given shaded backgrounds) need to be removed to make sense of the core text.



Layout necessary

Example 7: extracting data from tables

When wanting to make sense of the content in tables, layout is vital.

The identification of the table boundaries, its columns, rows, headers… is a very complex layout analysis sub-task.

The cognitive computing can then be used to answer questions like “was industrial production up or down in 2012?”

-on-year decline in turnover on the euro area market and the domestic market slowed over the first four months of the year.

was up just under 2%, ceasing the negative developments in the two previous months. The year-on-year decline in turnover diminished to just under 1%. After recording a sharp decline in the final quarter of last year, this year administrative and support service activities are recovering particularly well for the moment.

Economic Activity	2010	2011	2012	12 m. to Apr-13	2013 Mar.	2013 Apr.	2013 Apr.
			y-o-y in %				++
Industrial production: - total *	7.0	2.0	-0.7	-0.9	-3.3	**-1.1**	*2.0*
- manufacturing	7.4	1.8	-1.9	-2.2	-5.6	**-1.4**	*2.0*
Construction: - total **	-16.9	-24.8	-16.8	-18.9	-31.7	**-19.3**	*-6.4*
- buildings	-14.0	-39.7	-17.3	-26.0	-50.1	**-37.2**	*-2.8*
- civil engineering	-19.0	-15.3	-16.6	-13.8	-13.4	**-6.7**	*-6.6*
Trade (volume turnover)							
Total retail trade	-0.3	1.7	-2.3	-3.7	-5.9	**-3.7**	*-1.5*
Retail trade except automotive fuel	-1.6	-2.2	-4.7	-5.1	-6.3	**-1.4**	*-0.6*
- food, beverages, tobacco	-1.6	-2.9	-4.8	-4.3	-3.3	**-3.1**	*-0.8*
- non-food (except automotive fuel)	-1.6	-1.7	-4.9	-5.6	-9.2	**-0.2**	*-0.4*
Retail trade and repair of motor vehicles	11.9	7.6	-5.3	-5.6	-4.9	**4.7**	*1.0*
Private sector services *** +	6.3	3.2	-2.3	-2.6	-4.0	**-0.8**	*0.2*
Transport and storage +	19.6	8.0	0.7	-0.4	-5.2	**0.3**	*0.0*

Sources: SORS, Eurostat, Bank of Slovenia calculations.

Notes: Data are working days adjusted.

* Volume of industrial production. ** Real value of construction put in place. *** Excluding trade and financial services. + Nominal turnover.

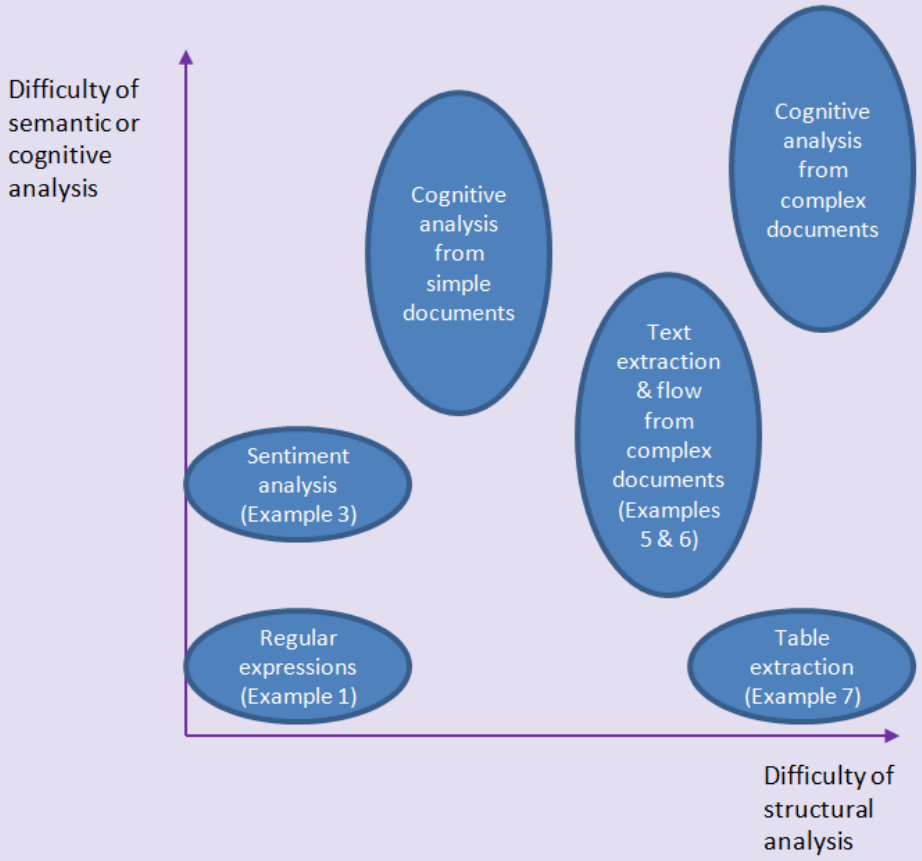
++: 3-month moving average compared to the corresponding average 3 months earlier. Data are seasonally and working days adjusted (except for construction where data are seasonally adjusted).

Conclusion

We contend that layout analysis is most useful when:

- the text for cognitive analysis is interspersed by graphics and tables
- key content is present in tables – which vary from one document to another
- text-flow is important

Finally, when looking at any new project, it's worth considering the following technical matrix:



In other words, it's useful to think of the feasibility of the cognitive component and the feasibility of the structural component independently.

© Zanran Ltd, 2017