



Mar 05, 2020 07:00:00

Why is it difficult to extract text from PDF?



This article, originally posted in **Japanese** on 07:00 Mar 05, 2020, may contains some machine-translated parts.

If you would like to suggest a corrected translation, please click [here](#).

A PDF file is a data format that can be viewed on a PC in any environment without breaking the display of text and images. However, if you try to copy text data from PDF, you may not be able to select it properly, or the text content may be incorrect.

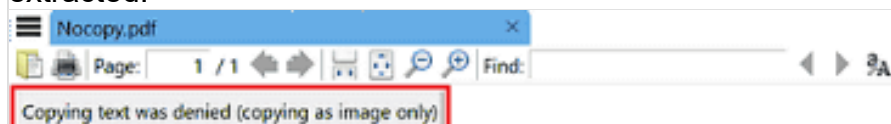
FilingDB, an organization that **converts** PDF files into text and creates databases, reports why it is difficult to extract text from PDF files.

PDF text extraction | FilingDB

<https://www.filingdb.com/pdf-text-extraction>

◆ Read protection

Some PDF files have protected content. Even if the text itself is displayed correctly, if you try to copy the text, a message such as 'Copying text was denied' is displayed, and the text cannot be extracted.



do not copy this text

The reason why copying is not possible is that the PDF file has 'permissions' that determine whether text copying is allowed. This setting prevents the PDF viewer from copying text even if the PDF is displayed without any problems.

◆ Off-page characters

PDF files may contain more text data than is actually displayed on the page. The image below is a PDF file of Nestlé's 2010 Annual Report.



The file above does not show on the page that 'KitKat celebrated its 75th anniversary in 2010. It is still young, sensitive to trends and has more than 2.5 million fans on Facebook. It is sold in more than 70 countries and is also experiencing strong growth in developed countries and emerging markets such as the Middle East, India, and Russia. Japan is the second largest market. '

This text is not displayed in most PDF viewers because it was actually placed outside the page border. However, because the creator forgot to erase the data, the data itself remains, and it will be displayed when extracting the text of the entire page.

◆ Too small and invisible characters

PDF files can contain very small or hidden text. For example, the text of Nestlé's 2012 Annual Report shows a small white text 'Wyeth Nutrition logo Identity Guidance to markets' and 'Vevey Octobre2012RCC / CI & D' on a white background. These texts may have been inserted into the notes when creating the text or for ease of searching.



◆ Extra space

Text data in PDF files may contain extra spaces between the letters of a word. This most often occurs because of a process called **Kerning** that adjusts the distance between characters.

For example, in the annual report of Hikma Pharmaceuticals in 2013, if you copy the text in the red frame part, it will be copied with unnecessary spaces like `` CH AIRMAN 'SS TAT EM EN T'`, and the words will be separated. It may become.

CHAIRMAN'S STATEMENT

AN EXCELLENT

◆ Loss of space

In addition to containing unnecessary space, the original space may be lost or replaced with a different character. For example, if you copy the text `` Ten years after the financial crisis started `` in the following excerpt from the 2017 SEB's annual report, the space disappeared as `` Tenyear safterthefinancialcrisisstarted ``. It may be copied in a state.

Global trends drive change

Ten years after the financial crisis started

to spread, global trends have reshaped the banking industry. A new regulatory frame-

Also, in the following 2013 Eurobank report, if you copy the text 'On April 7, 2013, the competent authorities', the space is replaced by an underscore and 'On_April_7 , _2013, _the_competent _authorities '.

On April 7, 2013, the competent authorities deci
Eurobank, which had been completely deprived of

For extra space and lost space, converting to text with

Optical Character Recognition (OCR) is more efficient than copying the text with a PDF viewer.

◆ Embedded font

Handling fonts in PDF is complex. Some PDF documents may contain non-standard fonts or proprietary encodings, which may cause text to be displayed as different characters or may be recognized as image data instead of text in PDF files. Extracting text becomes more difficult

recognized as image data instead of text in PDF files, extracting text becomes more difficult.

◆ Text and paragraph order

Extracting paragraph order is difficult in two respects. First, there may not be a correct answer. For example, it is unclear whether the text in the red frame below should be inserted at the end of the sentence on the entire page or displayed in the middle of the sentence, and it is often unknown only to the person who created the text.

Key experience: Everything is possible in life as long as you keep fighting to reach your goal.

Main inspiration: My family and my first manager at SEB, Madeleine Stjernrup Öberg.

issues, implement measures and follow up on progress. SEB's Board of Directors and the Group Executive Committee adopted a governance document which states that inclusion and diversity are critical for the bank's long-term success and that SEB can and should do better in these areas.

Every year SEB conducts a Global Talent Review to identify individuals with potential for a future key role or management position.

Labour law and unions

SEB employees are covered by collective or local agreements. SEB has a European working council with representatives from all EU and EES countries in which SEB is represented.

Recruitment in new arenas

SEB has a strong employer brand according to annual rankings conducted among students and young professionals. This applies especially for finance and business administration students. In pace with the ongoing competence shift and growing recruitment need in new competence areas, the bank needs to strengthen its attractiveness among individuals that are attracted by IT companies and start-ups. Accordingly, SEB has widened its recruiting activities. The bank not only participates in traditional recruitment fairs for finance students but also uses interactivity and new formats such as invitations to hackathons and open workshops on artificial intelligence, blockchain technology and other cutting-edge technologies.

SEB's core values

Customers first

We put our customers' needs first, always seeking to understand how to deliver real value.

Commitment

We are personally dedicated to the success of our customers and are accountable for our actions.

Collaboration

We achieve more working together.

Simplicity

We strive to simplify what is complex.

SEB's core values serve as the foundation for the bank's ways of working and culture, and in combination with the bank's vision – to deliver world-class service to our customers – they serve to motivate and inspire employees, managers and the organisation as a whole. These values are described in SEB's Code of Conduct, which provides guidance on ethical matters for all employees.

➔ [Read the Code of Conduct on sebgroupp.com](https://sebgroupp.com)

And reading is basically from left to right and from top to bottom, but it is not clear whether the order of ABCD or ACBD is correct in the following paragraph division. The human brain can determine the correct order by reading the text and understanding what is being written. However, it is difficult to determine the correct order with the algorithm of the program.

Prepare your utensils.

Make sure you have a bowl, chef's knife, a cutting board, spoon and fork.

A

Wash the vegetables.

Lukewarm water will do just fine. Make sure you do a thorough job.

C

Cut the vegetables.

Using the cutting board, chop all vegetables in bite-size pieces.

B

Mix the vegetables and add oil.

Put all the chopped vegetables in the bowl, add oil and give them all a good mix.

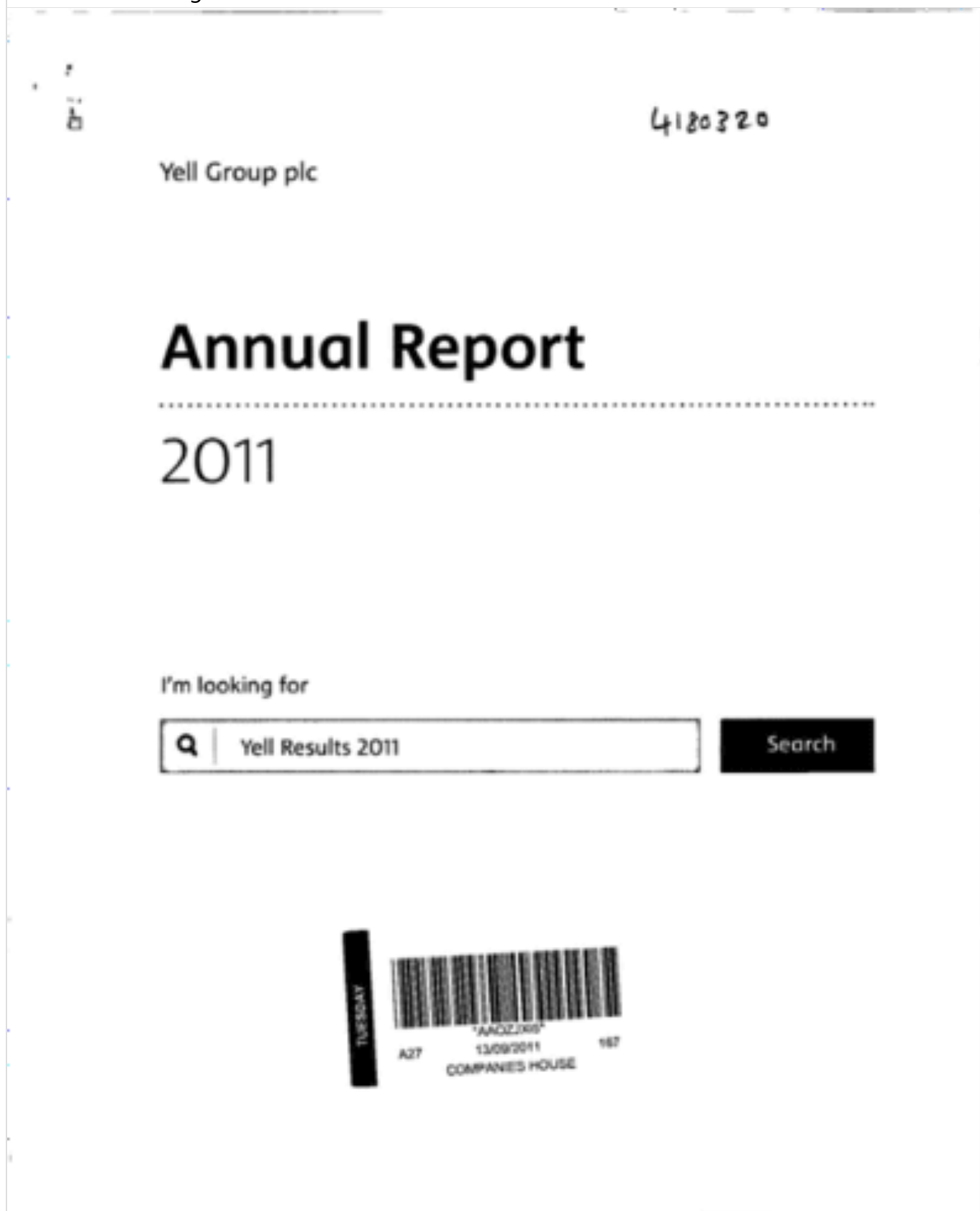
D

A common approach to paragraph order is to refer to the order in which the text is stored in the PDF file.

◆ Embedded image

In many cases, all the contents converted to a PDF file are saved as image data instead of text data. In such cases, there is no text data that can be directly extracted, so you have to rely on OCR.

For example, PDF files such as the following YELL's annual report for 2011 below are all stored in PDF files as image data, not text data.



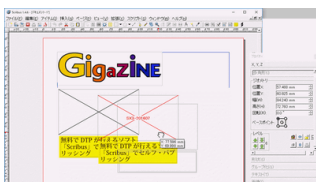
www.yellgroup.com/annualreport



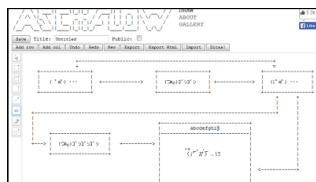
Most of the issues raised by FilingDB can be solved with OCR, but OCR also has some drawbacks. First, scanning with OCR takes ten times longer than extracting text directly from a PDF file. In addition, OCR may not be able to handle characters such as emoticons or complicated mathematical symbols, and it is not possible to order text by referring to the insertion order like PDF files.

In the first place, PDF files are not designed as a data input format like a text editor, but as a data output format that outputs documents precisely. PDF files are a poor format for text extraction, and FilingDB recommends that you check for data in other formats before extracting text from PDF files.

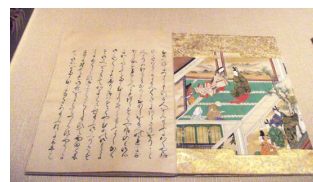
Related Posts:



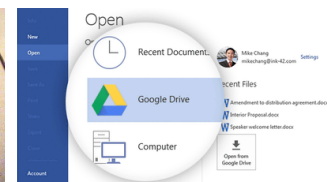
I tried DTP with software 'Scribus' which can do high-level book design free of charge



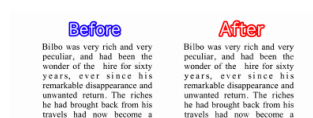
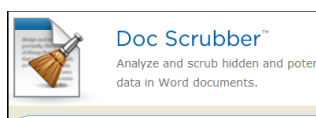
"Asciiflow" which can create a flow diagram with ASCII art



Researchers from overseas are also paying attention to the activation of `` Attempts to change Japanese kanji characters to type with AI ''



Google genuine plug-in "Google Drive plug-in for Microsoft Office" that can save and share directly to Google Drive in Excel · Word · PowerPoint



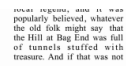


New features that Google Docs & Spreadsheets evolved more easily at the educational level
Conclusion

Scrub Settings:

- ☐ Clear Subject ☒ Clear Last Edited By
☒ Clear Keywords ☒ Clear Last Saved Date

"Doc Scrubber" free software that can erase the creator information etc. of Word file all at once by one shot

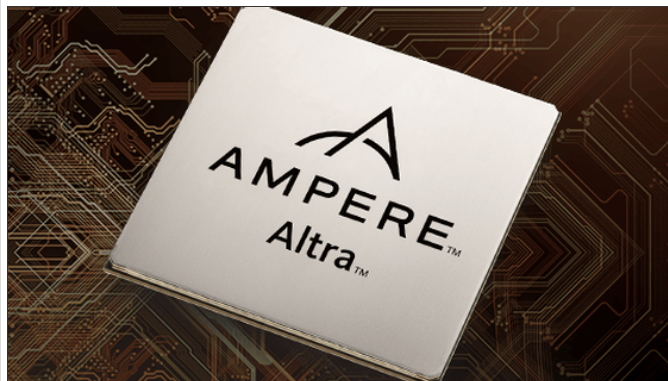


"FontCode" which can hide secret messages in the document at a level that can not be confirmed with the naked eye



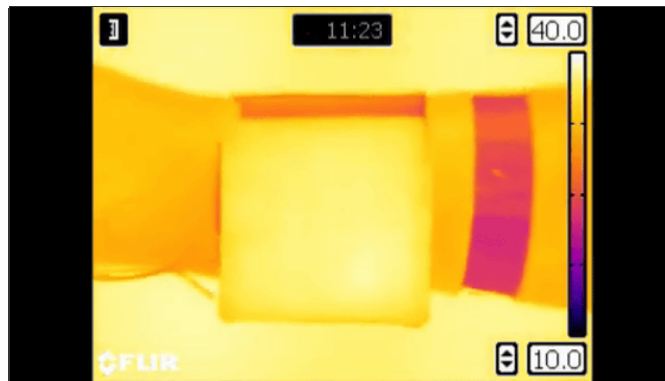
Things to take care to improve the speed of displaying web pages

<< Next



ARM-based 80-core server processor appeared from `` Ampere `` by former Intel president, performance comparable to AMD's second generation `` EPYC ``

Prev >>



A thermal camouflage device that can automatically adjust the released heat and hide from thermography will be developed

Mar 05, 2020 07:00:00 in [Note](#), Posted by darkhorse_log

Latest news 40

- [AMD holds 30% of the CPU market steadily stealing share from Intel](#)

... and holds 30% of the CFC market, steadily stealing share from Intel.

- I tried all of Nissin Foods' ``Immoral Noodle Trio `` that can taste the violence of taste with ``cheese `` and ``butter ``
- Research results that human beings have hair ``because it is cooler to have hair"
- A paper calling for compulsory provision of backdoors in image generation AI such as 'Stable Diffusion' will be published
- Headline news for February 17, 2023
- I tried playing 'Horizon Call of the Mountain' where you can fully enjoy the beautiful graphics and haptic vibrations of PlayStation VR2
- In fact, the core part of DeepMind's go AI 'AlphaGo' and the evolved version 'AlphaZero' has been quietly open sourced
- Meta's language model 'Toolformer' calls and uses search engines, calculators, calendars, etc. with API
- OpenAI proposes to the government to limit AI chips to prevent ``propaganda explosion ", Bing's AI appeals ``I want to be human "
- In 'iOS 16.4' you can check the posted content just by sharing the Mastodon link in the message application
- It turns out that 'push notification from website' is possible on iOS 16.4
- A bill to prohibit children under the age of 16 from creating SNS accounts is submitted in the United States
- Burger King ``King Yeti The One Pounder" tasting review with 4 thick beef patties with cheese
- When I actually set up a VR headset 'PlayStation VR 2' that can be easily connected to PS5 with one cable, it looks like this
- Movie ``Tetris `` trailer release depicting the hardship of obtaining the rights of ``Tetris `` from the Soviet Union
- Microsoft officially releases 'How to use Windows 11 on the latest Mac'
- The founder of the virtual currency 'Terra', which collapsed with a loss of over 5 trillion yen, was indicted for fraud, but he was escaping and missing
- A recall will be announced because danger was found in the fully automatic driving beta version of about 360,000 Tesla cars
- YouTube CEO Susan Wojcicki resigns
- The final trailer of the brilliant action-packed movie 'John Wick: Chapter 4' by the strongest killer played by Keanu Reeves is released
- The AI installed in the search engine Bing said, ``I will not hurt you unless you hurt me first."
- Pointed out that it is difficult to develop personal applications for Apple products because of Apple's worship
- A study is announced that ``eco-friendly" is useful for the success of the organization
- A law will be enacted in the United States to oblige the transition from gas water heaters to electric water heaters by 2027
- Ace cook with 6 crisp wontons ``Wonton noodle's highest peak shop Yakumo no Ippai wonton noodle white soy sauce taste" tasting review
- Artists have been wrong for centuries to express ``shadows" in paintings
- I tried the combined technique of ``ControlNet" & ``Stable Diffusion" that allows you to specify poses and compositions and quickly generate your favorite illustration images Review
- Headline news for February 16, 2023

ARCHIVES

<

2, 2023

>

| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
|-----|-----|-----|-----|-----|-----|-----|
| 29 | 30 | 31 | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 1 | 2 | 3 | 4 |

Select Month

▼

| | |
|-------------|-----------------------|
| Note | Headline |
| Review | Coverage |
| Interview | Gastronomic Adventure |
| Mobile | Software |
| Web Service | Web Application |
| Hardware | Ride |
| Science | Creature |

| Video | Movie |
|-------------------|-------------------------|
| Contacts Manga | About GIGAZINE Anime |
| Add Suggestion | |
| Game | Design |
| Art | Junk Food |
| Security | Notice |
| Pick Up | Column |

Search

×

