A RECOGNITA ÉS A PDF FÁJLFORMÁTUM

avagy

MI AZ A PDF ÉS MIÉRT SZERETJÜK?

A NEUMANN JÁNOS SZÁMÍTÓGÉP-TUDOMÁNYI TÁRSASÁG
INFORMATIKATÖRTÉNETI FÓRUMA (NJSZT ITF) ÉS
AZ ÓBUDAI EGYETEM

„NAGY SZÁMÍTÁSTECHNIKAI MŰHELYEK” SOROZATA

A 30 ÉVES RECOGNITA TÖRTÉNETE

2019. 11. 8.

Urbán Zoltán
zoltan.urban@kofax.com

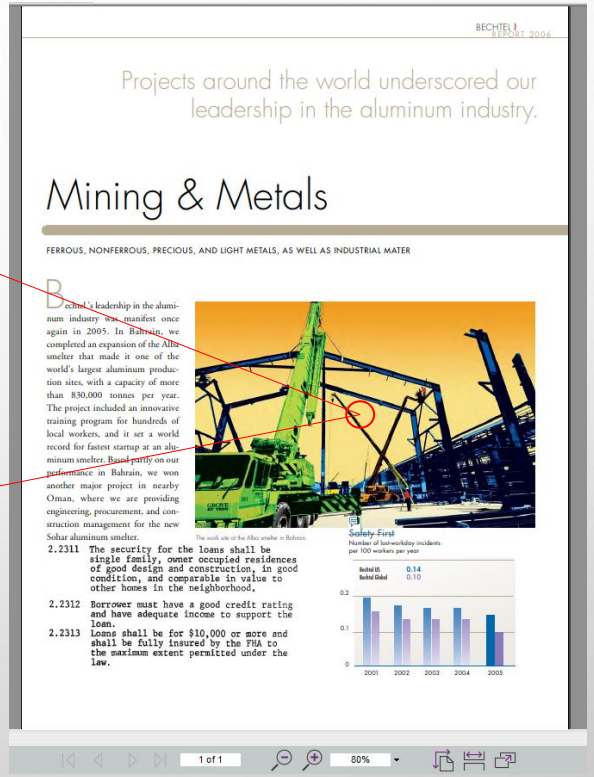
MI A PDF?

- A Portable Document Format (PDF) az Adobe Systems által kifejlesztett nyílt szabványú, dokumentumok tárolására alkalmas fájlformátum, amely alkalmas szöveget, ábrát és képeket tartalmazó dokumentum leírására eszközfüggetlen és felbontásfüggetlen formában.
- A legelterjedtebb technológia a végleges dokumentumok tárolására
  - A megjelenítés hardver-, operációs rendszer és felbontásfüggetlenül, képernyőn és nyomtatásban egyenrangúan eredethű
  - „Csomagformátum”: minden típusú kapcsolódó adat becsatolható a fájlba
  - Ideális formátum papíralapú dokumentumok digitalizálására
    - Az eredeti oldalak szkennelt képe látható
    - A mögöttes szöveg alapján kereshető, katalogizálható

A PDF FÁJL - ALAPISMERETEK

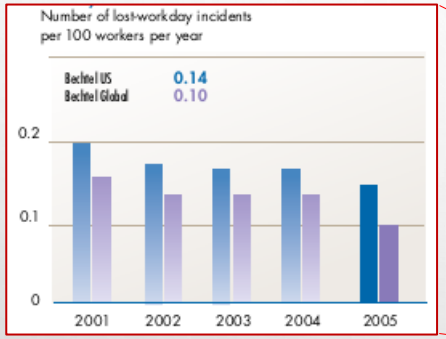
- A PDF fájlok tartalmazhatnak
  - Lapokat, rajtuk
    - Szövegeket
    - Vektoros ábrákat, vonalakat
    - Rasztergrafikákat
    - Annotációkat (kiemelés, szövegbuborék stb.)
  - A szövegekhez tartozó betűkészleteket és karakterkódolási információkat
    - Többé – kevésbé ☺
  - Címkéket (tags)
    - Egyes PDF alszabványokban kötelező
  - Dokumentuminformációkat (metadata)
  - Biztonsági elemeket (titkosítás, digitális aláírás)

A PDF FÁJL - ALAPISMERETEK

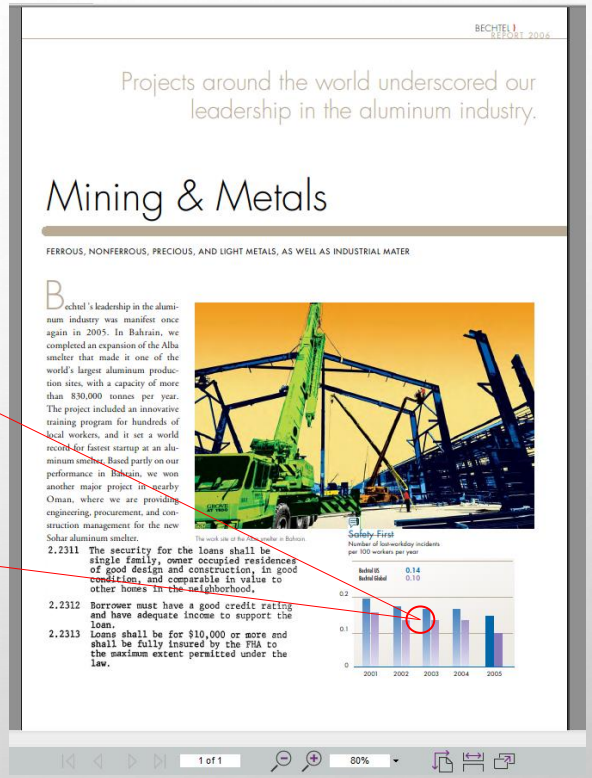


rasztergrafika

A PDF FÁJL - ALAPISMERETEK



vektorgrafika



A PDF FÁJL - ALAPISMERETEK

Bechtel 's leadership in the aluminum industry was manifest once again in 2005. In Bahrain, we completed an expansion of the Alba smelter that made it one of the world's largest aluminum production sites, with a capacity of more than 830,000 tonnes per year. The project included an innovative training program for hundreds of local workers, and it set a world record for fastest startup at an aluminum smelter. Based partly on our

szöveg

A PDF FÁJL - ALAPISMERETEK

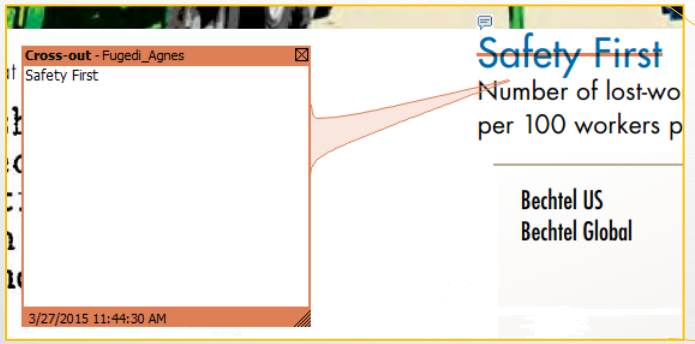
2.2311 The security for the loans shall be single family, owner occupied residences of good design and construction, in good condition, and comparable in value to other homes in the neighborhood.

2.2312 Borrower must have a good credit rating and have adequate income to support the loan.

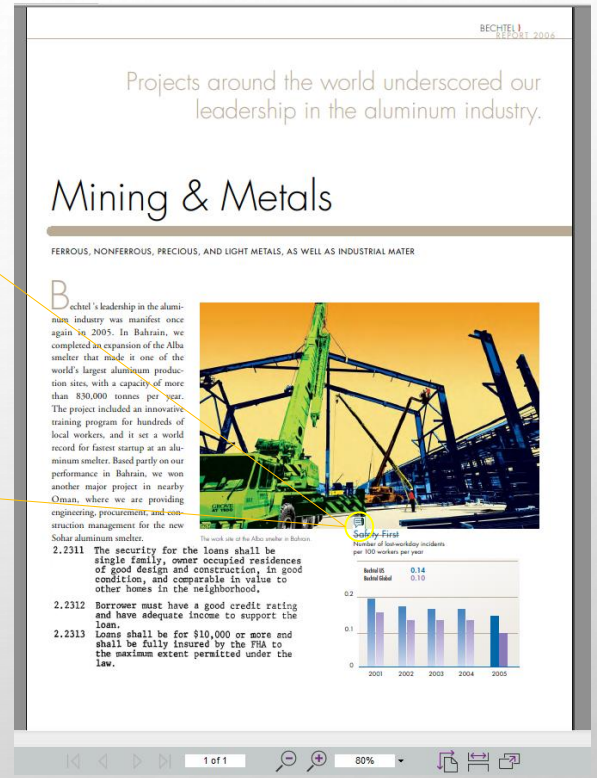
2.2313 Loans shall be for $10,000 or more and shall be fully insured by the FHA to the maximum extent permitted under the law.

Szkennelt (raszteres) szöveg

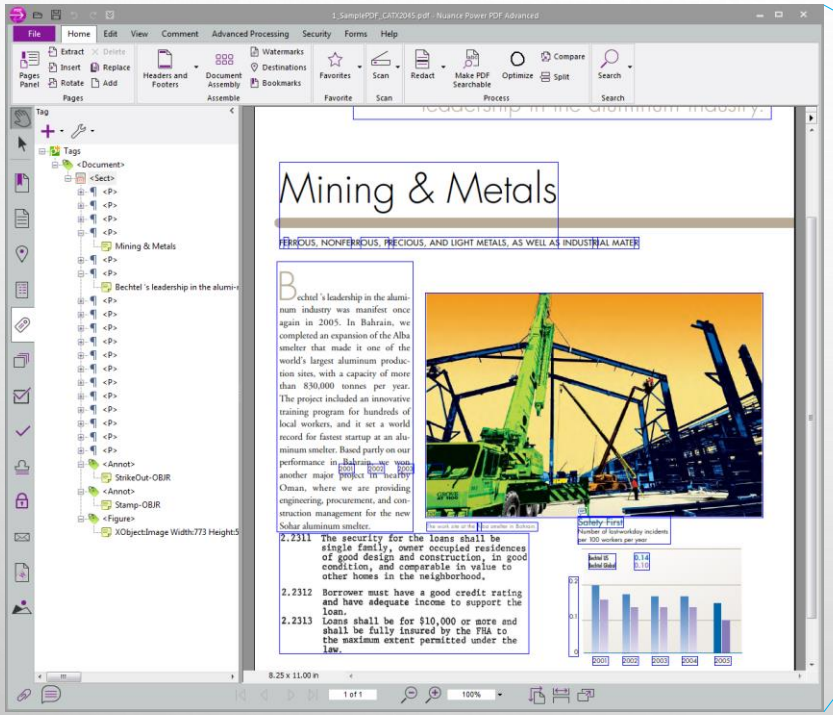
A PDF FÁJL - ALAPISMERETEK



Annotáció - áthúzás



A PDF FÁJL - ALAPISMERETEK



Címkék

HOGY KERÜL A (PDF) CSIZMA AZ (OCR) ASZTALRA?

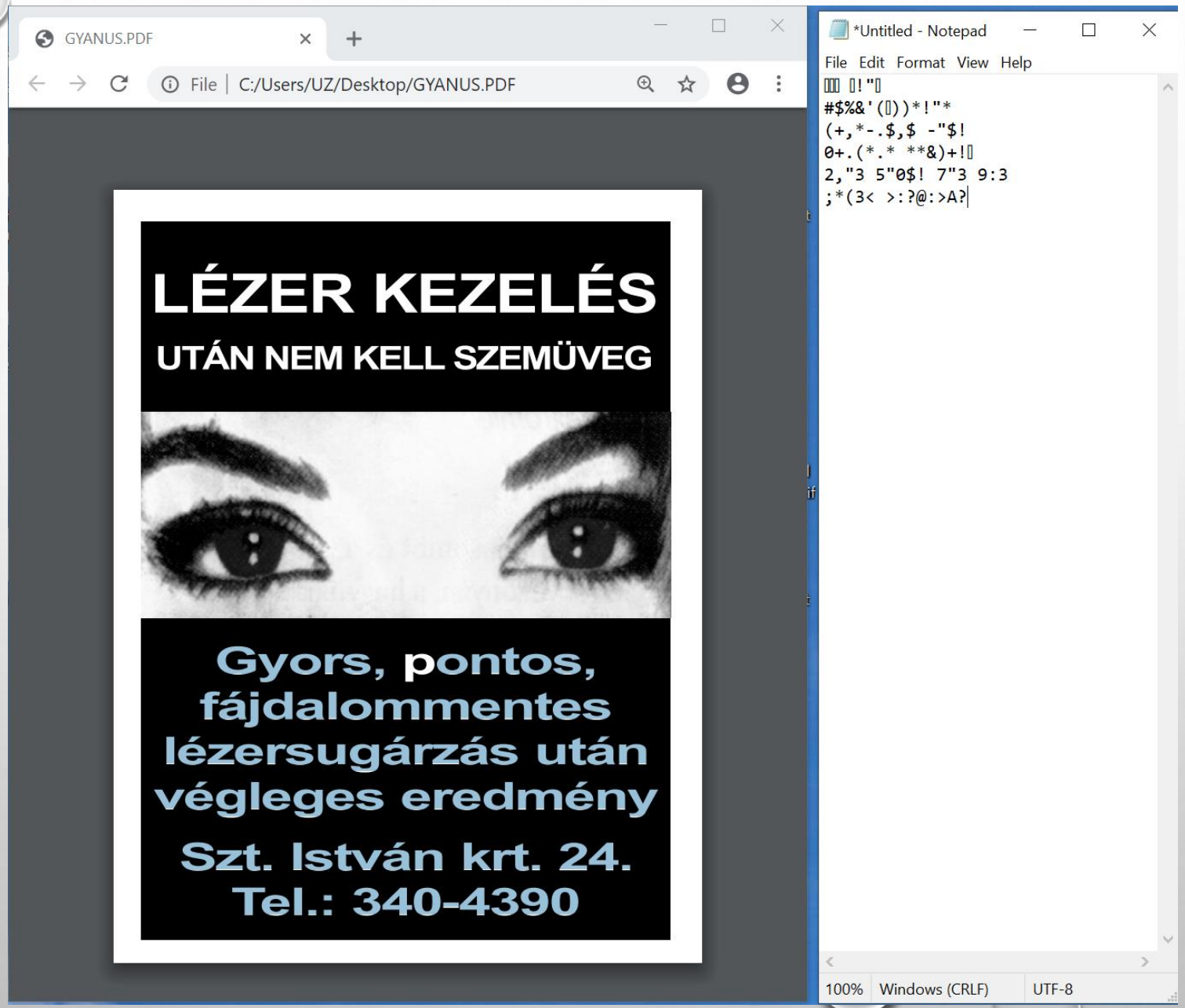
- PDF fájlok kezelésénél felmerülő feladatok
  - PDF készítése
    - Leggyakrabban „nyomtatás” útján
    - Papíralapú dokumentumok digitalizálása
  - PDF feldolgozása
    - Lapok törlése, beszúrása, átrendezése
    - Szöveg kinyerése
      - Gépi feldolgozás számára (pl. számlák)
      - Bemenet gyengénlátók felolvasó szoftvereinek
      - Redaktálás
    - Konvertálás (MS Word DOCX, ePub stb.)
      - „végleges” ⇔ szerkeszthető, tördelhető formátum
    - Címkézés (PDF/A-1a, PDF/UA)

PAPÍRALAPÚ DOKUMENTUMOK DIGITALIZÁLÁSA IMAGE-ON-TEXT PDF FÁJLOK

- Az egyes oldalakon a szkennelt képek látszanak
- A képek „mögött” nem látható módon az OCR által felismert szöveg pozícióhelyesen
- A hagyományos keresés a PDF-kezelő alkalmazásokban a szokott módon működik, mintha nem is a szkennelt képet látnánk
- Másolás / beillesztés is alkalmazható a szöveges részek kiexportálására
- További lehetőség: a szkennelt képek optimalizálása (MRC technológia)
  - Sokkal kisebb méret komoly minőségromlás nélkül
  - Alapja a szöveges és grafikus területek különválasztása (képfeldolgozás!)
    - A szöveges részek nagyobb felbontásban de monokromatikusan tárolva (jobb tömöríthetőség)
    - A grafikák kisebb felbontásban, veszteségesen tömörítve

SZÖVEGKINYERÉS, KONVERTÁLÁS, CÍMKÉZÉS

- Nem minden szöveg, ami annak látszik (pl. rasztergrafika feljebb)
- A PDF fájlok megjelenítéséhez nem szükséges a szövegkarakterek azonosíthatósága
- A PDF fájlok megjelenítéséhez nem szükséges a szövegrészek sorrendjének az ismerete
- A PDF fájlok nem kell, hogy tartalmazzanak logikai struktúrát
  - bekezdés, fejléc, lábléc, táblázatok, hasábok stb.
  - konvertálásnál, címkézésnél ezeket kell megtalálni és visszaadni a kimeneten
- A fentiek kezeléséhez szükség van képfeldolgozási és felismerési algoritmusokra!


GYANUS.PDF
File | C:/Users/UZ/Desktop/GYANUS.PDF
LÉZER KEZELÉS
UTÁN NEM KELL SZEMÜVEG
Gyors, pontos,
fájdalommentes
lézersugárzás után
végleges eredmény
Szt. István krt. 24.
Tel.: 340-4390
*Untitled - Notepad
File Edit Format View Help
#$%&'()*!"*
(+,*-.$,$ -"$!
0+.(*.* **&)+!
2,"3 5"0$! 7"3 9:3
;*(3< >:?@:>A?
100% Windows (CRLF) UTF-8

SZÖVEGKINYERÉS, KONVERTÁLÁS, CÍMKÉZÉS

- Nem minden szöveg, ami annak látszik (pl. rasztergrafika feljebb)
- A PDF fájlok megjelenítéséhez nem szükséges a szövegkarakterek azonosíthatósága
- A PDF fájlok megjelenítéséhez nem szükséges a szövegrészek sorrendjének az ismerete
- A PDF fájlok nem kell, hogy tartalmazzanak logikai struktúrát
  - bekezdés, fejléc, lábléc, táblázatok, hasábok stb.
  - konvertálásnál, címkézésnél ezeket kell megtalálni és visszaadni a kimeneten
- A fentiek kezeléséhez szükség van képfeldolgozási és felismerési algoritmusokra!

KÖSZÖNÖM A FIGYELMET!