**Szakdolgozat**



**Miskolci Egyetem**

**Interaktív megjelenítő eszköz pénzügyi adatok elemzéséhez**

**Készítette:**

#### Bencze Zsombor Programtervezőinformatikus

**Témavezető:**

#### Piller Imre

**Miskolc, 2023**

##### Miskolci Egyetem

Gépészmérnöki és Informatikai Kar

Alkalmazott Matematikai Intézeti Tanszék **Szám:**

**Szakdolgozat Feladat**

Bencze Zsombor (LP5J4B) programtervezőinformatikus jelölt részére.

**A szakdolgozat tárgyköre:** webalkalmazás, grafikus eszközök

**A szakdolgozat címe**: Interaktív megjelenítő eszköz pénzügyi adatok elemzéséhez

##### A feladat részletezése:

*A pénzügyi adatok (például tőzsdei árfolyamok) elemzéséhez elengedhetetlen, hogy a rendelkezésre álló információk a szakértők számára könnyen áttekinthető formában rendelkezésre álljanak.*

*A dolgozat célja olyan interaktív, dinamikus megjelenítési módok tervezésének, működésének és használatának a bemutatása, amelyekkel egyazon idősor különböző részei, különböző forrásból származó idősorok, az azokból származtatott értékek összehasonlíthatók, az aggregáláshoz használt paraméterek rugalmasan változtathatók.*

*Az alkalmazás elkészítése JavaScript nyelven történik, a megjelenítés és az események kezelése HTML5 és CSS elemeken keresztül kerül megvalósításra.*

**Témavezető:**Piller Imre (egyetemi tanársegéd)

##### A feladat kiadásának ideje:

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

szakfelelős

**Tartalomjegyzék**

1. Bevezetés 1
2. Elektronikus dokumentumok 2
   1. [Átvihetődokumentum formátumok . . . . . . . . . . . . . . . . . . . . 2](#_TOC_250021)
   2. [A dokumentum tartalma . . . . . . . . . . . . . . . . . . . . . . . . . . 3](#_TOC_250020)
   3. [Optikai karakterfelismerés . . . . . . . . . . . . . . . . . . . . . . . . . 3](#_TOC_250019)
   4. [Gépi tanulási módszerek . . . . . . . . . . . . . . . . . . . . . . . . . . 4](#_TOC_250018)
3. A dokumentum strukturális elemzése 5
   1. [A PDF dokumentum beolvasása . . . . . . . . . . . . . . . . . . . . . . 5](#_TOC_250017)
   2. [Margók becslése . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6](#_TOC_250016)
   3. [Paragrafusokra bontás . . . . . . . . . . . . . . . . . . . . . . . . . . . 8](#_TOC_250015)
   4. [Szavakra bontás 10](#_TOC_250014)
   5. [Karakter szintűelemzés 11](#_TOC_250013)
   6. [Bonyolultabb dokumentumok 12](#_TOC_250012)
   7. [Képek és szövegek megkülönböztetése 14](#_TOC_250011)
   8. [Karakterek felismerése 15](#_TOC_250010)
   9. [Küszöbértékek meghatározása 18](#_TOC_250009)
   10. [Képfelismerés neurális háló segítségével 19](#_TOC_250008)
   11. [Program tesztelése 19](#_TOC_250007)
       1. [Futási időbecslése 19](#_TOC_250006)
       2. [A paragrafusok feldolgozásának a helyessége és a szöveg helyes- ségének a viszonya 20](#_TOC_250005)
4. Partícionálás és klaszterezés 23
   1. [Oldalak partícionálása 23](#_TOC_250004)
   2. [Összefüggőképterületek detektálása 26](#_TOC_250003)
5. A vizsgálatokhoz készített programok 30
   1. [Képek vágása és további feldolgozása 30](#_TOC_250002)
   2. [Jupyter munkafüzetek használata 31](#_TOC_250001)
6. Összefoglalás 32

[Irodalomjegyzék 33](#_TOC_250000)

**Bevezetés**

A dokumentumaink egy jelentős része elektronikus formában érhetőel. Ennek közked- velt formátuma a PDF (*Portable Document Format*), amely segítségével különböző eszközökön közel azonos formában lehet megjeleníteni a tartalmakat.

A dokumentumok szerkezete lehet nagyon egyszerű, de igen komplikált is. A dol- gozat olyan módszereket mutat be, amelyek a képek strukturális elemeit ismerik fel. Alapvetően olyan elemekre gondolhatunk, mint a bekezdések, sorok és karakterek.

Az elemzésnek két főalternatívája lehet. Az egyikben a PDF API-k használatával lehetne kinyerni a fájlokból a dokumentum adatait. Másik lehetőségként a PDF kép- pé alakítása, majd képfeldolgozási módszerekkel való elemzése jöhet szóba. A dolgozat az utóbbit választja több okból kifolyólag. Ezt elsősorban az indokolja, hogy az így nyert tapasztalatok, elkészített algoritmusok felhasználhatóvá válnak majd szkennelt dokumentumok vizsgálatára is. A PDF egy elég bonyolult bináris formátum, de nem az egyetlen amelyet használnak. Azzal, hogy a feldolgozás elsőlényegi lépése a doku- mentum képpé alakítása lehetővé válik majd más formátumoknak a feldolgozása is, amennyiben azokat is képpé lehet konvertálni.

A vizsgálatokhoz és az algoritmusok fejlesztéséhez a Python programozási nyelv került kiválasztásra. Ehhez szabadon elérhetőek a képfeldolgozáshoz, adatelemzéshez, azok eredményeinek a megjelenítéséhez szükséges eszközök. Szerencsére ezek jól doku- mentáltak, rengeteg példát találni a használatukra.

A Python programozási nyelv tervezésénél szempont volt, hogy jól áttekinthető, könnyen olvasható és értelmezhetőlegyen a programkód. Ez lehetővé tette, hogy az algoritmusok bemutatása közvetlenül forrásód formájában szerepeljen a dolgozatban. Speciális jelölések és pszeudókód értelmezésére vonatkozó magyarázatok helyett köz- vetlenül láthatjuk, hogy hogy működnek maguk az algoritmusok.

A dolgozatban a különbözőszintűstrukturális elemek felismerésének módja és an- nak hatékonysága kerül bemutatásra. Az algoritmusok eredményei Python objektumok formájában válnak elérhetővé, amelyek a későbbi felhasználáshoz átalakíthatók lesznek majd XML vagy JSON formátumba is. A konverzióhoz, tároláshoz szükséges függ- vénykönyvtárak a Python kiadásoknak a részét képezik, így nem tünt célszerűnek a dolgozatban arra részletesen kitérni. A hangsúlyt a dolgozat az algoritmizálásra és azok eredményeinek a kiértékelésére helyezi.

**Elektronikus dokumentumok**

Az elektronikus dokumentumok fokozatosan kezdik felváltani a papír alapú megfele- lőiket. Teljesen nyilvánvaló, könnyen hozzáférhetőeszközök segítségével elvégezhető, hogy ilyen dokumentumokat megtekintsünk, létrehozzunk, szerkesszünk vagy valami- lyen módon megosszunk.

A fejezetben bevezetésképpen azt tekintjük át, hogy milyen elterjedt és szabvá- nyosított dokumentum formátumok vannak. A leírás célja ezen túlmenően az is, hogy megindokolja, hogy miért éppen a PDF formátumra esett a választás.