



PDF feldolgozás szemantikus keresőrendszerhez

Szakmai beszámoló

a szerző neve: Urbán Eszter Klára
a témavezető neve: Lakatos Róbert

DEBRECENI EGYETEM
Informatikai Kar
Debrecen, 2024

Tartalom

| | |
|---------------------------------------|-----------|
| Tartalom | 3 |
| 1 Bevezetés | 1 |
| 2 Előfeldolgozás | 3 |
| 2.1 PyPDF2 | 4 |
| 2.2 PDFMiner | 6 |
| 2.3 GROBID | 8 |
| 3 Struktúra kialakítása | 10 |
| 3.1 Meta adatok | 10 |
| 3.2 Bekezdések kinyerése | 11 |
| 3.3 Táblázatok feldolgozása | 12 |
| 3.4 JSON formátumok | 14 |
| 4 Interfész | 15 |
| 5 Eredmények | 17 |
| 6 Összefoglalás | 18 |
| Hivatkozások | 19 |
| Függelék | 21 |

Bevezetés

A generatív nyelvi modellek megjelenése új lehetőségeket nyitott a magasabb szintű felhasználói interakciót lehetővé tevő tudásbázis rendszerek fejlesztése előtt. Az ilyen rendszerekre példa az OpenAI [1] ChatGPT [2] vagy a Google [3] Gemini [4] szolgáltatása valamint a Microsoft [5] Bing [6] új típusú kereső rendszere. Azonban ezen rendszerek teljesítményét nagy mértékben meghatározza a nyelvi modellek számára rendelkezésre álló információ. Ezért az ilyen típusú rendszerek megfelelő minőségű működéséhez elengedhetetlen a nyelvi modellek finom hangolása, illetve egy szemantikus kereső rendszerrel történő összekapcsolása. A finomhangolás feladata, hogy a nyelvi modell kellően jól megtanulja a nyelvet és a különböző nyelvi feladatokat, mint például kérdés-válasz, szövegösszefoglalás vagy a szövegkiegészítés. Továbbá a finom hangolás során a nyelvi modellek elsajátítják a tanító adathalmaz által reprezentált tudásanyagot is.

Azonban a finom hangolás a nyelvi modellek képzésénél jelenleg magas számítási igénnyel jár és gyakori jelenség az úgynevezett hallucináció is, ami torzítja a modell által elsajátított tudás anyag visszaadásának pontosságát. Ezen problémák orvoslására a nyelvi modellekkel történő kommunikációnak a kontextusát kiegészítő információval látják el. A kiegészítő információ előállításához pedig egy szemantikusan kereshető adatbázist használnak. A szemantikusan kereshető adatbázisok gyors és pontos információ kinyerést tesznek lehetővé és folyamatos frissítésük és karbantartásuk nem jár nagy számítási költséggel. Ezen adatbázisok egyetlen gyenge pontja a bennük tárolt információk minősége.

Szakmai beszámoló anyagomban egy olyan szoftveres megoldást mutatunk be, amely segítségével a nehezen kezelhető tudományos anyagokat tartalmazó PDF [7] dokumentumokból kinyerem a releváns információkat oly módon, hogy azok utána a szemantikusan kereshető adatbázisba egyszerűen betölthető legyen. A 2. fejezetben bemutatom az általunk megvizsgált jelen-

leg elérhető szoftveres megoldásokat. A 3. fejezetben megmutatom, hogy a megvizsgált megoldásokra támaszkodva hogyan állítottuk elő saját megoldásunkat. Az 4. fejezetben leírom az általunk fejlesztett feldolgozó rendszerrel való kapcsolódást lehetővé tevő, külön erre a feladatra kialakított interfésziünket. A 5. fejezetben bemutatom megoldásunk eredményeit. Végezetül pedig a 6. fejezetben összefoglalom szakmai munkám legfontosabb részeit és hogy az általam készített AI alapú PDF feldolgozó rendszer hogyan járul hozzá a Debreceni Egyetem Informatikai Karának FIRCC projekt keretei között fejlesztésre kerülő AI alapú szakértő rendszerének (debai) jobb működéséhez.

Előfeldolgozás

A debai projektünk célja egy olyan szakértő rendszer létrehozása, amely hasonló képességekkel rendelkezik, mint a Google, OpenAI vagy a Microsoft szolgáltatásai. Azonban az Informatikai Kar saját infrastruktúráján fut, valamint rendelkezik olyan egyedi tudás anyaggal, ami a Kar saját tudományos gyűjteménye. Ezen dokumentumokra példa a 2.1. ábrán megtekinthető néhány példa, amely szemlélteti a tudományos szakirodalmak szerkezetének összetettségét.

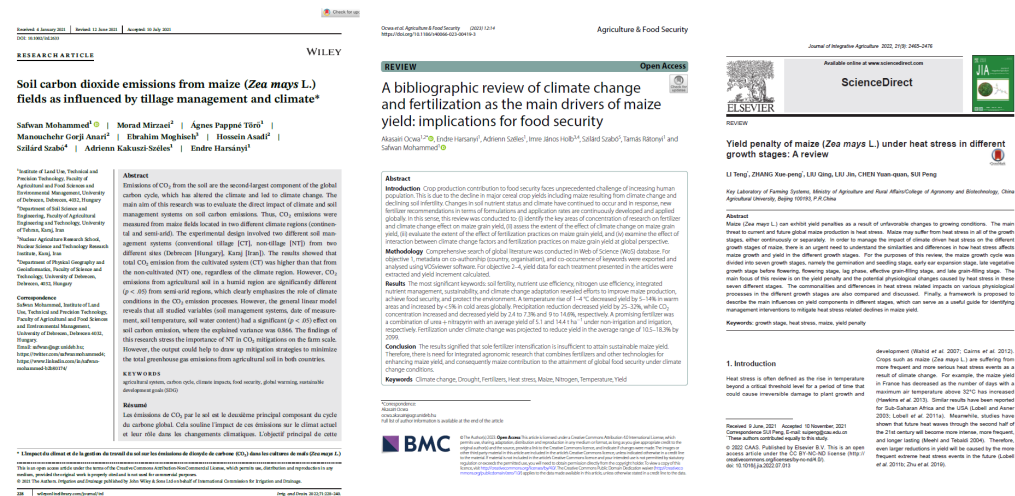


Figure 2.1: Peldák a feldolgozandó publikációkra.

Egy ilyen rendszer elengedhetetlen eleme, egy olyan modul, amely a tudásanyagokat tartalmazó PDF dokumentumok hatékonyan tudja feldolgozni. Ennek megvalósításához a népszerű PyPDF2 [8], PDFMiner [9] Python programozási nyelvhez [10] írt program csomagokat, valamint Generation

of Bibliographic Data (GROBID) [11] PDF feldolgozó rendszert vizsgáltuk meg. Kiemelt célunk volt, hogy a dokumentumokban lévő nehezen feldolgozható tudományos szerkesztési stílust alkalmazó táblázatokat (lásd 2.2. ábra) is össze tudjuk gyűjteni.

| Rank | Country | Documents | Total link strength |
|------|------------------------|-----------|---------------------|
| 1 | Peoples Republic China | 99 | 55 |
| 2 | USA | 81 | 64 |
| 3 | Germany | 43 | 37 |
| 4 | Kenya | 29 | 26 |
| 5 | Australia | 24 | 21 |
| 6 | India | 23 | 17 |
| 7 | Canada | 20 | 16 |
| 8 | England | 18 | 17 |
| 9 | Italy | 18 | 13 |
| 10 | The Netherlands | 16 | 16 |
| 11 | Ethiopia | 14 | 12 |
| 12 | Burkina Faso | 13 | 13 |
| 13 | Mexico | 13 | 13 |
| 14 | Pakistan | 12 | 11 |
| 15 | France | 11 | 11 |
| 16 | Spain | 11 | 9 |
| 17 | Ghana | 10 | 10 |
| 18 | Zimbabwe | 10 | 9 |
| 19 | Denmark | 9 | 7 |
| 20 | Mali | 9 | 9 |

MFCCSY: maize, fertilizer, climate change, soil, and yield

Figure 2.2: Példa egy táblázatra a feldolgozandó publikációkból.

2.1 PyPDF2

A PyPDF2 egy olyan program csomag, amely szöveges formátumban adja vissza a PDF tartalmát. A PyPDF2 esetében a tesztelés során a következő megállapításokat tettük:

1. A szöveget nem bekezdésekre tördeli, hanem, soronként választja el.
2. Amikor a szöveg oszlopszerűen egymás mellett halad azt tudja kezelni, hogy ne folyjon össze a szöveg. Viszont ebben az esetben sem bekezdéssenként tördelte azt.
3. Az élőfej és élőláb kezelése úgy történik, hogy minden oldalról ezeket a szövegtörzs elé teszi.

4. A szövegben többször előforduló indexeket nem tudja ugyanúgy viszaadni, hanem a számot csak a szavak után írja.
5. A tartalomjegyzékben való szereplés jelölését, ahogy az 2.3. ábrán látható egyáltalán nem tudja lekezelni, mint ahogy a képeket sem, de nem is jelöli semmilyen módon, hogy képnek kéne ott elhelyezkednie.

| | |
|---|---|
| <p>Ócwa et al. Agriculture & Food Security (2023) 12:14 https://doi.org/10.1186/s40066-023-00419-3 REVIEW Open Access</p> <p>© The Author(s) 2023. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data. Agriculture & Food Security</p> <p>A bibliographic review of climate change and fertilization as the main drivers of maize yield: implications for food security Akasairi Ocwa1,2* , Endre Harsanyi1, Adrienn Szélesi1, Imre János Holb3,4, Szilárd Szabó5, Tamás Rátönyi1 and Safwan Mohammed1</p> <p>Abstract Introduction Crop production contribution to food security faces unprecedented challenge of increasing human population. This is due to the decline in major cereal crop yields including maize resulting from climate change and declining soil infertility. Changes in soil nutrient status and climate have</p> | <p>RESEARCH ARTICLE Soil carbon dioxide emissions from maize (Zea mays L.) fields as influenced by tillage management and climate* Safwan Mohammed1 Morad Mirzaei2 /C19Agnes Pappné Törjői1 Manouchehr Gorji Anari2 Ebrahim Moghiseh3 Hossein Asadi2 Szil C19ard Szab C19o4 Adrienn Kakuszi-Szélesi1 Endre Hars C19anyi1 1Institute of Land Use, Technical and Precision Technology, Faculty of Agricultural and Food Sciences and Environmental Management, University of Debrecen, Debrecen, 4032, Hungary 2Department of Soil Science and Engineering, Faculty of Agricultural Engineering and Technology, University of Tehran, Karaj, Iran 3Nuclear Agriculture Research School, Nuclear Science and Technology Research Institute, Karaj, Iran 4Department of Physical Geography and Geoinformatics, Faculty of Science and Technology, University of Debrecen, Debrecen, 4032, Hungary</p> <p>Correspondence Safwan Mohammed, Institute of Land Use, Technical and Precision Technology, Faculty of Agricultural and Food Sciences and Environmental Management, University of Debrecen, Debrecen 4032, Hungary. Email: safwan@agr.unideb.hu; https://twitter.com/safwanmohammed4; https://www.linkedin.com/in/safwan-mohammed-b2b80174/Abstract</p> <p>Emissions of CO2 from the soil are the second-largest component of the global carbon cycle, which has altered the climate and led to climate change. The main aim of this research was to evaluate the direct impact of climate and soil management systems on soil carbon emissions. Thus, CO2 emissions were measured from maize fields located in two different climate regions (continental and semi-arid). The experimental design involved two different soil man-</p> |
|---|---|

Figure 2.3: PyPDF2 hibái.

6. A táblázatok rossz minőségben tudja feldolgozni. A fejléct az előtte lévő mondathoz írja hozzá. Az adatokat, amik egy sorban szerepelnek a táblából viszont szépen egymás mellé rendezi. Nagyobb táblázatoknál azonban ezt sem tudja tartani és a kinyert adatok nem lesznek megfelelő elhelyezkedésűek. (lásd 2.4. ábra)

```
MFCCSY: maize, fertilizer, climate change, soil, and yieldRank Country Documents Total link
strength
1 Peoples Republic China 99 55
2 USA 81 64
3 Germany 43 37
4 Kenya 29 26
5 Australia 24 21
6 India 23 17
7 Canada 20 16
8 England 18 17
9 Italy 18 13
10 The Netherlands 16 16
11 Ethiopia 14 12
12 Burkina Faso 13 13
13 Mexico 13 13
14 Pakistan 12 11
15 France 11 11
16 Spain 11 9
17 Ghana 10 10
18 Zimbabwe 10 9
19 Denmark 9 7
20 Mali 9 9Page 5 of 18
Ocw a et al. Agriculture & Food Security (2023) 12:14
```

Figure 2.4: PyPDF2-vel kinyert táblázat.

2.2 PDFMiner

A PDFMiner viselkedése nagyon hasonló a PyPDF2-hez. A PDFMiner esetében a tesztelés során a következő megállapításokat tettük:

1. A szöveg nem bekezdésekre bontja fel. Ahol PDF fájlban a sorok véget érnek ott érnek véget az egyes részek is. Viszont a PDFMiner a PyPDF2-vel ellentétben tudja értelmezni, hogy egy bekezdésnek mikor van vége és ott hagy egy entert a kettő között.
2. Az oszlopszerű tagolást ugyanúgy kezeli, mint a PyPDF2.

3. Az élőfej a szövegtörzs elé az élőláb pedig az adott oldalon található szöveg után kerül az outputba.
4. A képek és a hivatkozások ki vannak hagyva az outputból, semmi jelölés nem köti őket oda.
5. Ahogy az a 2.5. ábrán is látszik az indexeket sem tudja megfelelő módon lehivatkozni.

| | |
|--|---|
| <p>Öcwa et al. Agriculture & Food Security (2023) 12:14 https://doi.org/10.1186/s40066-023-00419-3</p> <p>Agriculture & Food Security</p> <p>REVIEW</p> <p>Open Access</p> <p>A bibliographic review of climate change and fertilization as the main drivers of maize yield: implications for food security Akasairi Ocwa1,2* Safwan Mohammed1</p> <p>, Endre Harsanyi1, Adrienn Széles1, Imre János Holb3,4, Szilárd Szabó5, Tamás Rátonyi1 and</p> <p>Abstract Introduction Crop production contribution to food security faces unprecedented challenge of increasing human population. This is due to the decline in major cereal crop yields including maize resulting from climate change and declining soil infertility. Changes in soil nutrient status and climate have continued to occur and in response, new fertilizer recommendations in terms of formulations and application rates are continuously developed and applied globally. In this sense, this review was conducted to: (i) identify the key areas of concentration of research on fertilizer and climate change effect on maize grain yield, (ii) assess the extent of the effect of climate change on maize grain yield, (iii) evaluate the extent of the effect of fertilization practices on maize</p> | <p>Received: 4 January 2021</p> <p>Revised: 12 June 2021</p> <p>Accepted: 10 July 2021</p> <p>DOI: 10.1002/ird.2633</p> <p>R E S E A R C H A R T I C L E</p> <p>Soil carbon dioxide emissions from maize (Zea mays L.) fields as influenced by tillage management and climate*</p> <p>Safwan Mohammed1 Manouchehr Gorji Anari2 Szil(cid:1)ard Szab(cid:1)o4 Morad Mirzaei2 (cid:1)Agnes Pappné Törjoi1 Ebrahim Moghiseh3 Hossein Asadi2 Adrienn Kakuszi-Széles1</p> |
|--|---|

Figure 2.5: PDFMiner hibái.

6. Gyakran nem determinisztikus a viselkedése. Ugyanis (lásd 2.6. ábra) a bekezdések egy részét kihagyja és azt a mondatrészt az adott bekezdés után helyezi el a kimenetbenben.

Results The most significant keywords: soil fertility, nutrient use efficiency, nitrogen use efficiency, integrated nutrient management, sustainability, and climate change adaptation revealed efforts to improve maize production, achieve food security, and protect the environment. A temperature rise of 1–4 °C decreased yield by 5–14% in warm areas and increased by < 5% in cold areas globally. Precipitation reduction decreased yield by 25–32%, while CO₂ concentration increased and decreased yield by 2.4 to 7.3% and 9 to 14.6%, respectively. A promising fertilizer was 1 under non-irrigation and irrigation, a combination of urea respectively. Fertilization under climate change was projected to reduce yield in the average range of 10.5–18.3% by 2099.

nitrapyrin with an average yield of 5.1 and 14.4 t ha⁻¹

Figure 2.6: A mondat egy részének rossz helyre illesztése.

7. A táblázatok nem tudja megfelelően kinyerni ugyanis nem tudja sorokba rendezni az adatokat, hanem minden egyes táblabeli értéket egymás alá ír. Ezt a jelenséget a 2.7. ábra szemlélteti.

| Rank | Country | Documents | Total link strength | |
|------|------------------------|-----------|---------------------|------------------------|
| 1 | Peoples Republic China | 99 | 55 | Peoples Republic China |
| 2 | USA | 81 | 64 | USA |
| 3 | Germany | 43 | 37 | |
| 4 | Kenya | 29 | 26 | Germany |
| 5 | Australia | 24 | 21 | |
| 6 | India | 23 | 17 | |
| 7 | Canada | 20 | 16 | Kenya |
| 8 | England | 18 | 17 | |
| 9 | Italy | 18 | 13 | |
| 10 | The Netherlands | 16 | 16 | |
| 11 | Ethiopia | 14 | 12 | Australia |
| 12 | Burkina Faso | 13 | 13 | |
| 13 | Mexico | 13 | 13 | |
| 14 | Pakistan | 12 | 11 | India |
| 15 | France | 11 | 11 | |
| 16 | Spain | 11 | 9 | |
| 17 | Ghana | 10 | 10 | Canada |
| 18 | Zimbabwe | 10 | 9 | |
| 19 | Denmark | 9 | 7 | |
| 20 | Mali | 9 | 9 | England |

MFCCSY: maize, fertilizer, climate change, soil, and yield

Figure 2.7: Eredeti táblázat és PDFMiner-rel kinyert táblázat.

2.3 GROBID

A GROBID PDF feldolgozó egy gépi tanulás alapú nyílt forrású rendszer, amelyet kifejezetten tudományos irodalomból származó adatok kinyerésére

és feldolgozására tervezték. A fejlesztők célja a GROBID létrehozásával a nagy mennyiségű fájlok hatékony feldolgozása illetve a félig strukturáltan PDF adatokat strukturált TEI (Text Encoding Initiative) [12] formátumú adattá alakítsa volt. A GROBID lehetővé teszi széles körű dokumentumelemek felismerését és a hozzájuk tartozó bibliográfiai adatok kinyerését.

A GROBID rendszerét összesen 111 darab tudományos PDF dokumentumra teszteltük. A GROBID esetében a futási idő átlagban nagyjából 183 másodpercet vesz igénybe. Ezeket az átalakított PDF-eket egy mappába írja ki XML [13] formátumban. A GROBID-ot úgy állítottuk be, hogy az egész dokumentumot feldolgozza, valamint a meta adatokat is, amik a dokumentumokat leíró adatokat tartalmazzák.

A tesztelés eredményeképpen megállapítottuk, hogy a GROBID hatékonyabban nyeri ki a szükséges információkat, mint a PyPDF2 és a PDFMiner. Ezért az adatkinyerés folyamatát a GROBID-ra építettük rá.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns:space="preserve" xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.tei-c.org/ns/1.0 https://raw.githubusercontent.com/kermitt2/grobid/master/grobid-home/schemas/xsd/Grobid.xsd"
  xmlns:xlink="http://www.w3.org/1999/xlink">
  <teiHeader xml:lang="en">
    <fileDesc>
      <titleStmnt>
        <title level="a" type="main">A bibliographic review of climate change and fertilization as the main drivers of maize yield: implications for
food security</title>
      <funder ref="#_EPeMmux">
        <orgName type="full">National Research, Development, and Innovation Fund</orgName>
      </funder>
      <funder ref="#_hqTgg6P">
        <orgName type="full">University of Debrecen</orgName>
      </funder>
      <funder>
        <orgName type="full">Ministry of Innovation and Technology of Hungary</orgName>
      </funder>
    </titleStmnt>
    <publicationStmnt>
      <publisher>Springer Science and Business Media LLC</publisher>
      <availability status="unknown"><p>Copyright Springer Science and Business Media LLC</p>
    </availability>
      <date type="published" when="2023-06-02">2023-06-02</date>
    </publicationStmnt>
    <sourceDesc>
      <biblStruct>
        <analytic>
          <author role="corresp">
            <persName><forename type="first">Akasairi</forename><surname>Ocwa</surname></persName>
            <email>ocwa.akasairi@agr.unideb.hu</email>
            <idno type="ORCID">0000-0003-4787-9270</idno>
          </author>
          <author>
            <persName><forename type="first">Endre</forename><surname>Harsanyi</surname></persName>
          </author>
        </analytic>
      </biblStruct>
    </sourceDesc>
  </teiHeader>

```

Figure 2.8: Példa a GROBID által feldolgozott XML-re.

Struktúra kialakítása

A GROBID hatékony eszköz a PDF fájlok feldolgozására azonban az általa generált kimenet még továbbra sem elégséges ahhoz hogy közvetlenül betöltsük a szemantikusan kereshető adatbázisunkban. Ezért a GROBID által készített kimenet további finomítására saját szövegtisztító modult fejlesztettünk. Saját megoldásunk kimenetének a JSON [14] fájl formátumot választottuk ugyanis az általunk használt FAISS (Facebook AI Similarity Search) [15] szemantikus keresést megvalósít adatbázis a JSON formátum feldolgozásával könnyen boldogul. A JSON létrehozásához a dokumentumokat három részre bontottuk. Egy meta rész, a bekezdések és a táblázatok. A következő fejezetekben ezen részek előállítását mutatjuk be.

Algorithm 1 PDF to XML converter

Require: *server_host, pdf_folder, xml_folder, json_folder*

Ensure: XML files converted from PDF

if $\neg \text{EXISTS}(\text{xml_folder})$ **then**

 CREATEDIRECTORY(xml_folder)

end if

 ▷ Makes a xml folder if needed

 Initialize GROBID with the right datas

 ▷ Uses GROBID

3.1 Meta adatok

A meta rész tartalmazza az cikk címét, készítőit, a megjelenési helyet és a dátumát. Ezeknél mind meg kellett vizsgálnunk, hogy a TEI formátumban hol helyezkednek ezen információk és azt kinyerni onnan. A megjelenési dátumnál a formázással is kellett foglalkozni, hogy az minden esetben egységes legyen. Az általunk választott formázási mód a *Év. Hónap. Nap.* lett amire minden egyéb formázási módot lecseréltünk. A szerzőknél a kereszt és a

vezetéknevet egyaránt meg kell adni így ezt is egy formázás követte, hogy megfelelően legyenek kiírva és azok elválasztva egy vessző segítségével. Ha a dokumentumnak valamelyik adata esetleg nem volt megadva akkor azt a „Nincs találat” értékkel kezeltük.

Algorithm 2 Process Metadata

Require: *xml_data*

Ensure: Get the metadata from the XML

title \leftarrow title

date \leftarrow date

journal \leftarrow journal

authors \leftarrow []

if *pers_name* is not empty **then**

forenames \leftarrow *pers_name*['forename']

surname \leftarrow *pers_name*['surname']

full_name \leftarrow concatenate *forename*['text'] with *surname* for each *forename* in *forenames*

if *authors* is not empty **then**

 concatenate ',' with *full_name* and append to *authors*

else

authors \leftarrow *full_name*

end if

end if

▷ Get the full name for the authors

return *title, date, journal, authors*

3.2 Bekezdések kinyerése

A bekezdések kinyerésénél figyelembe kellett vennünk, hogy van címe is egy bekezdésnek és egy címhez több szöveg blokk is tartozhat. Az absztraktokat a GROBID alapértelmezetten nem helyezi a bekezdések közé viszont az is tartalmazhat hasznos információkat így azt is a bekezdések közé szereztük be. Mivel maga a szöveg tartalmaz hivatkozásokat azonban azok nem releváns információk ezért ezeket a szögletes zárójelben lévő számokat kitisztítottuk a bekezdésekből a tisztább eredmény érdekében. A táblákat minden egyes szöveg blokkhoz külön írtuk ki. Ez azt jelenti, hogy ha egy bekezdés tartalmaz táblát az úgy jelenik meg a szövegben, hogy „Table n”, ahol az n az adott táblázat számát jelenítette meg. Ezt a részt a tisztítás

szempontjából átszerveztettük úgy, hogy helyette a tábla címe legyen olvasható a szövegben.

Algorithm 3 Process Paragraphs

Require: *xml_data*

Ensure: Get the paragraphs from the XML

```

paragraphs  $\leftarrow$  []
par_title  $\leftarrow$  title
par_text  $\leftarrow$  []
for p in text do
    original_text  $\leftarrow$  p.TEXT.STRIP()
    match  $\leftarrow$  search for 'Table n' in original_text
    if match then
        table_number  $\leftarrow$  integer part of match.GROUP(1)
        for table in table_data do
            if table['index'] == table_number and table['title'] then
                modified_text  $\leftarrow$  replace 'Table n' in original_text with
                table['title'] and append to paragraphs
            end if
        end for
         $\triangleright$  Replace the 'Table n' with the table title
    else
        modified_text  $\leftarrow$  remove all references with regex expression from
        original_text and append to paragraphs
         $\triangleright$  Remove the references
    end if
end for
title_text  $\leftarrow$  Abstract and append to first place of the paragraphs  $\triangleright$  Get
the abstract
return paragraphs

```

3.3 Táblázatok feldolgozása

A tábláknál az első formázás az volt, hogy a címe, a fejléc cellák és végül a kitöltő adatok szerepelnek. Mivel a dokumentumok tartalmaztak olyan táblákat is, amik nem fértek el egy oldalra a PDF-ben annak a táblának a címe a „continued” és ennek egyéb változatai volt a folytatása során. Az ilyen esetben ezeket a táblákat össze kellett olvasztani, hogy az arra való hivatkozás a bekezdéseknél megfelelően működjön. Ezért a bekezdésben való

hivatkozáshoz egy index értékkel is elláttuk az össze táblát, hogy egyszerűbb legyen ez alapján azonosítani a „Table n” résznél. A vektor adatbázisba való betöltéshez a táblákat át kellett alakítani string formátumú megjelenítésbe. A táblák tartalmazznak unicode-okat amelyek helyes feldolgozására külön ügyelnünk kellett.

Algorithm 4 Process Tables

Require: *xml_data*

Ensure: Get the tables from the XML

```

table_data  $\leftarrow$  []
previous_table_data  $\leftarrow$  None
continued_table_data  $\leftarrow$  None
table_index  $\leftarrow$  1
for table in tables do
    table_title  $\leftarrow$  table title
    table_rows  $\leftarrow$  find all rows in table
    if table_rows is not empty then
        table_head  $\leftarrow$  list of text of cells in first row of table_rows
        table_value  $\leftarrow$  list of lists of text of cells in rows excluding first
row of table_rows ▷ Get the tables data
        if "(continued)" in table_title
            append rows of table_value to rows of previous_table_data
        else
            append table_index, table_title, table_head, table_value as
dictionary to table_data
            table_index  $\leftarrow$  table_index + 1
        end if ▷ Repair of larger tables
    end if
end for
for table in table_data do
    formatted_table  $\leftarrow$  create a dictionary with "title", "head", and
"value" keys using corresponding values from table
end for ▷ Tables for the vector database
return table_data

```

3.4 JSON formátumok

A megtisztított adatokat különböző JSON formátokká konvertáltuk azért, hogy megtudjuk vizsgálni mely formátum a legmegfelelőbb egy mesterséges intelligencia alapú szemantikus keresőhöz illeszkedő vektor adatbázis létrehozásához. Először egy olyan JSON formátumot próbáltunk ki, amiben a bekezdések mondatokra voltak bontva és mindegyikhez egy index tartozott. Egy bekezdésnél pedig úgy lehet azonosítani, hogy mely mondatok tartoznak egybe, hogy az indexek egy „connects” részbe tárolva vannak. Minden PDF-hez külön készült egy ilyen formátumú JSON.

A kialakított JSON formátumról részletes mintákat a beszámoló Függelék részében mutatunk be.

Algorithm 5 Process XML to JSON

Require: *xml_data*

Ensure: Make the json structure

if \neg EXISTS(*json_folder*) **then**

 CREATEDIRECTORY(*json_folder*)

end if

\triangleright Makes a json folder if needed

for *filename* **in** LISTFILES(*xml_folder*) **do**

if *filename*.ENDSWITH('.xml') **then**

 Prepair the XML files

title, date, journal, authors \leftarrow PROCESSMETADATA(*xml_data*)

table_data \leftarrow PROCESSTABLES(*xml_data*)

paragraphs \leftarrow PROCESSPARAGRAPHS(*xml_data*, *table_data*)

formatted_date \leftarrow *date*.REPLACE('-', '.') + '.' \triangleright Make the date

format uniform

result_json \leftarrow {

 "meta": {

 "title": *title* if *title* else "No title found",

 "author": *authors* if *authors* else "No author found",

 "journal": *journal* if *journal* else "No journal found",

 "date": *formatted_date* if *date* else "No date found"

 },

 "paragraphs": *paragraphs*,

 "tables": *table_data*

 }

\triangleright Make the json structure

end if

end for

Interfész

Az interfész elkészítésénél a dokumentumok tárolásához a MinIO [16] objektum tárolót használtunk. Ez egy az Amazon [17] által kiadott S3 kompatibilis minden nyilvános felhőn elérhető objektumtároló. A fájlok tárolásához két tárolót hoztunk létre a *pdfs* és *results bucket*. A *pdfs* amiben a feltöltött PDF cikkek vannak eltárolva. Ezek közül lehet kiválasztani, hogy melyikkel szeretnénk a vektor adatbázist létrehozni. A *results bucket* ahova az interfész a kész átalakított JSON fájlokat hozza létre. Ahhoz, hogy a rendszereink tudjanak egymással kommunikálni a FastAPI-t [18] használtuk. Ez egy magas teljesítményű webes keretrendszer API-k építéséhez Pythonban. Ez segít megnyitni azokat a végpontokat, amiket majd használni fogunk a számunkra megfelelő szerveren.

Az rendszerünk biztonsági adatainak a védelem érdekében a szükséges belépési jelszavak a különböző felületekre, mint a GROBID-ba és a MinIO-ba ki lett szervezve egy konfigurációs fájlba így ezek nem a kódba vannak integrálva hanem egyszerűbben változtatható valamint így a saját adatok is védve vannak. Ilyen adat például a MinIO-nál az elérési URL, a titkos és az engedélyezési kulcs, a bucket-ek nevei és a GROBID szerver címe is.

A PDF feldolgozáshoz egy végpontra volt szükségünk. Itt az első lépésben a S3 kiinduló tárolóból (*pdfs*) az összes adatot az ID-ja alapján beolvastunk és ezt a pdf mappában lokálisan eltároltuk. Ezeket a meghívott PDF feldolgozó osztály, ami magába foglalja a mappák létrehozását, a GROBID működését és az összes JSON formázást átalakítja a kívánt formára és eltárolja a JSON mappába. Ezután a kész JSON fájlokat feltölti a végső S3 tárolóba (*results*) ugyanazzal a névvel ellátva amivel az beolvasásra került.

Algorithm 6 Using a config file and starting FastAPI

Require: *config* ▷ Configuration dictionary

Ensure: initialize MinIO and GROBID variables

Initialize Minio client *s3* with *config['s3Url']*, *config['s3Access']*,
config['s3Secret']

srcBucket \leftarrow *config['srcBucket']*

resultBucket \leftarrow *config['resultBucket']*

server_host \leftarrow *config['grobidHost']*

Start FastAPI service with at host '-' and port xxxx

Algorithm 7 Process endpoint logic

Require: *body* ▷ Request body containing source IDs

Ensure: Process PDFs and store results in S3

pdf_folder \leftarrow 'pdf'

xml_folder \leftarrow 'xml'

json_folder \leftarrow 'json'

▷ Usage folders

for each *sourceID* in *body.source* **do**

 Get PDF data from S3 for *sourceID*

 Save PDF data to *pdf_folder/sourceID.pdf*

end for

▷ Get pdfs from the s3 bucket

Initialize PDF to JSON converter *converter* with *server_host*,
pdf_folder, *xml_folder*, *json_folder*

for each *sourceID* in *body.source* **do**

 Upload JSON result to S3 for *sourceID*

end for

▷ Upload final jsons to the s3 bucket

Eredmények

Az előfeldolgozás folyamán alaposan tanulmányoztuk három különböző PDF feldolgozási módszer működését: PyPDF2, PDFMiner és GROBID. Mindhárom esetben részletesen elemeztük a rendszerek viselkedését, különös tekintettel arra, hogy mennyire hatékonyan tudnak kezelni a tudományos dokumentumokat, beleértve a szövegek struktúráját, táblázatokat és egyéb formázási elemeket.

Az általunk választott GROBID által nyújtott XML alapú adatkinyerési módszertanunk eredménye egy olyan JSON állomány, amelyben megtalálhatóak a releváns információk, amelyek kinyerhetők a PDF-ekből. Ezen információk segítségével hatékony vektor adatbázis létrehozása válik lehetővé, amely ideális alapul szolgálhat mesterséges intelligencia alapú szemantikus keresések számára. A Faiss által támogatott szemantikus keresési funkciók segítségével könnyen és hatékonyan kereshetünk az adatbázisban.

Az interfész elkészítése során sikeresen implementáltuk a MinIO objektum tárolót és a FastAPI-t. Ez a kombináció lehetővé teszi számunkra a PDF feldolgozás és az adatkinyerés folyamatának automatizálását, valamint a kész adatok tárolását és elérését. A konfigurációs fájl segítségével biztosítottuk a rendszer biztonságát és rugalmasságát, hiszen könnyen változtathatóak és védelmezhetőek az érzékeny adatok.

Az elért eredmények alapján sikeresen teljesítettük célkitűzéseinket: egy hatékony szakértő rendszert hoztunk létre, amely képes feldolgozni és kinyerni az információkat a tudományos dokumentumokból. Ennek eredményeként hozzájárulhatunk az Informatikai Kar tudományos gyűjteményének gazdagításához és a kutatók munkájának támogatásához.

Összefoglalás

A DebAI projekt célja egy olyan szakértő rendszer létrehozása, amely ötvözi a OpenAI Google és a Microsoft szolgáltatásainak képességeit. Az Informatikai Kar saját infrastruktúráján fut, és egyedi tudástartalommal rendelkezik, amely a Kar saját tudományos anyagából származik. Az ehhez szükséges PDF dokumentumok hatékony feldolgozásához különböző eszközöket vizsgáltunk, mint a PyPDF2-t, a PDFMiner-t és a GROBID-ot.

A PyPDF2 és a PDFMiner szöveges formátumban adja vissza a PDF tartalmát, de nem tökéletesen kezelik a bekezdéseket, táblázatokat és egyéb strukturált adatokat. A GROBID egy gépi tanulás alapú eszköz, amely hatékonyan képes kinyerni a szükséges információkat a tudományos irodalomból. A tesztelés során megállapítottuk, hogy a GROBID jobban teljesít az adatkinyerés terén, ezért a rendszerünkben ezt használjuk.

Az adatkinyerési folyamat során a PDF dokumentumokat három részre bontottuk: meta adatokra, bekezdésekre és táblázatokra. A meta adatok a címeket, szerzőket és megjelenési adatokat veszi figyelembe. A bekezdéseket és táblázatokat is strukturált formában tároljuk. A táblázatokat is átalakítottuk, hogy a vektor adatbázis hatékonyan kezelhesse azokat.

A GROBID által kinyert adatok alapján létrehoztunk egy JSON állományt, amely tartalmazza a releváns információkat, és alkalmas egy mesterséges intelligencia alapú szemantikus kereső vektor adatbázis létrehozására. Összességében a GROBID segítségével hatékonyan és strukturáltan tudjuk kezelni a PDF dokumentumokban lévő információkat a debai projektben.


Terveink között szerepel továbbá a táblázatok hatékonyabb feldolgozásának fejlesztése, mivel a jelenlegi GROBID rendszer nem mindig képes tökéletesen kezelni azokat. Emellett a bekezdések tisztítása is további finomítást igényel a bekezdésbeli hivatkozások során.


Hivatkozások

- [1] OpenAI, “Openai,” 2024.
- [2] OpenAI, “Chatgpt,” 2024.
- [3] Google, “Google,” 2024.
- [4] G. Team, “Gemini: A family of highly capable multimodal models,” 2023.
- [5] Microsoft, “Microsoft,” 2024.
- [6] Microsoft, “Bing kereső,” 2024.
- [7] I. O. for Standardization, “Document management - portable document format (pdf) - part 2: Extensions,” 2020.
- [8] M. Thoma, “Pypdf2,” 2024.
- [9] Y. Shinyama, “pypdf,” 2024.
- [10] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
- [11] “Grobid.” <https://github.com/kermitt2/grobid>, 2008–2024.
- [12] T. Consortium, “Tei p5: Guidelines for electronic text encoding and interchange,” dec 2016.
- [13] X. W. Group, “Xml fájlformátum,” 2024.
- [14] J. W. Group, *JSON fájlformátum*, 2024.
- [15] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, “The faiss library,” 2024.

- [16] MinIO, Inc., “MinIO: High performance, kubernetes native object storage.” <https://min.io>, 2024.
- [17] Amazon, “Amazon,” 2024.
- [18] S. Ramírez, “FastAPI: Fastapi documentation.” <https://fastapi.tiangolo.com>, 2024.

Függelék

**agronomy**



Article

Analysis of the Content Values of Sweet Maize (*Zea mays* L. Convar Saccharata Koern) in Precision Farming


Cintia Demeter ¹, János Nagy ^{1,*}, László Huzsvai ², Annabella Zelenák ¹, Átala Szabó ¹ and Adrienn Széles ¹

¹ Institute of Land Use, Engineering and Precision Farming Technology, Faculty of Agricultural and Food Sciences and Environmental Management, University of Debrecen, 138 Bősörményi Str., H-4032 Debrecen, Hungary; cintia.demeter@gmail.com (C.D.); zelenak@ag.unideb.hu (A.Z.); szaboatala@ag.unideb.hu (A.S.); szeles@ag.unideb.hu (A.S.)

² Faculty of Economics and Business, Institute of Statistics and Methodology, University of Debrecen, 138 Bősörményi Str., H-4032 Debrecen, Hungary; huzsvai.laszlo@econ.unideb.hu

* Correspondence: nagyjanos@ag.unideb.hu; Tel: +36-06-30-417-1737


Abstract: The global precision farming area is constantly increasing, and precision sweet maize production developed the most. Sweet maize yield is above average in precision farming. Additionally, its role in healthy nutrition is becoming increasingly important due to new hybrids with high carotenoid content. Precision farming techniques are needed to produce healthy food. In particular, nutrient supply and irrigation, sowing, crop management and harvesting need to be carried out with precision techniques. These factors are all prerequisites for effective and healthy growing and processing. The aim was to use the yields of the four sweet maize hybrids grown on the largest area to examine their nutritional values and concentrations (mg kg⁻¹ dry matter) and to analyse their yield per hectare. Concentration is important for the consumer because K, P, Mg, Ca, Fe, Zn, and Na play an important role in metabolism, skin protection, and bone and tooth health. The new results obtained show that the amount of lutein and zeaxanthin per hectare is important for the processing industry, especially for use in food supplements. Their anti-inflammatory effects and their role in disease prevention (cardiovascular diseases, Age-Related Macular Degeneration (AMD)) have been demonstrated. Consumers choose sweet maize mainly on the basis of its palatability, which is why the sugar content of the hybrids was also studied. We assumed that the element concentration in the yield of new hybrids with higher yield per hectare does not decrease with increasing yield. The concentrations of zeaxanthin, β-cryptoxanthin and β-carotene appear in one principal component and they are in close positive correlation with each other. The lutein concentration was independent of the former three compounds. The independence of the lutein concentration means that it is not possible to estimate its amount based on the other three components. For yield per unit area, the correlation is one-dimensional. Yield determines the lutein, zeaxanthin, β-cryptoxanthin and β-carotene concentrations per hectare.

**check for updates**

Citation: Demeter, C.; Nagy, J.; Huzsvai, L.; Zelenák, A.; Szabó, Á.; Széles, A. Analysis of the Content Values of Sweet Maize (*Zea mays* L. Convar Saccharata Koern) in Precision Farming. *Agronomy* **2021**, *11*, 2596. <https://doi.org/10.3390/agronomy11122596>

Received: 28 October 2021
Accepted: 14 December 2021
Published: 20 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: sweet maize; precision farming; minerals; sugars; lutein; zeaxanthin

1. Introduction

Sweet maize is either eaten fresh or used as a raw material for food processing and the industry. Farmers are interested in achieving the highest income per unit area. This in turn depends on who they sell their product to.

Consumers are interested in buying the most nutrient-rich food possible. As the amount of food consumed per day is limited, the physiological effect of more “concentrated” products is better (functional foods). If consumers are willing to pay more for such products, it is preferable to grow hybrids with a high nutrient concentration.

For hybrids grown for the processing industry, however, quantity is the most important factor, as higher yields provide higher revenues. If the aim is to extract a component,

Agronomy **2021**, *11*, 2596. <https://doi.org/10.3390/agronomy11122596> <https://www.mdpi.com/journal/agronomy>


```

{
  "meta": {
    "title": "Analysis of the Content Values of Sweet Maize (Zea mays L. Convar Saccharata Koern) in Precision Farming",
    "author": "Cintia Demeter, János Nagy, László Huzsvai, Annabella Zelenák, Atala Szabó, Adrienn Széles",
    "journal": "MDPI AG",
    "date": "2021. 12. 20."
  },
  "paragraphs": [
    {
      "title": "Abstract",
      "div": [
        {
          "text": "The global precision farming area is constantly increasing, and precision sweet maize production developed the most. Sweet maize yield is above average in precision farming. Additionally, its role in healthy nutrition is becoming increasingly important due to new hybrids with high carotenoid content. Precision farming techniques are needed to produce healthy food. In particular, nutrient supply and irrigation, sowing, crop management and harvesting need to be carried out with precision techniques. These factors are all prerequisites for effective and healthy growing and processing. The aim was to use the yields of the four sweet maize hybrids grown on the largest area to examine their nutritional values and concentrations (mg kg-1 dry matter) and to analyse their yield per hectare. Concentration is important for the consumer because K, P, Mg, Ca, Fe, Zn, and Na play an important role in metabolism, skin protection, and bone and tooth health. The new results obtained show that the amount of lutein and zeaxanthin per hectare is important for the processing industry, especially for use in food supplements. Their anti-inflammatory effects and their role in disease prevention (cardiovascular diseases, Age-Related Macular Degeneration (AMD)) have been demonstrated. Consumers choose sweet maize mainly on the basis of its palatability, which is why the sugar content of the hybrids was also studied. We assumed that the element concentration in the yield of new hybrids with higher yield per hectare does not decrease with increasing yield. The concentrations of zeaxanthin, β-cryptoxanthin and β-carotene appear in one principal component and they are in close positive correlation with each other. The lutein concentration was independent of the former three compounds. The independence of the lutein concentration means that it is not possible to estimate its amount based on the other three components. For yield per unit area, the correlation is one-dimensional. Yield determines the lutein, zeaxanthin, β-cryptoxanthin and β-carotene concentrations per hectare.",
          "table": []
        }
      ]
    },
    {
      "title": "Introduction",
      "div": [
        {
          "text": "Sweet maize is either eaten fresh or used as a raw material for food processing and the industry. Farmers are interested in achieving the highest income per unit area. This in turn depends on who they sell their product to.",
          "table": []
        },
        {
          "text": "Consumers are interested in buying the most nutrient-rich food possible. As the amount of food consumed per day is limited, the physiological effect of more \"concentrated\" products is better (functional foods). If consumers are willing to pay more for such products, it is preferable to grow hybrids with a high nutrient concentration.",
          "table": []
        },
        {
          "text": "For hybrids grown for the processing industry, however, quantity is the most important factor, as higher yields provide higher revenues. If the aim is to extract a component, e.g., lutein, the quantity per unit area is important. This is described as the element concentration multiplied by yield.",
          "table": []
        },
        {
          "text": "The correlation of the content values was examined, and a true multivariate statistical method (Principal Component Analysis (PCA)) was used to group the hybrids under study. This was also done for the concentration data and for the element yields multiplied by yield.",
          "table": []
        },
        {
          "text": "Fresh sweet maize is an increasingly popular food mainly because of its high valuable content values and taste. New results in the precision production of sweet maize are expected, mainly due to its increasing role in healthy nutrition, however, breeding and production technology challenges are also expected in the context of sustainability and climate change.",
          "table": []
        }
      ]
    }
  ]
}

```

these two broad approaches, the application of herbicides is the prevalently used one. However, the use of herbicides involves several drawbacks. Applying herbicides to the entire field is very expensive. Herbicides cost roughly \$60 per acre which is 10% of the expected market revenue of the corn as per the 2021 Purdue Crop Cost and Return Guide [10]. Excessive use of herbicides is also detrimental to soil fertility, the aquatic ecosystem, and human health. Furthermore, weeds develop resistance to herbicides over time. Selective spraying of herbicides would address these shortcomings and also cut down the cost of the herbicides. Selective spraying of herbicides or the removal of weeds requires precise identification of weeds. Hence, the identification of weeds plays an important role in the management and control of weeds. Given the scale of the problem, manual identification of weeds is either untenable or impractical in many situations. ML techniques are successfully applied for the precise identification of weeds. The use of machine learning techniques also made the automation of weed control and management possible.

This review surveyed the various ML approaches that were applied over the years for the identification of weeds in cornfields. We also describe in full technical detail, the type of ML problem solved (classification, object detection etc.), the type of weeds identified, the type of data used, the type of error metrics used to evaluate the performances of these approaches. These ML approaches are grouped into three major categories namely, SVM, Neural Networks, and Miscellaneous. Section 3 describes the first category, SVM, Section 4 discusses Neural Network approaches, and Section 5 elaborates on the miscellaneous ML techniques used in the past for the identification of weeds in cornfields. Section 6 explains the importance of data for the performance of the ML techniques and the various metrics that are used to evaluate the performance of these techniques. Section 7 briefly discusses the conclusion and future research directions of ML-based identification of weeds. Table 1 summarises the abbreviations that are used in this study.

2. Machine learning

Machine Learning is a class of Artificial Intelligence (AI) that focuses on aiding the computers in learn the underlying relationship between inputs and outputs from the given data and make accurate predictions [11]. ML algorithms employ statistical methods to learn from the exposed data without any explicit programming instructions [12]. The workflow of a typical ML model is as depicted in Fig. 1 and consists of the following phases:

Data acquisition – gathering data (open-source datasets, sensors, etc.)
Data pre-processing – involves cleaning the data, making the data

Table 1
List of abbreviations.

| Abbreviation | Explanation |
|--------------|--|
| AI | artificial intelligence |
| ANN | artificial neural network |
| ASM | active shape models |
| BP | backpropagational network |
| CCM | color co-occurrence method |
| CDC | canonical discriminant classification |
| CNN | convolutional neural network |
| DA | discriminant analysis |
| DHT | double hough transform |
| DT | decision tree |
| DWT | discrete wavelet transform |
| EOH | edge of histogram |
| FFT | fast fourier transform |
| FIP | fast image processing |
| FLDA | fisher linear discriminant analysis |
| GA | genetic algorithms |
| GAN | generative adversarial network |
| GLCM | gray level co-occurrence matrix |
| GMM | gaussian mixture model |
| HIS | hue, intensity, saturation |
| HT | hough transform |
| IOU | intersection over union |
| KNN | k-nearest neighbor |
| LBP | linear binary pattern |
| LDA | linear discriminant analysis |
| LIDAR | light detection and ranging |
| LMC | linear margin classifier |
| LR | linear regression |
| LS-SVM | least square-support vector machine |
| MOG | mixer of gaussian |
| ML | machine learning |
| NDVI | normalised difference vegetation index |
| PCA | principal component analysis |
| PCANet | principal component analysis network |
| PDF | probability density functions |
| PNN | probabilistic neural network |
| RBF | radial basis function |
| RCRD | robust crop row detection |
| RF | random forest |
| RGB | red, green, blue |
| ROI | region of interest |
| RVI | ratio vegetation index |
| SMH | shape matrix histogram |
| SOM | self-organizing map |
| SPCA | sparse principal component analysis |
| SVTD | support vector data description |
| SVM | support vector machine |
| SWLDA | stepwise linear discriminant analysis |
| VI | vegetation indices |
| WIR | weed infestation rate |

```
{
  "text": "This review surveyed the various ML approaches that were applied over the years for the identification of weeds in cornfields. We also describe in full technical detail, the type of ML problem solved (classification, object detection etc.), the type of weeds identified, the type of data used, the type of error metrics used to evaluate the performances of these approaches. These ML approaches are grouped into three major categories namely, SVM, Neural Networks, and Miscellaneous. Section 3 describes the first category, SVM, Section 4 discusses Neural Network approaches, and Section 5 elaborates on the miscellaneous ML techniques used in the past for the identification of weeds in cornfields. Section 6 explains the importance of data for the performance of the ML techniques and the various metrics that are used to evaluate the performance of these techniques. Section 7 briefly discusses the conclusion and future research directions of ML-based identification of weeds. List of abbreviations. artificial neural network summarizes the abbreviations that are used in this study.",
  "table": {
    "List of abbreviations. artificial neural network"
```

```
"tables": [
  {"title": "List of abbreviations. artificial neural network", "head": [{"active shape models"}, {"value": [{"backpropagational network"}, {"color co-occurrence method"}, {"canonical discriminant classification"}, {"convolutional neural network"}, {"discriminant analysis"}, {"double hough transform"}, {"decision tree"}, {"discrete wavelet transform"}, {"edge of histogram"}, {"fast fourier transform"}, {"fast image processing"}, {"fisher linear discriminant analysis"}, {"genetic algorithms"}, {"generative adversarial network"}, {"gray level co-occurrence matrix"}, {"hue, intensity, saturation"}, {"hough transform"}, {"intersection over union"}, {"k-nearest neighbor"}, {"linear binary pattern"}, {"linear discriminant analysis"}, {"light detection and ranging"}, {"linear margin classifier"}, {"linear regression"}, {"least square-support vector machine"}, {"mixer of gaussian"}, {"machine learning"}, {"normalised difference vegetation index"}, {"principal component analysis"}, {"principal component analysis network"}, {"probability density functions"}, {"probabilistic neural network"}, {"radial basis function"}, {"robust crop row detection"}, {"random forest"}, {"red, green, blue"}, {"region of interest"}, {"ratio vegetation index"}, {"shape matrix histogram"}, {"self-organizing map"}, {"sparse principal component analysis"}, {"support vector data description"}, {"support vector machine"}, {"stepwise linear discriminant analysis"}, {"vegetation indices"}, {"weed infestation rate"}]}],
}]
```

Table 2
Summary of studies that employed SVMs for the identification of weeds.

| Study | Research problem | Dataset | Accuracy |
|-------|---|---|--|
| [17] | Detection of weed and nitrogen stress in corn | 20 data points of 9 treatments consisting of 4 replicates thereby resulting in a data set of 720 entries. 50% of the data was used for training purposes while the remaining 50% was used for testing. Hardware used: A Compact Airborne Spectrographic Imager | 10-fold cross-validation used (testing data set). SVM: 66% to 76% for combined weed and nitrogen application rates. 73% to 83% accuracy. 83% to 93% accuracy, respectively for weed and nitrogen treatments separately. |
| [57] | Classification of weed and corn seedlings using textural features | 66 color images (30 corn seedlings, 36 weed images). 60% used for training, 40% for testing Hardware used: A digital camera (resolution of 640×480 pixels). | SVM with different feature selections produced 92.31 to 100%. |
| [42] | Using shape parameters to identify corn/weed seedling in fields | 64 color images (40-training set, 24-testing set) Hardware used: A digital camera (resolution of 640×480 pixels). | SVM (Sigmoid-96.5%, RBF-67.67% and Polynomial-90%, respectively) |
| [43] | Studying local binary pattern for automated weed classification | 200 images (100 each of broadleaf and grass, respectively). Dataset is divided into 10 subsets. 1 subset used as the testing set and 9 subsets for training. Hardware used: A digital camera (resolution 1200×768 pixels) | SVM: 98.5% |
| [44] | Categorize weed seedlings into groups for spot spraying and weed scouting | 400 features rows for training and verification. 240 external data sets were used for testing (100 data of <i>amaranthus palmeri</i> weeds and 100 of other weeds). Weed species: <i>Phyllanthus Urinaria</i> , <i>Agerantum Conyzoides</i> sp., <i>Amaranthus palmeri</i> sp., and other weeds (dicotyledon and monocotyledon) Hardware used: Logitech c615 Webcam (resolution of 1920×1050 pixels) | SVM: True positive true value for all the groups (100%), for second variant group for <i>Agerantum Conyzoides</i> (66.7%). |
| [56] | Classifying weed images using Wavelet Transform | 1200 images (500 of broad category, 500 of narrow category, and 200 of unknown category, respectively). Training: 600 images (250 of broad leaves, 250 of narrow and 100 unknown weeds). Testing: Remaining 600 images (250 images of broad leaves, 250 of narrow and 100 unknown weeds). | Synlet wavelet family: 98.1% |

Table 2 (continued)

| Study | Research problem | Dataset | Accuracy |
|-------|---|---|---------------------------------------|
| [19] | Classification of maize and weed | Hardware used: Not mentioned 1000 images (500 of crop, 500 of weed). 450 of each were used for Training, 100 for Testing. | 82% |
| [46] | Performance comparison of algorithms used for identifying weeds | Hardware used: Not mentioned 2560 images. 1155 of each class (weed and crop) used for training and 125 images per class used to validate the trained model. Hardware used: A 10 MP digital camera | SVM: 100% for crop and 83.2% for weed |

matrices are first subjected to convolution, followed by pooling.

Convolution is mainly performed for feature extraction, and it is done using filters or kernels in the form of matrices. Kernel matrices are of a considerably smaller dimension and are chosen appropriately based on the nature of the problem. Convolved features are obtained by taking Hadamard product between the image and the kernel matrices. As the kernel matrices of smaller dimensions compared to image matrices, convolved features are generated by sliding the kernel matrices from left to right and top to bottom and taking the Hadamard product at each position of the kernel matrix on the image matrix. The sliding of the kernel matrix is defined in terms of 'strides'; for example, a stride of 1 allows the kernel filter to shift one column left and one row down. In addition to striding, convolution also involves padding, which adds additional rows and columns of zeros to the input matrices so that the pixel information present in the edges of image matrices is not lost.

The features extracted from convolution are sensitive to the location, and to achieve a translation invariance (less sensitive to the location) of these features, a downsampling operation called 'pooling' is carried out. Like the kernel filter, the pooling filter is also of smaller size compared to the feature maps. The size of the feature maps is usually halved when using pooling filters. For example, the size of 4×4 will be converted to 2×2 . Max pooling and average pooling are the two most common types of pooling filters used in CNN. Average pooling involves the extraction of the average value of map features, whereas max-pooling extracts the maximum value. The choice of pooling depends upon the nature of the given data. Average pooling tends to smoothen the image, whereas max-pooling tends to brighten or select bright pixels from the image. After pooling, a fully connected layer is formed as a single column vector for each example and fed into the neural networks, and, trained using a backpropagation algorithm. For classification problems, CNNs commonly use ReLU, eLU, and tanH for hidden layers and SoftMax activation functions for the output layer.

Mozhou et al. [54] proposed a new neural network architecture, SOM where the neurons are associated with local linear mappings for the classification of crop and weed from their near-infrared reflectance spectra which were obtained with the help of an imaging spectrograph. The dataset consisted of 80 corn samples, 77 samples of the buttercup (*Ranunculus repens*), 79 samples of Canada thistle (*Girium arvense*), 75 samples of charlock (*Sinapis arvensis*), 73 samples of chickweed (*Stellaria media*), 76 samples of dandelion (*Taraxacum officinale*), 80 samples of grass (*Poa annua*), 78 samples of redshank (*Polygonum persicaria*), 75 samples of stinging nettle (*Urtica dioica*), 78 samples of wood sorrel (*Onalis europaea*) and 75 samples of yellow trefoil (*Medicago lupulina*) resulting in a dataset of 766 and 88 reflectance spectra for weed and corn, respectively. A separability index was used to obtain five principal components and the following wavelengths: 539, 540, 542, 545, 549, 557, 565, 578, 585, 596, 605, 639, 675, 687, 703, 814 and

"{\title": "Summary of studies that employed SVMs for the identification of weeds.\", \"head\": {\Study\", \"Research problem\", \"Dataset\", \"Accuracy\", \"value\": [{\"[17]\", \"Detection of weed\", \"20 data points of 9\", \"10-fold cross-\", [{\"\", \"and nitrogen stress in\", \"treatments consisting of\", \"validation used\"], [{\"\", \"corn\", \"4 replicates thereby\", \"(testing data set).\", [{\"\", \"\", \"resulting in a data set of\", \"SVM: 66% to 76% for\"], [{\"\", \"\", \"720 entries. 50% of the\", \"combined weed and\"], [{\"\", \"\", \"data was used for\", \"nitrogen application\"], [{\"\", \"\", \"training purposes while\", \"rates.\"], [{\"\", \"\", \"the remaining 50% was\", \"73% to 83% accuracy\"], [{\"\", \"\", \"used for testing.\"], \"83% to 93%\"], [{\"\", \"\", \"Hardware used: A\", \"accuracy\", [{\"\", \"\", \"Compact Airborne\", \"respectively for weed\"], [{\"\", \"\", \"Spectrographic Imager\", \"and nitrogen\"], [{\"\", \"\", \"\", \"treatments\"], [{\"\", \"\", \"\", \"separately.\"], [{\"[57]\", \"Classification of\", \"66 color images (30 corn\", \"SVM with different\"], [{\"\", \"weed and corn\", \"seedlings, 36 weed\", \"feature selections\"], [{\"\", \"seedlings using\", \"images). 60% used for\", \"produced 92.31 to\"], [{\"\", \"textural features\", \"training, 40% for testing\", \"100%\"], [{\"\", \"\", \"Hardware used: A\", [{\"\", \"\", \"digital camera\", \"\", [{\"\", \"\", \"(resolution of 640x480\", \"\", [{\"\", \"\", \"pixels\"], [{\"\", \"\", \"\", \"[42]\", \"Using shape\", \"64 color images (40-\", \"SVM (Sigmoid-\", \"\", \"parameters to\", \"training set, 24-testing\", \"96.5%, RBF-67.6%\"], [{\"\", \"identify corn/weed\", \"set\"], \"and Polynomial-90%\"], [{\"\", \"seedling in fields\", \"Hardware used: A\", \"respectively\"], [{\"\", \"\", \"digital camera\", \"\", [{\"\", \"\", \"(resolution of 640x480\", \"\", [{\"\", \"\", \"pixels\"], \"\", [{\"\", \"\", \"\", \"[43]\", \"Studying local\", \"200 images (100 each of\", \"SVM: 98.5%\"], [{\"\", \"binary pattern for\", \"broadleaf and grass\", \"\", [{\"\", \"\", \"automated weed\", \"respectively). Dataset is\", \"\", [{\"\", \"\", \"Classification\", \"divided into 10 subsets.\"], [{\"\", \"\", \"\", \"1 subset used as the\", \"\", [{\"\", \"\", \"testing set and 9 subsets\", \"\", [{\"\", \"\", \"for training.\"], [{\"\", \"\", \"\", \"Hardware used: A\", \"\", [{\"\", \"\", \"\", \"digital camera\", \"\", [{\"\", \"\", \"\", \"(resolution 1200x768\", \"\", [{\"\", \"\", \"\", \"pixels\"], [{\"\", \"\", \"\", \"[44]\", \"Categorize weed\", \"400 features rows for\", \"SVM: True positive\"], [{\"\", \"seedlings into groups\", \"training and\", \"true value for all the\", [{\"\", \"for spot spraying and\", \"verification. 240\", \"groups (100%\", for\", [{\"\", \"weed scouting\", \"external data sets were\", \"second variant group\"], [{\"\", \"\", \"used for testing (100\", \"for Agerantum\"], [{\"\", \"\", \"data of amarathus\", \"Conyzoides (66.7%).\", [{\"\", \"\", \"palmer weeds and 100 of\", \"\", [{\"\", \"\", \"other weeds\"], \"\", [{\"\", \"\", \"Weed species:\", \"\", [{\"\", \"\", \"Phyllanthus Urinuria\", \"\", [{\"\", \"\", \"Agerantum Conyzoides\", \"\", [{\"\", \"\", \"\", \"sp., Amaranthus palmeri\", \"\", [{\"\", \"\", \"\", \"sp., and other weeds\", \"\", [{\"\", \"\", \"\", \"dicotyledon and\", \"\", [{\"\", \"\", \"\", \"monocotyledon\", \"\", [{\"\", \"\", \"\", \"Hardware used:\", \"\", [{\"\", \"\", \"Logitech c615 Webcam\", \"\", [{\"\", \"\", \"\", \"resolution of\", \"\", [{\"\", \"\", \"\", \"1920x1050 pixels\", \"\", [{\"\", \"\", \"Classifying weed\", \"1200 images (500 of\", \"Symlet wavelet\", \"\", \"images using\", \"broad category, 500 of\", \"family: 98.1%\"], [{\"\", \"Wavelet Transform\", \"narrow category, and\", \"\", [{\"\", \"\", \"200 of unknown\", \"\", [{\"\", \"\", \"category, respectively\"], \"\", [{\"\", \"\", \"Training: 600 images\", \"\", [{\"\", \"\", \"\", \"(250 of broad leaves\", \"\", [{\"\", \"\", \"250 of narrow and 100\", \"\", [{\"\", \"\", \"unknown weeds\"], \"\", [{\"\", \"\", \"Testing: Remaining 600\", \"\", [{\"\", \"\", \"images (250 images of\", \"\", [{\"\", \"\", \"broad leaves, 250 of\", \"\", [{\"\", \"\", \"narrow and 100\", \"\", [{\"\", \"\", \"unknown weeds\"], \"\", [{\"\", \"\", \"\", \"Hardware used: Not\", \"\", [{\"\", \"\", \"mentioned\", \"\", [{\"\", \"\", \"\", \"[19]\", \"Classification of\", \"1000 images (500 of\", \"82%\"], [{\"\", \"maize and weed\", \"crop, 500 of weed). 450\", \"\", [{\"\", \"\", \"of each were used for\", \"\", [{\"\", \"\", \"\", \"Training, 100 for\", \"\", [{\"\", \"\", \"Testing\"], \"\", [{\"\", \"\", \"\", \"Hardware used: Not\", \"\", [{\"\", \"\", \"mentioned\", \"\", [{\"\", \"\", \"\", \"[46]\", \"Performance\", \"2560 images. 1155 of\", \"SVM: 100% for crop\", \"\", \"comparison of\", \"each class (weed and\", \"and 83.2% for weed\"], [{\"\", \"\", \"algorithms used for\", \"crop used for training\", \"\", [{\"\", \"\", \"identifying weeds\", \"and 125 images per class\", \"\", [{\"\", \"\", \"used to validate the\", \"\", [{\"\", \"\", \"trained model\", \"\", [{\"\", \"\", \"\", \"Hardware used: A 10\", \"\", [{\"\", \"\", \"\", \"MP digital camera\", \"\", \"\"]}]\".