Miskolci Egyetem Alkalmazott Matematikai

Gépészmérnöki és Informatikai Kar Intézeti Tanszék

Általános Informatikai Intézeti Tanszék

Természetes nyelvű szövegek témakör szerinti elemzése gépi tanulás segítségével

Szakdolgozat

**Készítette:**

**Név:** Farkas Ádám

**Neptunkód:** FE019W

**Szak:** Mérnökinformatikus BSc

Korszerű web technológiák szakirány

|  |  |  |  |
| --- | --- | --- | --- |
| **Miskolci Egyetem**  **Gépészmérnöki és Informatikai Kar** |  | | **Általános Informatikai Intézeti Tanszék**  3515 Miskolc-Egyetemváros |
| Szak**: Mérnökinformatikus** | |  | Szakdolgozat azonosító: **IAL/FE019W/BSc/2023** |
| Szakirány: korszerű web-technológiák | |  | **Intézmény azonosító: FI 87515** |
|  | |  |  |

**SZAKDOLGOZAT FELADAT**

**Farkas Ádám**

BSc mérnökinformatikus jelölt részére

|  |  |
| --- | --- |
| A tervezés tárgyköre: | **ide a tárgykör** |
| A feladat címe: | **Természetes nyelvű szövegek témakör szerinti elemzése gépi tanulás segítségével** |
| **A feladat részletezése:**  Az aktuálisan elérhető idegen nyelvű szövegek lehetővé teszik, hogy azokon gépi tanulási módszerek felhasználásával a tartalomra vonatkozó elemzéseket hajtsunk végre. A dolgozat célja egy olyan idegennyelv tanulását segítő eszköz fejlesztése, amely egy szövegkorpusz automatizált feldolgozásával statisztikai kimutatásokat, a szavak témakör szerint csoportosítását, illetve egy tartalmi összefoglalót is megfogalmaz belőle. A dolgozat ezen, alapvetően klaszterezési probléma megoldásával foglalkozik. Bemutatja a természetes nyelvi elemzés matematikai hátterét, hasonló alkalmazási területeken a készen elérhető megoldásokat. Az elemzések Python programozási nyelv, és a hozzá elérhető adatelemző függvénykönyvtárak segítségével készülnek. | |

|  |  |
| --- | --- |
| Tervezésvezető:  Piller Imre | Tanszék, beosztás:  Alkalmazott Matematikai Intézeti Tanszék |
| Konzulens(ek):  idejönakonzulensneve | Cég, beosztás:  idejönakonzcégeésbeosztás |
| A szakdolgozat kiadásának időpontja: | 2023.02.15. |
| A szakdolgozat beadásának határideje: | 2023.11.17. |
| Miskolc, 2023.11.30. | **Prof. Dr. Kovács László**  tanszékvezető egyetemi tanár |

1. A szakmai gyakorlat helye: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. A szakmai gyakorlat vezetőjének neve: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. A szakdolgozat módosítása: szükséges (a módosítást külön lap tartalmazza)   
 nem szükséges (a megfelelő rész aláhúzandó)

Miskolc, \_\_\_\_\_\_\_\_\_\_\_\_\_

tervezésvezető aláírása

4. A tervezést ellenőriztem: (1) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(2) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(3) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(4) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dátum, tervezésvezető aláírása

5. A szakdolgozat beadható

nem adható be

Miskolc, \_\_\_\_\_\_\_\_\_\_\_\_\_

konzulens aláírása tervezésvezető aláírása

6. A szakdolgozat ….. szövegoldalt,

….. db rajzot,

….. db CD mellékletet

….. egyéb mellékletet tartalmaz.

7. A szakdolgozat bírálatra: bocsátható

nem bocsátható

A bíráló neve, címe: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Miskolc, \_\_\_\_\_\_\_\_\_\_\_\_\_

tanszékvezető aláírása

8. Osztályzat: a bíráló javaslata: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

a tanszék javaslata: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

a Záróvizsga Bizottság döntése: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Miskolc, \_\_\_\_\_\_\_\_\_\_\_\_\_

a Záróvizsga Bizottság elnökének aláírása

# Köszönetnyilvánítás

Szeretnék köszönetet mondani Piller Imrének, a tervezésvezetőmnek, aki rengeteg hasznos tanáccsal, és figyelmes, segítőkész hozzáállással egyengette a dolgozatom útját. Szeretném megköszönni a barátnőmnek a végtelen türelmet és támogatást, amelyet kaptam tőle a dolgozat megírása közben. Szeretnék még köszönetet mondani a családomnak, akik nélkül sosem jutok el oda, ahol ma tartok.

# 1. Bevezető

Szakdolgozatom témájaként egy gépi tanulással elkészített elemzést választottam. Az aktuálisan elérhető idegen nyelvű szövegek lehetővé teszik, hogy azokon gépi tanulási módszerek felhasználásával a tartalomra vonatkozó elemzéseket hajtsunk végre. A dolgozatom célja egy olyan idegennyelv tanulását segítő eszköz fejlesztése, amely egy szövegkorpusz automatizált feldolgozásával statisztikai kimutatásokat készít, emellett a benne szereplő szavakat témakör szerint csoportosítja, illetve egy tartalmi összefoglalót is megfogalmaz belőle. A munkám ennek a klaszterezési problémának a megoldásával foglalkozik. Bemutatja a természetes nyelvi elemzés matematikai hátterét, és hasonló alkalmazási területeken a készen elérhető megoldásokat. Az elemzéseket Python programozási nyelven, és az ehhez elérhető adatelemző függvénykönyvtárak segítségével készítettem el. Az alábbiakban egy rövid leírást adok arról, hogy pontosabban mivel is fogok foglalkozni a dolgozatban.

A gépi tanulás és a mesterséges intelligencia területe az utóbbi néhány évben hatalmas népszerűségre tett szert. Ha gazdasági szempontból nézzük, ez a folyamat ugyan a 2010-es évek második felében még nem hatott látványosan a termelékenység növekedésére [1], viszont nem hagyhatjuk figyelmen kívül például a tudományos, kulturális életre gyakorolt hatását sem. Ez az egyre fokozódó, robbanásszerű fejlődés termesztésen kihatott a digitális formátumban már meglévő szövegekkel kapcsolatos leíró, elemzési és összehasonlítási feladatokra is. Ezek a vizsgálatok közelebb vihetnek minket az egyes nyelvek, illetve nyelvek közötti kapcsolatok jobb megértéséhez.

A digitális korszak előrehaladtával az idegen nyelvű szövegek óriási mennyisége vált elérhetővé. Ez a számítógépek által rendelkezésünkre bocsátott egyre hatalmasabb számítási kapacitás felhasználásával már eddig is lehetőséget teremtett a különböző témájú szövegkorpuszok mélyebb megismerésére, rendszerezésére. Ugyanakkor ezeknek a szövegeknek a nagysága sokszor már szinte áttekinthetetlen, és manuális, vagy egyszerű digitalizált formában történő elemzésük nemcsak időigényes, de általában teljeskörűen lehetetlen is. Ezt a folyamatot fejleszti tovább a gépi tanulás alkalmazása, amely lehetővé teszi ilyen nagy mennyiségű szövegbázis rendszerezett és minden eddiginél gyorsabb feldolgozását, illetve statisztikai, nyelvtani, és valamilyen szinten már a jelentésbeli elemzését is.

A szakdolgozatom középpontjában egy olyan eszköz kifejlesztése állt, amely amellett, hogy statisztikai elemzésekkel segíti az idegen nyelvek tanulását, a szövegeket automatizált módon témakörök mentén is csoportosítja. A vizsgálatom fontos része az a klaszterezési probléma, melynek megoldása révén egy olyan rendszer hozható létre, amely a rendelkezésre álló szövegkorpuszokat, jelen esetben a közismert Harry Potter könyvsorozat első regényének szövegét tematikus struktúrába szervezi. Mindezek mellett egy, a könyv tartalmába, fontosabb mozzanataiba betekintést nyújtó összefoglalót is elkészítészítettem.

A könyv szövegének statisztikai alapú vizsgálatát négy különböző nyelven alkottam meg. A tartalmi és a klaszterezési elemzéseket az angol nyelvű változaton készítettem el.

Ahogy már említettem, elemzéseim során a Python programozási nyelvet alkalmaztam, illetve felhasználtam az ehhez elérhető adatelemző függvénykönyvtárakat. Ezzel az erős technológiai bázissal megkíséreltem egy olyan hatékony eszköz létrehozását, amely segítséget nyújt azoknak, akik szeretnének mélyebben elmerülni a nyelvtanulás hagyományostól eltérő lehetőségeiben, illetve a különféle nyelvek mélyrehatóbb tanulmányozásában.

A jövőre nézve szerintem fontos alkalmazási terület lehet még az összehasonlító nyelvészet, amelyet szintén nagymértékben segíthet a természetes nyelvű szövegek gépi tanulás és mesterséges intelligencia segítségével történő elemzése. Erről a felhasználási területről is igyekszem egy kisebb ízelítőt adni a következő fejezetben.

# 2. Az NLP problémakör áttekintése

2.1. Az NLP

Az elemzéseim megalapozásához fontos volt, hogy a természetes nyelvek feldolgozása (Natural Language Processing, a továbbiakban NLP) témakörét próbáljam alaposabban megérteni. Ez a terület jelenleg már megkerülhetetlen a technológiai fejlődés és az adatmennyiség hihetetlen mértékű növekedésének köszönhetően. Az NLP a mesterséges intelligencia és a nyelvészet metszetében áll. Célja nem kevesebb, mint az emberi és a gépi kommunikáció közötti szakadék áthidalása. Ebben a fejezetben szeretnék áttekintést adni az NLP alapelveiről, fejlődéséről és lehetséges jövőbeli alakulásáról.

A természetes nyelv megértéséhez az emberi kommunikáció szerteágazó bonyolultságával kell megküzdenünk. Mindezt igen nehéz feladat számítógépes formába átültetni. A korai NLP-rendszerek nagymértékben támaszkodtak különféle nyelvi szabályokra és kézzel készített nyelvtanokra. Chomsky generatív nyelvtana, amelyet a Syntactic Structures című korai művében adott közre [2], elméleti alapot adott a mondatok szintaktikai struktúráinak elemzésére és megértésére. Ez a megközelítés csak korlátozottan alkalmazható az emberi nyelv hatalmas változatosságának kezelésére.

A statisztikai módszerek és a gépi tanulás felé való elmozdulás jelentős löketet adott az NLP fejlődésének. A fontos eredmények között találjuk a rejtett Markov-modellek (HMM) bevezetését a beszédrészek címkézésére [3], valamint a statisztikai gépi fordítás témakörében elért előretörést. Ezeből a példákból láthatóvá vált, hogy az adatvezérelt módszerek felhasználhatók összetett nyelvi feladatok megoldására is. Mindez hozzájárult a gépi tanulás elterjedéséhez az NLP-ben.

2.2. Alapvető feladatok az NLP-ben

2.2.1. Szövegosztályozás és érzelemelemelemzés

Az NLP egyik alapvető feladata a szövegosztályozás, ahol a cél egy dokumentum előre meghatározott osztályokba sorolása. Pang és Lee "Opinion Mining and Sentiment Analysis" című 2008-as munkájukban a szövegosztályozás egyik részterületét, a hangulatelemzést vizsgálták [4]. Ez a feladat a szövegben kifejezett érzelmek meghatározását jelenti. Minél pontosabban tudjuk meghatározni a szövegek, szavak mögött megbúvó érzelmi töltetet, annál közelebb kerülünk például a felhasználói vélemények jobb megértéséhez az értékelésekben, a közösségi médiában és a vásárlói visszajelzésekben, hogy csak néhány példát említsek.

2.2.2 Nevezett entitások felismerése (NED)

A megnevezett entitások, például emberek, szervezetek vagy helyszínek neveinek azonosítása alapvető a lényeges információk megszerzéséhez. Erre remek példa a A CoNLL-2003, amely egy megosztott feladat részeként kiadott, nyelvfüggetlen nevesített entitásfelismerési adathalmaz [5]. Az adatok nyolc fájlból állnak össze, amelyek az angol és a német nyelvet fedik le. Mindkét nyelvhez tartozik egy gyakorlófájl, egy fejlesztőfájl, egy tesztfájl és egy nagy fájl megjegyzések nálküli adatokkal.

Az ezirányú kitatások ösztönözték a NED fejlődését, különböző megközelítésekkel. Többek között a szabályalapú rendszerek, illetve olyan gépi tanulási modellek tartoznak ide, mint például a feltételes véletlen mezők (CRF, vagyis Conditional Random Field) és újabban már a deep learning módszerek is ide sorolhatók.

2.2.3 Gépi fordítás

A szövegek egyik nyelvről a másikra történő fordításának lehetősége már régóta az NLP egyik fontos célja. A neurális gépi fordítás (NMT vagyis Neural Machinr Translation) megjelenése, amelyet Sutskever et al. (2014) és Bahdanau et al. (2014) mutatott be, paradigmaváltást jelentett a kérdéskörben [6]. Az NMT egyik legfontosabb újítása, a figyelemmechanizmusok bevezetése lehetővé tette, hogy a modellek a bemeneti adathalmaz meghatározott részeire összpontosítsanak, ami jelentősen javította a fordítás minőségét.

2.3. Deep learning és NLP

2.3.1. Szóbeágyazások

A szóbeágyazások megjelenése forradalmasította a szavak NLP-rendszerek általi reprezentációját [7]. A beágyazások a szavak közötti szemantikai kapcsolatokat rögzítik, ez pedig lehetővé teszi a modellek számára a kontextus és a jelentés megértését. Például a Word2Vec, a GloVe és a fastText technikákat lehet érdemes kiemelni, amelyek a különböző NLP-alkalmazások szerves részévé váltak, és alkalmazásukat a gyakorlatban is látni fogjuk az általam hozott szövegek elemzése során.

2.3.2. Rekurrens neurális hálózatok (RNN) és hosszú rövidtávú memória (LSTM)

A nyelvek szekvenciális jellege vezetett ahhoz a felfedezéshez, hogy a rekurrens neurális hálózatok (RNN vagyis Recurrent Neural Network) jól felhasználhatók az NLP-feladatok megoldásához.

Hochreiter és Schmidhuber 1997-ben publikált LSTM-architektúrája [8] megoldotta az eltűnő gradiens problémáját, amely lehetővé téve az RNN-ek számára, hogy megragadják a szekvenciák hosszú távú függőségeit. A réshosszra való viszonylagos érzéketlensége az előnye más RNN-ekkel, rejtett Markov-modellekkel, illetve más szekvenciatanulási módszerekkel szemben. Célja, hogy az RNN számára olyan rövid távú memóriát biztosítson, amely több ezer időlépésre képes, tehát nevezhető „hosszú rövidtávú memóriának”. Többek között az idősorokon alapuló adatok osztályozására alkalmazható, illetve feldolgozására és előrejelzésére, például a kézírásbeszédfelismerés, gépi fordítás, beszédaktivitás-felismerés, robotvezérlés, videojátékok és az egészségügy területén. Természetesen meg kell említenem, hogy az LSTM használata sikeresen beépíthető olyan feladatokba is, mint a nyelvi modellezés és a gépi fordítás.

2.3.3. Transzformátorok és figyelemfelkeltő mechanizmusok

A Transformer architektúra Vaswani és társai [9] általi bevezetése (2017) fontos mérföldkő volt az NLP fejlődésében. A Transformerek az önfigyelési mechanizmusokra támaszkodva felülmúlták a korábbi modelleket a kontextusfüggő információk megragadásában. A Transformer architektúra lett az alapja az olyan korszerű és közismertté váló modelleknek, mint a BERT [10] a kontextuális szóbeágyazások és a GPT (Generative Pre-trained Transformers) [11] a nyelvi modellezés területén.

2.4. Kihívások és jövőbeli irányok

2.4.1 Etikai megfontolások az NLP-ben

Ahogy az NLP-technológiák egyre jobban elterjednek, úgy természetesen felmerülnek az elfogultsággal, a méltányossággal és a magánélet, illetve érzékeny vagy személyes adatok védelmével kapcsolatos etikai aggályok. Például érdemes megemlíteni a szóbeágyazásokban megjelenő részrehajlást, tehát megfelelő odafigyeléssel kell kezelnünk a nyelvi adatokban akaratlanul is megjelenő társadalmi elfogultságokat, előítéleteket, kategorizálásokat [12]. Ha belegondounk, egyértelmű, hogy az NLP-rendszerek felelős fejlesztésére és figyelmes gondozására van szükség, ha a lehetőségekhez mérten igazságos és elfogulatlan eredményeket szeretnénk biztosítani.

2.4.2 Multimodális NLP

A többféle modalitás, például a szöveg, a képek és a beszédhang integrálása a kutatások egyre növekvő területéhez tartozik. A legújabb munkák, mint például a CLIP, a modalitások közötti összevont tanulásban rejlő lehetőségeket tárgyalják, a fenti példában a szövegek és a hozzájuk tartozó képek összeillesztése területén [13]. A multimodális NLP hozzájárul a különféle típusú tartalmak átfogóbb megértéséhez és generálásához, és alkalmazásai lefedik többek között a képfeliratozást segítő technológiákat.

Mint az végigkövethettük, a természetes nyelvfeldolgozás figyelemre méltó átalakuláson ment keresztül az utóbbi évtizedekben, és átölelte a szabályalapú rendszereket, az adatvezérelt megközelítéseket, a deep learning megjelenésével bezárólag. A nyelvészeti elméletekre, statisztikai módszerekre és kortárs neurális architektúrákra támaszkodva az NLP egyre sikerekesebben alkalmazható az emberi nyelvek megértésében és generálásában. Azonban továbbra is rengeteg kihívás áll fenn, és folyamatos interdiszciplináris együttműködésre van szükség az etikai aggályok kezelése, az alkalmazkodóképesség javítása és a multimodális megértés fejlesztése érdekében. Ahogy az NLP tovább fejlődik, például az egészségügyi ellátástól az oktatásig számos területen kifejtett hatása azt bizonyítja, hogy a mesterséges intelligencia egyre bővülő területének egyik fontos bázisává vált.

2.5. NLP a gyakorlatban

2.5.1. Az elemzések vázlata

A természetes nyelvi feldolgozás és az adatelemzés metszéspontjában elhelyezkedő szövegelemzés az utóbbi években jelentős teret nyert. Ez a tudományág, mint már láthattuk, a strukturálatlan szöveges adatokból értelmes meglátások, minták és tudás kinyerésével foglalkozik.

A Python egy sokoldalú és nagy teljesítményű programozási nyelv, a szövegelemzés területén dolgozó szakemberek és kutatók egyik legelterjedtebb eszközévé vált. A Python szövegelemzésre való felhasználása robusztus és hatékony megközelítést kínál a hatalmas mennyiségű szöveges adat feldolgozásához, elemzéséhez és értékes információk kinyeréséhez. A Python-hoz átfogó könyvtárak és eszközök állnak rendelkezésre, az érzelmek elemzésétől és a nevesített entitásfelismeréstől kezdve a dokumentum klaszterezésen és a témamodellezésen át, így ideális választásnak tartottam az elemzéseim elkészítéséhez.

2.5.2. Az alkalmazott megközelítés

Dipanjan Sarkar „Text Analytics with Python: A Practitioner’s Guide to Natural Language Processing” [14] című könyvét kiválóan tudtam hasznosítani az NLP témakörének gyakorlati szempontú megismeréséhez. A könyv a későbbiekben részletezésre kerülő statisztikai és gépi tanulásos szövegelemzésekhez is remek fogódzóként szolgált.

Mint azt a bevezetőben már megállapítottam, a modern társadalom számára a hatalmas információmennyiség kezelése egyre komolyabb, és egyre több haszonnal kecsegtető kihívást jelent, és az NLP kiemelkedő játszik ebben a folyamatban. Sarkar könyvének gyakorlatias megközelítése révén rálátásom nyílt a szövegekből való információkinyerés, dokumentumösszegzés és témamodellezés alapvető fogalmaira.

A könyv részletezi a különböző technikákat, amelyek lehetővé teszik az információk hatékony kinyerését a szöveges adatokból. Az NLP terén elért eddigi eredmények felhasználása segíti az információk gyorsabb és hatékonyabb feldolgozását, elősegítve a releváns tartalmak azonosítását a rengeteg redundáns és lényegtelen adat között. A szinguláris értékbontás és az alacsony rangú mátrixközelítés matematikai elveit felhasználva bemutatható, hogyan lehet integrálni az NLP algoritmusokat egy adott szövegkorpusz mélyebb feldolgozásába.

A feldolgozás során, mint azt hamarosan látni fogjuk, részletesen elemeztem az NLP problémakör három stratégiáját: a statisztikai elemzéseket, a témák szerinti modellezést és automatikus dokumentum összefoglalást kulcskifejezésekkel és párbeszédekkel. Az NLP-ben alkalmazott fejlett kulcskifejezés összegyűjtési módszerek, illetve a téma szerinti klaszterezési technikák segítik a nagy adatmennyiségben való tájékozódást és a releváns információk azonosítását. Azt is részletezni fogom, hogyan lehetséges finomítani a témák szerint kialakított klasztereket, és hogyan lehet meghatározni szövegfüggően az optimális témák számát.

Természetesen a kapott eredmények értelmezése, megértése és felhasználása is kritikus szempont. Az automatizált dokumentum-összefoglalás kapcsán ismertetett látens szemantikai elemzés és a TextRank módszerek gyakorlati példákon keresztül segítik majd megérteni, hogyan alkalmazhatók ezek a módszerek valós adatokon, és milyen hasznos eredmények érhetők el velük.

Bár már valamivel régebbi forrás, viszont az általam fejlesztett alkalmazáshoz szintén hasznos vezérfonalat nyújtott a "Natural Language Processing with Python" című könyv [15], amely a fentebb említett mű mellett kiváló alapot szolgáltat a feladatmegoldásom során gyakran használt NLTK (Natural Language Toolkit) nevű könyvtár eszközkészletének megértéséhez.

Kutatásom továbbá kiterjedt a különböző témamodellezési technikák fejlesztésére is, beleértve a latens szemantikus indexelést, a latens Dirichlet-kiosztást, és a nemrégiben bevezetett nem-negatív mátrix-faktorizációt a Gensim és a Scikit-Learn keretrendszerek segítségével. Az NLTK-ról és az utóbb említett könyvtárakról a következő fejezetben lesz szó részletesebben.

2.6. Összehasonlító nyelvelemzés

A gépi tanulás és a természetes nyelvi feldolgozás (NLP) számos átalakító eszközt ad nekünk az összehasonlító nyelvi elemzéshez. Ez új lehetőséggel szolgál a kutatók számáraba különböző adathalmazok nyelvi árnyalatainak mélyebb vizsgálatára. Ezek a technológiák a minták kinyerésének automatizálásával megkönnyítik a nyelvhasználat hasonlóságainak és különbségeinek azonosítását, és egyszerűsítik az összehasonlító elemzés folyamatát. Két fontosabb, új eredményt szereték itt megemlíteni, amelyek bizonyítják az imént megfogalmazott állításokat.

Az első, fontos tanulmány 2013-ból származik, amelyben Mikolov és társai bemutatták a Word2Vec-et, egy úttörőnek számító modellt a természetes nyelvi feldolgozás témakörében [16]. A tanulmány a szóreprezentációk vektortérben történő hatékony becslésével foglalkozik. Részletesen bemutatják a Word2Vec skip-gram és „continuous bag of words” (CBOW) architektúráját. Kiderül, hogy a modell hogyan ragadja meg a szemantikai kapcsolatokat a szavak többdimenziós vektortérbe való leképezésével. Ez lehetővé teszi a szemantikai hasonlóság mérését és felszínre hozza a nyelvi árnyalatokat. A tanulmány jelentősen befolyásolta a későbbi kutatásokat a témakörben, és alapját képezte az NLP hatékonyabb alkalmazásának. Ide tartozik például az összehasonlító nyelvi elemzés és az időbeli nyelvi változások folyamatának megértése is.

A másik érdekes publikációt Gulordava és munkatársai adták közre 2018-ban. Ők az emberi nyelv hierarchikus szerkezetét vizsgálták rekurrens neurális hálózatok (RNN) segítségével [17]. A tanulmány mélyebben elemzi a neurális hálózatoknak azt a képességét, hogy szintaktikai struktúrákat tanuljanak és reprezentáljanak. Ezen kívül bemutatnak olyan módszereket, amelyekkel e hálózatok „szondázhatók”, azzal a céllal, hogy megértsük a hierarchikusan kódolt nyelvi információkat. Bár a munka elsősorban a szintaktikai struktúrákra összpontosít, emellett azonban értékes betekintést nyújt a fejlett neurális hálózati modellek segítségével végzett összehasonlító nyelvi elemzés területére is. Kutatásuk hozzájárul annak megértéséhez, hogy a deep learning modellek hogyan rögzítik és dolgozzák fel a különböző nyelvekben a hierarchikus nyelvi jellemzőket.

# 3. Python fejlesztőkörnyezet összeállítása

3.1. Programozási nyelvek [18] [19] [20] [21]

Szerencsére az elmúlt évtizedek során programozási nyelvek rendkívül sokszínű tárháza vált elérhetővé, amelyek különféle feladatokra és projektekre lettek kialakítva. Mielőtt rátérnék az általam választott Python fejlesztőkörnyezet bemutatására és részletezésére, szeretnék egy átfogó áttekintést nyújtani néhány ismertebb, szélesebb körben használatos programnyelvről, amelyek hasznosnak bizonyulhatnak gépi tanulással kapcsolatos feladatok, elemzések elvégzéséhez. A listát nem más nyelvvel, mint a Pythonnal kezdem.

3.1.1. A Python

A Python nyelv a sokoldalúságával és könnyen érthető szintaxisával emelhető ki a többi programnyelv közül. Alkalmazási területei közé tartozik például a webfejlesztés, de általánosan adatelemzési és gépi tanulási feladatokra használják. Az olvasható, letisztult kód és a gyors fejlesztési idő miatt ideális lehet mind a kezdő, mind a tapasztaltabb fejlesztők számára. Gazdag rendelkezésre álló könyvtárkészletét és széleskörű közösségi támogatását érdemes még kiemelni, amelyek fontos érvek voltak a Python használata mellett.

Egyes statisztikák és elemzők szerint a teljesítménye nem mindig a legkiemelkedőbb, különösen erőforrásigényesebb alkalmazásokban (lásd a hivatkozásokat).

3.1.2. Az R nyelv

Az R programozási nyelv legfontosabb felhasználási területei közé küzé tartozik a statisztikai analízis és adatvizualizáció. A statisztikai modellezés és a gazdag vizualizációs lehetőségek révén segíti az adatelemzési feladatokat. Hátrányai közé sorolható például, hogy a szintaxisa néha bonyolult lehet, és inkább a tisztán statisztikai területeken találhatók meg a fő alkalmazási területei.

3.1.3. A C++

A C++ nyelv legerősebb tulajdonságai közé tartozik a hatékonyság és a teljesítmény, ezeken a területeken emelkedik ki elsősorban. Leginkább rendszerprogramozásban és játékfejlesztésben alkalmazzák, és rendkívül erőteljes és sokoldalú nyelvnek számít. Mindezek ellenére a C++ alaposabb megtanulásához szükséges idő hosszabb lehet, illetve a manuális memóriakezelés nagy komplexitást eredményezhet.

3.1.4. A Java

A Java platformfüggetlen, így ideális választás nagy rendszerek és vállalati alkalmazások fejlesztésére. Robosztus és biztonságos. Hátránya, hogy a kód hosszabb, terjengősebb lehet, és a bonyolult lekérdezések néha kihívást jelenthetnek.

3.1.5. A JavaScript

A JavaScript a kliensoldali webfejlesztés alapnyelve. Dinamikus funkciókat hoz a böngészőkbe. A különböző böngészők támogatása miatt előfordulhatnak kompatibilitási problémák.

3.1.6. Összehasonlítás

Ha a Python főbb jellemzőit szeretnénk összehasonlítani a többi felsorolt programnyelv tulajdonságaival, akkor a következő fontosabb megállapításokra juthatunk:

Ha az R nyelvvel hasonlítjuk össze, akkor a Python általánosabb feladatokra, így gépi tanulásra, de akár webfejlesztésre is ideális lehet. Az R inkább statisztikai vizsgálatokra specializálódott, jó vizualizációs képességekkel.

A C++ nyelvhez képest a Python egyszerűbb, gyorsabb benne a fejlesztés, emellett sokoldalú is. Ezzel szemben előbbi inkább teljesítményorientált, komplex és erőteljes hardverközeli fejlesztést tesz lehetővé.

Ha a Java nyelvvel kellene szembeállítanom a Pythont, akkor azt emelném ki, hogy a Python egyszerű szintaxissal rendelkezik, és emellett széles körű alkalmazhatóságot kínál. A Java nagy előnye szintén a platformfüggetlenség, ami vállalati alkalmazások, robosztus rendszerek fejlesztésénél hasznos, viszont mint már volt róla szó, terjengősebb a szintaktikáj, mint a Pythonnak.

Ha a JavaScript viszonylatában vizsgálódunk, akkor abban lehet a fő különbséget meglátni, hogy a Python inkább a szerveroldali fejlesztéshez alkalmasabb, illetve a gépi tanulás, és a ttudományos számításokhoz testhezállóbb. A JavaScript pedig eredetileg kliensoldali webfejlesztésre lett optimalizálva.

Ezeknél az összehasonlításoknál fontos hangsúlyozni, hogy természetesen minden programnyelvnek megvannak a maga előnyei és ideális alkalmazási területei. Az optimális választás mindig a konkrét feladat vagy projekt céljától és a követelményektől függ. Emellett egy adott feladat megoldására többféle programozási nyelv is alkalmas lehet.

3.1.7. A Python előnyei

Konklúzióként az alábbi előnyök sorolhatók fel, amelyek arra ösztönöztek, hogy a Pythont válasszam a szakdolgozatom szövegelemzési feladatainak programozási nyelveként: a Pythonban írt kódok könnyen értelmezhetők és olvashatók, és ez elősegíti a hatékony munkát és a karbantarthatóságot. Emellett a Python a webfejlesztéstől kezdve, adatelemzésen át egészen a gépi tanulásig sokféle területen használható. Továbbá gazdag és sokoldalú könyvtárkészlettel rendelkezik, többek között az NLP területén, amely jelentősen felgyorsítja a fejlesztést. A Python szerencsére aktív közösséggel rendelkezik, könnyen hozzáférhetők dokumentációk, segítségek és fejlesztői erőforrások.

Ezek a tulajdonságok együttesen teszik a Pythont az egyik legkedveltebb választássá a fejlesztők és tudósok körében. Mint már említettem, a jól olvasható kódja és a gyors fejlesztési idő nagy előnyt biztosít, amikor gépi tanulásos problémákat, feladatokat kell megoldani, implementálni, illetve a megoldások különböző adatkészletekkel és paraméterekkel való tesztelésében.

3.2. Fontosabb Python fejlesztőkörnyezetek [22] [23] [24]

Most, hogy a választott programozási nyelv már adott, a Python fejlesztőkörnyezetünk (IDE) kiválasztása szintén fontos lépés a hatékony fejlesztéshez, és számos lehetőség közül választhatunk a projektek jellegétől függően. Az alább összegyűjtött Python IDE-k mindegyikében találhatunk előnyös és hátrányos tulajdonságokat. Igyekeztem kiemelni, hogy miért esett végül a választásom a Spyder-re. De lássuk az általam vizsgált fejlesztői környezetek legfontosabb aspektusait.

3.2.1. DataCamp Workspace

A DataCamp Workspace egy olyan Python fejlesztőkörnyezet, amelyet kifejezetten az online tanuláshoz terveztek. A platform interaktív kurzusokat és gyakorlatokat ad, amelyekhez beépített fejlesztői környezet áll rendelkezésre. Ennek előnye, hogy a felhasználók azonnal alkalmazhatják a tanultakat a gyakorlatban, anélkül, hogy saját fejlesztőkörnyezetet kellene telepíteniük. Könnyen elérhetők a kurzusokhoz kapcsolódó kódok és gyakorlatok, és azonnal láthatók a változtatások eredményei.

Hátránya az offline használat korlátozottsága. Továbbá a platformot elsősorban oktatási célokra fejlesztették ki, a fejlesztőknek kevésbé van lehetőségük testre szabni a környezetet az egyéni igényeikhez. Fontos szempont az is, hogy az online környezetek általában kisebb teljesítményűek lehetnek, különösen nagyobb projektek vagy adatfeldolgozási feladatok esetén.

3.2.2. Jupyter Notebook

A Jupyter Notebook egy interaktív környezet, amely lehetővé teszi a kód, a szöveg és a képek kombinálását egy dokumentumban. Széles körben alkalmazzák tudományos kutatásokhoz és adatelemzési projektek során. Az interaktivitás révén a felhasználók lépésről lépésre fejleszthetik és dokumentálhatják a kódot, miközben megmarad a projekt is. Többek között a Python, R és Julia programozási nyelveket támogatja a platform, így sokoldalúan használható tudományos tevékenységekre. A Jupyter Notebook nem mindig optimális nagyobb projektek létrehozására.

3.2.3. PyCharm

A PyCharm-ot a JetBrains kifejezetten Python fejlesztéshez tervezte. Fontos funkciói közé tartozik a kódok ellenőrzése, a refaktorálás és a kódelemzés. A PyCharm támogatja például a Django és a Flask keretrendszereket. Az integrált debugger és a beépített unit tesztelési eszközök segítik a hibakeresést és a kódminőség fenntartását. A teljes verzió díjköteles, bár létezik ingyenes verzió is.

3.2.4. Google Colaboratory

A Google Colaboratory egy felhőalapú Jupiter Notebook környezet, amely a Google fiókkal rendelkező felhasználóknak ingyenesen elérhető. Ideális a közös munkához és a projektek megosztásához, mivel a felhasználók egyszerűen megoszthatják a Colab dokumentumaikat, és közösen dolgozhatnak rajtuk. A Colab GPU-támogatása lehetővé teszi a gyors és hatékony gépi tanulási projektek futtatását a felhőben. Azonban használata korlátozott a Google Drive-hoz való kötöttség miatt.

3.2.5. Visual Studio Code

A Visual Studio Code egy könnyűsúlyú, mégis erőteljes kódszerkesztő, amely jól kiterjeszthető és testreszabható. Támogatja a Python programozási nyelvet, és számos kiterjesztés közül választhatunk, amelyek segítenek a produktivitás növelésében. Gyors és reszponzív, és alkalmas kisebb projektekhez és nagyobb alkalmazásokhoz egyaránt. Hátrányként említhető, hogy levésbé hangsúlyozza a Python-specifikus fejlesztést.

3.2.6. Atom

Az Atom egy ingyenes, nyílt forráskódú szövegszerkesztő, amelyet a GitHub fejlesztett ki. Könnyen testreszabható és bővíthető, és számos kiterjesztéssel szolgál a Python fejlesztők számára. A közösség által fejlesztett kiegészítők révén személyre szabható a felhasználók igényeinek megfelelően, és segíti a hatékony kódszerkesztést.

3.2.7. Thonny

A Thonny egy könnyen használható fejlesztőkörnyezet, kifejezetten a kezdő programozók számára tervezték meg. Ez az egyszerűsített környezet tartalmazza a Python interpretert és egy alapvető kódszerkesztőt, így minimálisra csökkenti a konfigurálási folyamatot. A Thonny ideális választás lehet azoknak, akik gyorsan el szeretnének kezdeni a Python programozást, anélkül, hogy sok időt töltenének beállításokkal és konfigurációval. Azonban kevésbé alkalmazható nagyobb projektekhez.

3.2.8. JupyterLab

A JupyterLab a Jupyter Notebook kiterjesztése, további funkcionalitásokkal és testreszabási lehetőségekkel. Ez a környezet lehetővé teszi a felhasználók számára, hogy több ablakban és lapon dolgozhassanak egyszerre, optimalizálva ezzel a munkafolyamatot. A beépített fájlkezelő, kiterjesztett widgetek és tematikus kiegészítők sorolhatók még az eszköztárához. Nem támogat minden Jupyter Notebook kiegészítőt.

3.2.9. DataSpell

A DataSpell egy integrált fejlesztőkörnyezet, amely támogatja a nagy adathalmazokon végzett feldolgozásokat és az adatvizualizációt. Ez a környezet magában foglalja a Jupyter Notebookot és olyan egyéb eszközöket, amelyek az adatelemzésben és gépi tanulási projektekben nyújtanak segítséget. A DataSpell gyors kódellenőrző és hibakereső eszközökkel rendelkezik, elősegítve ezzel a hatékony fejlesztést. Hátránya a kisebb ismertség és szűkebb közösség.

3.2.10. Spyder

A Spyder (Scientific Python Development Environment) egy olyan tudományos számítási fejlesztőkörnyezet, amelyet kifejezetten adatvizualizációs és adatelemzési feladatokhoz terveztek. Az előnyei közé tartozik a könnyű telepítés és konfiguráció, valamint a gyors és interaktív adatvizualizáció lehetősége. A Spyder tartalmazza a SciPy, NumPy és a Matplotlib könyvtárakat, ami megkönnyíti az adatelemző munkát. Az Ipython konzol mellett a beépített változóinspektor és a változók értékeinek vizualizációja segíti a változók állapotának folyamatos nyomon követését és a hibakeresést. Emellett a Spyder támogatja a Jupyter Notebookot, így a projektjeinkben kombinálhatjuk a két környezet előnyeit. A Spyder további funkcionalitásokkal bővíthető a kiterjeszthető plugin rendszer segítségével, így az egyéni igényeinkhez testreszabható fejlesztőkörnyezetet kapunk.

A Spyder előnyei közé tartozik a könnyű használat, az adatelemző eszközök gazdagsága és az IPython konzol által biztosított interaktivitás. Ideális választás lehet azok számára, akik az adatelemzésben és a tudományos számításokat szeretnének végezni, és egy olyan fejlesztőkörnyezetet keresnek, amely támogatja mind a fejlesztési, mind az elemzési folyamatokat. Ugyanakkor fontos azt is megemlíteni, hogy a Spyder közössége és bővíthetősége kisebb lehet más, nagyobb közösséggel rendelkező IDE-khez képest.

Mindezt összefoglalva, azért esett végül a választásom a Spyder használatára, mert elég kompaktnak és jól felépítettnek éreztem a moduljait, amelyekkel hatékonyan tudtam végigkövetni, illetve tesztelni a fejlesztés során fellépő köztes állapotokat. Végül az elkészült elemzések ereményeit is integráltan kézhez kaptam. Ezt a folyamatos interaktivitást rendkívül produktívnak éreztem. A sokféle rendelkezésre álló modul ellenére szerintem kellően letisztult a felhasználói felület. Végig azt éreztem, hogy jól kezemben tudom tartani a különböző funkciókat, és teljeskörű támogatást kapok a kódjaim karbantartásához, teszteléséhez, és az eredmények kiértékeléséhez. Konklúzióként kijelenthető, hogy a Spyder a tudományos jellegű Python fejlesztés terén hasznos eszköz, különösen, ha nagyobb mennyiségű adat elemzésére, interaktív feldolgozására és vizualizációjára van szükség.

3.3. Függvénykönyvtárak [25] [26] [27]

Most pedig térjünk rá azoknak a Python függvénykönyvtáraknak a bemutatására, amelyeket vizsgálataim során alkalmaztam. Ezek a könyvtárak kifejezetten az NLP területén végzett kutatómunkát támogatják, stabil és kiterjedt matematikai hátteret adnak a kívánt szövegelemzések elvégzéséhez.

3.3.1. Matplotlib [28]

A Matplotlib fejlett matematikai algoritmusokat alkalmaz látványos és informatív ábrák létrehozásához, adatvizualizációhoz. Tartalmaz algoritmusokat adatpontok interpolációjához, görbeillesztéshez és statisztikai vizualizációhoz. A könyvtár integrációja a NumPy-al hatékony numerikus műveletekre ad lehetőséget, ami növeli a kigenerált vizualizációk matematikai hasznosságát.

3.3.2. Pandas [29]

A Pandas a statisztikai modellezés és az idősor elemzés integrálásával bővíti matematikai eszköztárunkat. A könyvtár tartalmaz gördülő statisztikákhoz, exponenciálisan súlyozott mozgóátlagokhoz és autokorrelációhoz készült függvényeket, ezzel időfüggő adatok elemzését teszi lehetővé. Emellett a Pandas matematikai optimalizációs algoritmusokat használ a hatékony adatkiigazításhoz és aggregáláshoz, optimalizálva a számítási teljesítményt.

3.3.3. spaCy [30]

A spaCy fejlett matematikai modelleket, például transzformátor-architektúrákat használ. Figyelemmechanizmusok és mély neurális hálózatok alkotják a nevesített entitások felismerésének és a beszédrészek címkézési képességeinek alapját. A könyvtár matematikai kifinomultsága kiterjed a beágyazásokra is, ami lehetővé teszi a szavak és kifejezések folyamatos vektortérben történő reprezentációját, így képessé téve arra, hogy megragadja a szemantikai kapcsolatokat.

3.3.4. NLTK (Natural Language Toolkit) [15]

Az NLTK túlmutat az alapvető nyelvfeldolgozáson, mivel olyan matematikai eszközökkel rendelkezik, mint az entrópia-alapú jellemzőválasztás információelmélete és az olyan valószínűségi modellek, mint a Naive Bayes osztályozó. A különböző távolságmérők implementációja lehetővé teszi a hasonlóságelemzést és a klaszterezést a szövegbeli különbözőség matematikai mérőszámai alapján. Az NLTK sokoldalúságát az adja, hogy statisztikai technikákat használ a szövegosztályozáshoz és a hangulatelemzéshez.

3.3.5. NumPy [31]

A NumPy matematikai arzenálja kiterjed a lineáris algebrára, illetve szinguláris értékek felbontását, mátrix faktorizációt és sajátérték-számításokat tartalmaz. Fourier-transzformációs funkciói megkönnyítik a frekvenciatartományok elemzését, míg a numerikus integrációs és differenciálási eszközök hozzájárulnak a tudományos szimulációkhoz. A NumPy átfogó matematikai függvénykészlete megalapozza a nagy léptékű tudományos számításokhoz szükséges numerikus hatékonyságot.

3.3.6. Gensim [32]

A Gensim továbbfejleszti matematikai eszköztárát az online tanuláshoz és a szöveges adatok áramlásához szükséges inkrementális algoritmusokkal. Megvalósítja a latens szemantikai elemzés (LSA) és a latens Dirichlet-allokáció (LDA) algoritmusait, amelyekkel közelebb kerülünk a szemantikai kapcsolatok szövegből való kinyeréséhez. Mátrix-faktorizációs technikákat foglal magában a témamodellezéshez és a hasonlóságelemzéshez, illetve matematikai optimalizálási módszereket használ a modellképzéshez. A Gensim képes nagy szövegkorpuszok látens struktúráinak megragadására, ami értékes eszközzé teszi a rejtett szemantikai kapcsolatok feltárása kapcsán.

3.3.7. scikit-learn [33]

A Scikit-learn matematikai alapja magában foglalja a modellképzés és paraméterhangolás optimalizálási technikáit. A többdimenziós skálázáson és spektrális beágyazáson alapuló sokrétű tanulási algoritmusok implementációja geometriai perspektívát ad a többdimenziós adatelemzéshez. A scikit-learn továbbá integrálja a hipotézisvizsgálat és a modellértékelés statisztikai módszereit, így átfogó csomagot nyújt a gépi tanulással foglalkozó szakemberek számára.

3.3.8. NetworkX [34]

A NetworkX eszköztárába tartoik a fejlett gráfelméleten alapuló moduláris optimalizáció és a spektrális klaszterezésen alapuló közösségfelismerő algoritmusok. Péládul a Katz-centralitás és a sajátvektor-centralitás segítségével képes számszerűsíteni egy hálózatban található csomópontok fontossági súlyát. A könyvtár matematikai gazdagsága kiterjed a hálózati motívumok azonosítására és az irányított gráfok maximális folyamának megtalálására szolgáló algoritmusokra is. Mindez mutatja a könyvtár komplex hálózatelemzésben való hasznosságát.

3.3.9. SciPy [35]

A SciPy tovább bővíti matematikai eszköztárunkat optimalizációs algoritmusokkal, például a korlátozott és korlátlan minimalizálási technikákkal. A könyvtár jelfeldolgozó modulja szűrési és spektrumanalízis módszereket integrál, illetve statisztikai funkciói hipotézistesztek és valószínűségi eloszlások széles skáláját ölelik fel. A SciPy a tudományos számítások és adatelemzés átfogó eszközeként szerepel a palettán.

Összefoglalva, ezek a Python-könyvtárak együttesen hatékony környezetet adnak a tudományos számításokhoz, az adatelemzéshez és a természetes nyelvi feldolgozáshoz, felhasználva a rendelkezésre álló matematikai eszközöket és algoritmusokat. Az idézett irodalmi hivatkozások további betekintést nyújtanak e könyvtárak fejlesztésébe és alkalmazásába.

# 4. A Spyder konfigurációja

…

# 5. Szövegkorpuszok összegyűjtése és statisztikai elemzése

5.1. A szövegek kiválasztása

A statisztikai és gépi tanulásos vizsgálataim elvégzéséhez a Harry Potter könyvsorozat első kötetének, a Harry Potter és a Bölcsek Köve című regénynek szövegét választottam ki. Ez már kellően nagy szövegkorpusznak bizonyult a releváns feldolgozások számára. Első lépésként felkutattam a kötet angol, német, francia és spanyol nyelvű, interneten elérhető változatait pdf formátumban.

5.2 A statisztikai vizsgálatok célja

A négy különböző nyelvű szöveg statisztikai elemzésének egyrészt az volt a célja, hogy összehasonlíthatók legyenek az egyes nyelvek felépítési sajátosságai. Másrészt a könyvben előforduló szókincs vizsgálata olyan szempontból is érdekes, hogy hogyan segítheti a nyelvtanulást a szavak gyakoriság szerinti összegyűjtése.

A következő vizsgálatokat végeztem el mind a 4 nyelvű változatra: a szövegek sorainak a száma, a szavak száma, a szavak hosszának előfordulása, az első 100 szó listázása gyakoriság szerint, illetve ugyanezek a vizsgálatok fejezetenként. Ezekből az információkból több következtetést is levonhatunk a könyvre vonatkozóan:

A szöveg sorainak és szavainak száma segít felmérni a könyv hosszát és általános terjedelmét. A mondatok számából kiderülhet, hogy inkább jobban tagoltak, vagy összetettek a mondatok. Az egyes szavak hosszának előfordulása meghatározhatja a szókincs gazdagságát és hogy mennyire lehetnek jelen szakterület-specifikus vagy nehéz szavak a szövegben. Így jobban meghatározhatjuk egy szöveg értelmezhetőségét. A szavak eloszlásának fejezetenkénti változása információt nyújthat a fejezetek különböző témáiról vagy fókuszairól.

Ha ezek az információk több különböző nyelvű változat esetén is rendelkezésre állnak, az további előnyökkel jár a nyelvtanulás szempontjából. Például a különböző nyelvű változatok segíthetnek a szóhasználat és a kifejezések összehasonlításában, ami elősegítheti a szókincs bővítését és hatékonyabbá teheti a nyelvtanulást. A különböző nyelveken írt változatok összehasonlítása lehetővé teszi a nyelvtani szerkezetek és kifejezések jobb megértését, ami segíthet a nyelvtanulóknak, hogy jobban megértsék egy adott nyelv sajátosságait. Végül a különböző nyelvű változatok megismerése által a nyelvtanulók bepillantást nyerhetnek a különböző kultúrákba és nyelvi kontextusokba is, amelyek esetleg nem lennének olyan szembetűnőek, ha szeparáltan vizsgálnánk a különböző nyelvű változatokat.

5.2. A választott szövegek előfeldolgozása

A szövegkorpusz meghatározása után megkezdődhetett a szöveg elemezhető formába alakítása. Először is a begyűjtött ’pdf’-eket ’txt’ formátumba konvertáltam[[1]](#footnote-1). Ezután a kapott ’txt’ fájlokat megvizsgálva arra a következtetésre jutottam, hogy a kapott négy különböző nyelvű szövegnek korántsem egységes a tagolása, így az általános előfeldolgozási lépések mellett speciális, csak egy-egy adott dokumentumra alkalmazható formázásra is szükség lesz.

Az előfeldolgozást először még egyszerű Python kóddal, illetve a beépített ’re’ modul segítségével kezdtem el, ami reguláris kifejezésekkel való operációkat tartalmaz. Ez a fázis egyben a Python nyelvvel való ismerkedésként is szolgált.

Az angol szövegben elég sok helyen fordult elő, hogy bizonyos sorok végén elválasztásjellel kerültek elválasztásra bizonyos szavak. Ezeket a szavakat egyesítenem kellett.

# 6. A könyv szövegének tartalmi összefoglalója

# 7. Témakörök szerinti elemzés

# 8. Paraméterek vizsgálata, a kapott eredmények rendszerezése

# 8. Összefoglaló

# 9. Summary

# 10. Irodalomjegyzék

[1] Szalavetz Andrea. Mesterséges intelligencia és technológiavezérelt termelékenységemelkedés. Külgazdaság, 2019, 63.7-8: 53-79.

[2] Noam Chomsky: Syntactic Structures. Mouton & Co. 1957.

[3] Church, Kenneth. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In Proceedings of the Second Conference on Applied Natural Language Processing (pp. 136-143).

[4] Pang, Bo; Lee, Lillian. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.

[5] Tjong Kim Sang, Erik F.; De Meulder, Fien. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 (pp. 142-147).

[6] Sutskever, Ilya; Vinyals, Oriol; Le, Quoc V. (2014). Sequence to Sequence Learning with Neural Networks. In Advances in Neural Information Processing Systems 27 (pp. 3104-3112).

[7] Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

[8] Hochreiter, Sepp; Schmidhuber, Jürgen. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

[9] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; Polosukhin, Illia. (2017). Attention is All You Need. In Advances in Neural Information Processing Systems 30 (pp. 5998-6008).

[10] Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[11] Radford, Alec; Narasimhan, Karthik; Salimans, Tim; Sutskever, Ilya. (2018). Improving Language Understanding by Generative Pretraining. arXiv preprint arXiv:1701.00160.

[12] Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James Y.; Saligrama, Venkatesh; Kalai, Adam T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Advances in Neural Information Processing Systems 29 (pp. 4349-4357).

[13] Radford, Alec; Chen, Jonathon; et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.

[14] Dipanjan Sarkar: Text Analytics with Python. A Practitioner’s Guide to Natural Language Processing. Second Edition. Apress, Bangalore, Karnataka, India, 2019

[15] Steven Bird, Ewan Klein, and Edward Loper: Natural Language Processing with Python. O’Reilly, 2009

[16] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013)

[15] Parisotto, Emilio; Salakhutdinov, Ruslan. (2017). Neural Map: Structured Memory for Deep Reinforcement Learning. arXiv preprint arXiv:1702.08360.

[16] Gabriel Wittum, Michael Hoffer, Babett Lemke, Robert Jabs, Arne Nägel: Automated methods for the comparison of natural languages. Computing and Visualization in Science, 2020. 23:7, <https://doi.org/10.1007/s00791-020-00325-2>

[17] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, Marco Baroni: Colorless green recurrent networks dream hierarchically. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018.

Programnyelvekhez:

[18] https://careerfoundry.com/en/blog/data-analytics/best-machine-learning-languages/#what-skills-are-important-for-machine-learning

[19] https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/

[20] https://www.springboard.com/blog/data-science/best-language-for-machine-learning/

[21] <https://neptune.ai/blog/programming-languages-machine-learning>

Fejlesztőkörnyezetekhez:

[22] https://www.datacamp.com/tutorial/data-science-python-ide

[23] https://www.projectpro.io/article/best-python-ide-for-data-science-and-machine-learning/812#:~:text=PyCharm%20is%20one%20of%20the,syntax%20highlighting%2C%20and%20error%20checking.

[24] <https://www.upgrad.com/blog/python-ides-for-data-science-machine-learning/>

Library-khez:

[25] https://medium.com/activewizards-machine-learning-company/comparison-of-top-6-python-nlp-libraries-c4ce160237eb

[26] https://sunscrapers.com/blog/9-best-python-natural-language-processing-nlp/#textblob

[27] <https://bitbytesoft.com/top-nlp-libraries-python/#CoreNLP>

[28] John D. Hunter: Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

[29] WesMcKinney: Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

[30] a spaCy hivalos weboldala: <https://spacy.io/usage/spacy-101>

[31] Stefan van der Walt, S. Chris Colbert, Gael Varoquaux: The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, 13(2), 22-30.

[32] Radim Řehůřek, Petr Sojka: Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45-50.

[33] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[34] Aric A. Hagberg, Daniel A. Schult, Pieter J. Swart: Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference, 11-15.

[35] Eric Jones, Travis Oliphant, Pearu Peterson: SciPy: Open Source Scientific Tools for Python. 2001.

# 10. Mellékletek

1. A konvertálást a <https://www.pdf2go.com/pdf-to-text> oldal segítségével végeztem el. Utolsó elérés időpontja: 2023.11.29. [↑](#footnote-ref-1)