# 1. Bevezető

Mérnökinformatikai egyetemi tanulmányaim során a korszerű web-technológiák szakirányt választottam, és érdeklődöm is a webes világ iránt, illetve a munkámban is ilyen technológiákkal foglalkozom. Ennek ellenére megfogalmazódott bennem egy olyan gondolat, hogy szakdolgozatomnak olyan témát kellene választanom, amely nem egy „hagyományos” webes alkalmazás elkészítéséből és bemutatásából áll. Emellett a dolgozatom alapötletét az a gondolat is alakította, hogy valahogyan szerettem volna összekötni a programozás világát a szövegek elemzésével, mivel a nyelvek és a nyelvtan nagy rajongója vagyok. Továbbá az is segítette az ilyen irányú orientálódásomat, hogy történész végzettséggel is rendelkezem, így közel áll hozzám a hosszabb terjedelmű szövegek, tanulmányok olvasása, értelmezése. Mindez ahhoz vezetett, hogy a természetes nyelvű szövegek feldolgozása (NLP, vagyis Natural Language Processing) területén választottam témát. Ez pedig nem más, mint hosszabb természetes nyelvű szövegek témakör szerinti elemzése gépi tanulás segítségével.

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]](#footnote-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 számunkra, és ez lehetőséget teremtett a különböző kultúrák mélyebb megismerésére. Ugyanakkor ezeknek a szövegeknek a nagysága sokszor szinte áttekinthetetlen, és manuális elemzésük nemcsak időigényes, de általában teljeskörűen lehetetlen is. Ezen a ponton kap szerepet a gépi tanulás hatékonysága, amely lehetővé teszi az összetett információk rendszerezett és gyors feldolgozását.

A szakdolgozatom középpontjában egy olyan eszköz kifejlesztése állt, amely amellett, hogy statisztikai elemzésekkel segíti az idegen nyelvek tanulását, a szövegeket automatizált módon témakörök mentén csoportosítja. A vizsgálat fókuszában egy klaszterezési probléma áll, 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 (TODO: csak a statisztikai készült el, vagy a klaszterezés is) vizsgálatát négy különböző nyelven készítettem el. Ezek az angol, német, francia és spanyol nyelven elérhető szövegek.

A dolgozatomban nemcsak a gépi tanulás különféle módszereinek alkalmazására koncentráltam, hanem igyekeztem rálátással szolgálni a természetes nyelvek elemzésének matematikai hátterére is. Továbbá áttekintést nyújtok azokról a már bevált megoldásokról, melyek sikeresen alkalmazhatók hasonló területeken.

Elemzéseim során a Python programozási nyelvet alkalmaztam, illetve felhasználtam az ehhez elérhető adatelemző függvénykönyvtárakat. Ezzel az erős technológiai bázissal egy olyan hatékony eszköz létrehozását kíséreltem meg, 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. (TODO: kéne vmi hivatkozás ide)

# 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]](#footnote-2), elméleti alapot adott a mondatok szintaktikai struktúráinak elemzésére és megértésére. Ezt 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]](#footnote-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 munkájukban a szövegosztályozás egyik részterületét, a hangulatelemzést vizsgálták[[4]](#footnote-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]](#footnote-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]](#footnote-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]](#footnote-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]](#footnote-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 hazsnálata sikeresen beépíthető olyan feladatokba is, mint a nyelvi modellezés és a gépi fordítás.

2.3.3. Transzformátorok és figyelemfelkeltő mechanizmusok

A Transformer architektúra Vaswani et al. (2017) általi bevezetése[[9]](#footnote-9) 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]](#footnote-10) (Devlin et al., 2018) a kontextuális szóbeágyazások és a GPT (Generative Pre-trained Transformers)[[11]](#footnote-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, és így a nyelvi adatokban akaratlanul is megjelenő társadalmi elfogultságok, előítéletek, kategorizálások megfelelő odafigyeléssel való kezelésének fontosságát[[12]](#footnote-12). Ha belegondounk, egyértelmű, hogy az NLP-rendszerek felelős fejlesztésére és figyelmes gondozására van szükség a lehetőségekhez mérten igazságos és elfogulatlan eredmények biztosítása érdekében.

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]](#footnote-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. Az érzelmek elemzésétől és a nevesített entitásfelismeréstől kezdve a dokumentum klaszterezésen és a témamodellezésen át a Python-hoz átfogó könyvtárak és eszközökállnak rendelkezésre, így ideális választásnak tartottam az elemzéseim elkészítéséhez.

2.5.2. A fő ihletforrások

Dipanjan Sarkar „Text Analytics with Python: A Practitioner’s Guide to Natural Language Processing”[[14]](#footnote-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 alkalmazás a "Natural Language Processing with Python" című könyvre is támaszkodik[[15]](#footnote-15), amely a fentebb említett mű mellett kiváló forrás az 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. (TODO: behivatkozni)

# 3. Python fejlesztőkörnyezet összeállítása

# 4. Szövegkorpuszok összegyűjtése elemzéshez

4.1. A szövegek kiválasztása

4.2. A választott szövegek előfeldolgozása

# 5. A könyv szövegének tartalmi összefoglalója

# 6. Témakörök szerinti elemzés

# 7. Paraméterek vizsgálata, a kapott eredmények rendszerezése

# 8. Összefoglaló

# 9. Summary

# 10. Irodalomjegyzék

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

Noam Chomsky: Syntactic Structures. Mouton & Co. 1957.

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).

Brown, Peter F.; Pietra, Vincent J. Della; Pietra, Stephen A. Della; Mercer, Robert L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19(2), 263-311.

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

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).

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).

Bahdanau, Dzmitry; Cho, Kyunghyun; Bengio, Yoshua. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the 3rd International Conference on Learning Representations.

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

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

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).

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.

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

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).

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

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

Dömösi, Pál, Falucskai, János, Horváth, Géza, Mecsei, Zoltán, Nagy, Benedek: Formális Nyelvek és Automaták. Forrás: <https://gyires.inf.unideb.hu/KMITT/b24/>

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

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

Sayan Putatunda: Practical Machine Learning for Streaming Data with Python. Design, Develop, and Validate Online Learning Models. Apress, Bangalore, India, 2021

# 10. Mellékletek

1. Szalavecz 2019. [↑](#footnote-ref-1)
2. Chomsky, 1957. [↑](#footnote-ref-2)
3. Church, 1988. [↑](#footnote-ref-3)
4. Pang et al. 1988. [↑](#footnote-ref-4)
5. Tjong Kim et al, 2003. [↑](#footnote-ref-5)
6. Sutskever et al, 2014; Bahdanau et al, 2014. [↑](#footnote-ref-6)
7. Mikolov et al, 2013. [↑](#footnote-ref-7)
8. Hochreiter, Schmidhuber, 1997. [↑](#footnote-ref-8)
9. Vaswani et al, 2017. [↑](#footnote-ref-9)
10. Devlin et al, 2018. [↑](#footnote-ref-10)
11. Radford et al, 2018. [↑](#footnote-ref-11)
12. Bolukbasi et al, 2016. [↑](#footnote-ref-12)
13. Radford et al, 2021. [↑](#footnote-ref-13)
14. Dipanjan, 2019. [↑](#footnote-ref-14)
15. Steven et al, 2009. [↑](#footnote-ref-15)