

TDK-dolgozat

Sánta Péter

Szerző-attribúció gépi tanulási eszközökkel

Sánta Péter

III. évf. programtervező informatikus BSc

Témavezetők: Kicsi András, Dr. Vidács László

MTA-SZTE Mesterséges Intelligencia Kutatócsoport

Tartalomjegyzék

Absztrakt.....	4
1. Bevezetés.....	5
2. A szerző-attribúció akadályai.....	7
3. Korábban próbált módszerek áttekintése	9
4. Saját kísérletek, tapasztalatok a korábbi módszerekkel	10
4.1 Egyszerű nyelvtani jellemzőkészlettel való tapasztalatok.....	10
4.2 Stabilitás alapú jellemzőhalmaz bemutatása	12
4.3 N-gram alapú jellemzőhalmaz bemutatása.....	13
4.3.1 A módszer kiértékelése.....	15
4.3.2 Kísérletek szógramokkal	16
4.3.3 Kísérletek karaktergramokkal.....	17
4.3.4 Kísérletek bájtgramokkal.....	18
4.3.5 Kísérletek a Stylo csomaggal.....	19
5. Mélytanuló illetve BERT modellekkel folytatott kísérletek	20
6. Összefoglalás, konklúzió.....	28
7. További irányok, kitekintés.....	29
Irodalomjegyzék.....	30

Absztrakt

Megállapítani két szövegről, hogy ugyanaz írta-e, egy közismert és gyakran felmerülő feladat, aláírásoktól egészen a kriminológia világáig terjed a felhasználása. Az egyik legegyszerűbb és legismertebb megközelítése a feladatnak a külalak, az íráskép alapján történő beazonosítás. Abban az esetben azonban, amikor ez nem elégséges, egy logikus következő lépés lehet a szövegben mutatkozó személyiségjegyek alapján azonosítani. A napjainkban jellemzőbb, digitális szövegek esetén pedig az utóbbi lépés különös fontossággal bír, a külalaki jellemző háttérbeszorulásával.

A kutatás centrális kérdése, hogy pontosan milyen jellemzők beszélnek a szöveg szerzőjéről, és ezeket pontosan hogyan is kezeljük. A feladatot gépi tanulási módszerekkel közelítem meg, hagyományos és korszerűbb módszereket egyaránt vizsgálva. Lineáris gépekkel és n-grammokkal, illetve különböző nyelvi jellemzőkkel kísérletezek, ezen felül BERT modellek teljesítményét is felmérem. A kutatás főképp korábbi, angol nyelvű kutatásokra épít a témából, ezeknek a magyar nyelvbe való integrálása a fő megközelitésem.

A pontos feladat, amivel a kutatás foglalkozik két fő részre osztható. Az egyik részben a feladat, hogy szerzők egy **zárt halmazából** kiválasszuk azt a szerzőt, aki a kérdéses szöveget írta.

A másik részben egy általános modellel is kísérletezünk, mely **a nyílt halmazos** verziót bővíti, annyival, hogy ha adott két új szöveg, azonos szerzőkkel, akkor azt a kettőt is tudjuk társítani egymással, és nem elég mindkettőre egyszerűen azt mondani, hogy „egyéb szerzőtől” származik. Az irodalomban ez a konkrét feladat már sokkal kevésbé jelenik meg.

A kísérleteink során mindkét verzióban születtek a gyakorlatban is felhasználható eredményeink. Zárt halmazon a legjobb eredmény 0.2-es fedésű és 0.95-ös pontosságú, mely bár korábbi zárt halmazos eredményekhez képest rosszabb, a kutatás során a korábbiaknál eltérő jellegű szövegekkel történt a munka: rövid fenyegetőleveleken. Az eredmények szignifikánsan rövidebb szövegeken születtek, mint korábbi kutatásokban, illetve a szerzők ezeknél az írásoknál aktívan igyekeztek elfedni az identitásukat. A bővített nyílt halmazos verzióban a legjobb modellre bár igenleges kimenet esetén nem lehet hagyatkozni, nemleges válasz esetén 78%-ban tényleg különböznek a szerzők, és 22%-ban téved, 964 db fenyegetőlevélen. Ezen a verzión korábban kevésbé voltak publikálva eredmények.

1. Bevezetés

„Mondd, mikor ér haza a férjed? Ha hazaért, közöld vele, hogy FIZESSE MEG, AMIVEL TARTOZIK... máskülönben...”

Mitévő legyen az ember, ha egy ismeretlen tettes megfenyegeti, akár rendszeresen ehhez hasonló üzeneteket kap? Látszólag a tettet megvédi az anonimitás, az áldozat tehetetlen... vagy mégsem? Lehetséges lenne valahogy azonosítani a szerzőt és igazságot szolgáltatni?

A 2013 tavaszán publikált *The Cuckoo's Calling* című művet Robert Galbraith írta, egy kevésbé ismert szerző, akiről nem sokat tudni... legalábbis ezt gondolták az emberek, amíg ugyanezen év júniusában egy anonim személy fel nem fedte, hogy a rejtélyes név mögött valójában a mára már mindenki által ismert, J.K. Rowling állt. [\[1\]](#) Ezt később a neves író meg is erősítette.

A szerző azonosításához kulcsszerepet játszhat azt megállapítani, hogy két szöveget ugyanaz az ember írta-e, ez a szerző-attribúció problémája és cikkünk témája. A problémának két ismertebb, és egy kevésbé ismert verziója szerepel az irodalomban: a legkönnyebb a zárt halmazos, melyben szerzők egy zárt halmazából kell kiválasztanunk azt a személyt, aki az adott szöveget írta. Ennél egy szinttel nehezebb a nyílt halmazos változat, ahol azt is meg kell állapítanunk, ha a korábban említett zárt halmazból senki sem társítható az adott szöveghez. A kutatásunk során mi a nyílt halmaznak egy általánosabb változatát vesszük szemügyre, ahol nincs előre rögzítve szerzők listája, csupán egy szövegpáros a bemenet, amiről meg kell mondani, ugyanaz-e a szerző. Azaz, ha két új szöveg mögött olyan szerző áll, akitől még nem láttunk korábban szöveget, akkor is képesnek kell lennünk megmondani, hogy a két szöveg szerzője azonos-e.

A probléma felmerülhet történelmi kontextusban is, erre egy példa volt a *Hitler Diaries*, amiket állítólag Hitler maga írt, viszont végül kiderült, hogy valójában hamisítványok voltak Konrad Kujau tollából [\[2\]](#). Ezért négy és fél évnyi börtönbüntetést is kapott (bár három év után kiengedték) [\[3\]](#). A szerző identitása ebből látszódnóan komoly következményekkel is járhat, akár bűncselekményt is követhet el az író, amilyen esetben pedig szükséges a pontos azonosítás.

A hamisítók elfogásánál természetesen súlyosabb motivációja is lehet a szerző azonosításának, például fenyegetőlevelek esetében kritikus fontossággal bír az írónak nem csupán a pontos, de számtalan esetben a gyors azonosítása is. A rendőrségnek, vagy akár terrorelhárító szervezeteknek igyekezniük kell a fenyegetést mihamarabb elhárítani, ugyanakkor rendkívül fontos, hogy a megfelelő embert kapják el, ezért súlyos felelősséggel járhat azt mondani, hogy „ezt a szöveget ez az ember írta”. Ezen okból kifolyólag kutatásom során különös figyelmet kapott az, hogy az adott módszer mennyire pontos, mennyire megbízható. Személy szerint

preferáltam azt, ha inkább abban jeleskedik a módszer, hogy „ártatlanságot” (azt, hogy nem egy adott illető írta a szöveget) megmutassa, mintsem abban, hogy megmutassa, hogy két szöveget tényleg ugyanaz az ember írt. Ezzel a megközelítéssel, ha a módszer alapján arra a válaszra érkezünk, hogy a két szöveg szerzője azonos, akkor azt figyelmen kívül hagyhatjuk, viszont ha arra érkezünk, hogy eltérőek, akkor azt figyelembe vehetjük és abban meg is bízhatunk.

Az ideális eset természetesen az lenne, ha mindkét fajta válaszban nagyon megbízhatnánk, azonban a feladat nehézsége miatt már a „félíg megbízható” modellek, módszerek is fontos mérföldkövek, és jó eredmények. A szerző azonosítás téma még napjainkig is aktív kutatási téma, és lényegében megoldatlan kérdés, annak ellenére is, hogy mekkora igény van a megfejtésére. Zárt, viszonylag szűk halmazokon már régebbi (2000-2010 közötti) módszerekkel is születtek már kiváló eredmények (különbféle nyelveken, nem csupán angol nyelvű szövegeken) a múltban [4], modern módszerekkel pedig lenyűgöző eredményekre is volt már példa angol korpuszokon (Reuters_50_50 és Gutenberg korpuszok, utóbbin 89,2%-os pontosság) [5].

Modern módszerekkel nyílt halmazon is születtek már jó eredmények [6], 2018-ban még 95%-os pontosságot is sikerült elérni [7] nagyszámú szerzővel és szöveggel (3000 novella és 500 szerző). Ezek rendkívül ígéretes eredmények, azt mutatják, hogy a téma kutatása eredményes és progresszív, azonban sajnálatos módon vannak hátulütők. Az első – és vitathatóan a legszembetűnőbb – probléma, hogy egyik eredmény sem olyan nyílt halmazon született, melyen új szövegpárosokat is tudni kell egymással társítani (innen: teljes nyílt halmaz), hanem a klasszikus nyílt halmazos problémán. A feladat csak addig terjedt ki, hogy „állapítsuk meg, hogy adott csoportból melyik szerző társítható az új szöveggel, vagy egyáltalán társítható-e bármelyikük is”, és ezekben a verziókban nem kellett megállapítani két vagy több új szöveg esetén, hogy az új szövegek szerzői egyeznek-e (még ha az eredeti szerzőhalmazban nem is voltak benne). A második fő probléma, hogy ezek a kutatások kifejezetten hosszú szövegekre fókuszáltak, több ezer szavasokra, és a felhasznált módszerek kevésbé – vagy egyáltalán nem is – működnek olyan rövid terjedelmű szövegek esetén, mint mondjuk a korábban említett fenyegetőlevelek. A harmadik fő probléma pedig, hogy a szóban forgó módszerek mennyire alkalmazhatóak az angoltól különböző nyelveken, például a kutatásomban vizsgált nyelven, a magyaron. Ezeken kívül sok egyéb probléma is felmerül, melyeket a következő szakaszban fejtek ki bővebben.

Kutatásom célja egy áttekintő képet adni a témáról, hogy egészen pontosan hol is tart most a világ a szerző azonosítás kérdésével, mire tudjuk már felhasználni, mennyire hatékonyan,

milyen megszorításokkal. Ötleket nyújtok arról, milyen irányokba lehet érdemes haladni a jövőben, és hogy ezek az irányok mi mindent implikálnak. Ezen felül kezdeti kísérleteket, eredményeket is bemutatok a teljes nyílt halmazos szerzőazonosítást illetően, magyar nyelven. Munkámat főképp a korábbi, főképp angol nyelvű kutatásokra alapozom, ezeket vizsgálom meg mélyebben, magyar nyelven.

2. A szerző-attribúció akadályai

A cikk 2., 3. és 4. fejezetei több ponton hagyatkoznak Patrick Juola 2008-as cikkjére [\[4\]](#), mely leírta a szerző-attribúció helyzetét, illetve összefoglalta egy 2004-es szerző-attribúció vetélkedő (*Ad-hoc Authorship Attribution Competition*, innentől csak vetélkedő) eredményeit is, melynek módszerei a kutatás során tesztelve lettek magyar nyelven. Erről bővebben később esik szó, a 3-4. fejezetekben.

A korábban említett problémák mellett egy komoly problémája a feladatnak annak nyitottsága, általánossága. Rengeteg fajta szöveg létezik a világon. Egyrészt vannak nagyon rövid szövegek, vannak hosszabbak, vannak egész cikknyi szövegek, de vannak egész könyvek is. Nem triviális feladat olyan módszert találni, amely rövidebb és hosszabb szövegekre is jól működik, különösen az előbbi komoly kihívást jelent, a korábbi kutatások is rendszerint hosszabb szövegekkel foglalkoztak, például novellákkal. Magas az esélye, hogy különböző módszereket lesz érdemes használni különböző hosszú szövegek esetén a jövőben is, nem elég csak egyetlen módszert ismerni, implementálni és alkalmazni. Extrém rövid szövegek esetén akár lehetetlen feladatba is ütközhetünk, általánosságban például a 4-5 szavas szövegek nem hordoznak kellő mennyiségű információt. Ez még a kisebbik probléma természetesen, ez várható akadály, azonban a leggyakoribb szövegek ennél nem sokkal hosszabbak, például e-mailek vagy chat üzenetek, és ezekre is komoly kihívás jól működő módszert találni.

A szövegek tematikája is nagyon sokoldalú, például a fentebb említett példákkal élve a *The Cuckoo's Calling* egy krimi novella, míg a Hitler Diaries naplószerű, ezek az írások nagyon nehezen hasonlíthatók össze intuitívan is. Ebből származik egy veszélyes csapda is, melybe könnyű beleesni, miszerint az írásokat egy módszer tematika alapján hasonlít össze inkább, és nem „szerző alapján”, ez inkább „tematikai azonosítás”. Az, hogy a „szerző alapján” pontosan mit is takar már önmagában is egy nehéz kérdés, ez a szerző-attribúció egy centrális kérdése is, hogy pontosan mi alapján is lehet elkülöníteni egymástól két szerzőt. Egy szerző sok különböző

helyzetben, sok különböző időben, sok különböző témában írhat szöveget, amik között közös pontot találni vitathatóan az egyik legnehezebb feladata a kutatási témának. Ebből külön kutatási irányokba is el lehet haladni, meg lehet például figyelni, hogy egy szerző stílusa hogyan változik az idők során.

A szöveg jellege szintén nagyon változatos, helyzettől függően az egyén írhat nagyon formálisan, vagy nagyon lazán is. Például az emberek többsége ha egy hivatalos levelet ír, akkor sokkal precízebben és megfontoltabban fogalmaz, mint akkor, mikor egy barátjával írogat egy chat alkalmazáson keresztül. A kutatásom során megtervezett szövegeken teszteltem (fenyegetőleveleken), azonban a tanító korpuszom lazább, kevésbé megtervezett szövegekből állt (kommentekből), ennek oka, hogy az utóbbihoz találtam megfelelő jellegű és méretű korpuszt, az előbbihez kis mennyiségű anyag állt rendelkezésemre. Korábbi kutatások főleg megtervezett, hivatalosabb jellegű szövegeken kísérleteztek, de előfordult lazább jellegű is. A gyakran használt Reuters korpusz [7] hírszövegekből áll, a Gutenberg korpusz [8] elektronikus könyvekből, míg 2008-ban a vetélkedő alatt nagyon változatos korpuszokat használtak: esszék, költemények, levelek, novellák is előfordultak, de kötetlen beszédekről is felhasználtak beszédleiratokat többek között. Felmerülhet a kérdés, hogy mennyire függ a különböző módszerek teljesítménye attól, hogy megtervezett vagy inkább lazább az írásnak a jellege. Ez a cikk ebbe nem megy bele mélyen, mert a kötött korpuszom egyrészt kisméretű volt, másrészt egy különleges eset.

Fenyegetőlevelek mögött álló szerzők azonosítása kutatásom fő célja. Ahogy említettem, ezek a szövegek különlegesek, ugyanis ezeknél az írásoknál a szerző nem csupán anonim, de lehet rá számítani, hogy szándékosan félrevezető is. Tekintve mennyire súlyos következményei lennének annak, ha azonosítanák, feltételezhető hogy egyes ravaszabb szerzők szándékosan máshogy írnak, mint egyébként szoktak, és kivételesen precízen, pontosan tervezik meg nem csupán azt, hogy mit írnak, de még azt is, hogyan. Ebben az esetben felmerül egy újabb akadály, amivel számolni kell, mégpedig az, hogy a szerző-attribúció során olyan szempontok alapján kell társítanunk a szerzőket az írásaikhoz, amiket nem triviális meghamisítaniuk. Olyan szempontok, amikre túl egyszerű odafigyelni, kevésbé működnek, sőt veszélyesek is lehetnek, kifejezetten félrevezethetnek. Absztraktabb, komplexebb jellemzők kellenek, amik elárulják a szerzőt akkor is, ha egyébként megpróbál félrevezetni. Jó hír az, hogy a korábbi kutatásokban a jellemzők többsége ilyen; a szerző-attribúcióhoz a tapasztalatok alapján kifejezetten bonyolult jellemzők kellenek egyébként is, bár van példa egyszerűbb jellemzőre is. Ennek súlyos

hátránya, hogy ez a komplexitás sokszor azzal jár, hogy az eredmények „miért”-je nehezen értelmezhető, a módszerből fekete doboz lesz, amit nehéz megfejteni, illetve elmagyarázni. Ez a kriminológiában különösen hátrányos, ugyanis ezen a területen kiemelten fontos, hogy megalapozott gyanú legyen a tettes identitásáról, ehhez pedig hiteles eredmény kell. Nem elég indok egy ember meggyanúsítására az, hogy „azt mondta a gép”.

3. Korábban próbált módszerek áttekintése

Amikor hagyományos, kézzel írt szövegekhez próbálunk egy arcot társítani, akkor ezt az íráskép alapján tesszük. Digitális szövegek esetén is hasonló módon szeretnénk eljárni. Ha azonosítani tudjuk, hogy pontosan miktől lesz egy szövegnek X a szerzője, és mitől Y, onnantól ezeket a jellemzőket megfelelően súlyozva meg tudjuk állapítani – vagy legalábbis magas valószínűséggel meg tudjuk mondani –, hogy a két szöveget ugyanaz a személy írta-e. A centrális kérdés múltbeli kutatások során – és a sajátom alatt is –, hogy pontosan mik ezek a jellemzők. Sok ötletelés történt már a múltban, hogy mi alapján lehet azonosítani. Egyik ilyen jellemző a szöveg szavainak átlagos hossza, erről azonban később kiderült, hogy egyrészt nem is konzisztensek egy szerző különböző írásai között, másrészt nem megfelelő különböző szerzők elkülönítésére sem. Egyéb jellemzők, amikkel próbálkoztak többek között a mondatok átlagos hossza, a szavak magánhangzóinak átlagos száma, szófajok eloszlása, type/token arány (lexikai változatosság), a Simpson Diversity Index illetve Yule karakterisztikus K-ja volt. Ezek közül azonban egyik jellemző sem bizonyult hatékonynak a feladat megoldásában. Történt próbálkozás az RVO (Relative Vocabulary Overlap) metrikával is, mely azt vizsgálja, hogy két írás szókincsileg mennyire hasonlít egymáshoz, mennyire merítenek ugyanazon szótárból. Ennek súlyos hátulütője, hogy a szöveg témája erősen befolyásolja. Ha két szöveg tematikailag egyezik – vagy nagyon hasonló –, akkor várhatóan jobban fog egyezni a szókincsük, mint két tematikailag eltérő szövegnek, még akkor is, ha az utóbbi párost egyébként ugyanaz írta, míg az előbbit két eltérő személy. Ez a korábban említett „tematikai azonosítás” csapdájának egy tipikus példája. Egy másik próbálkozás szinonimapárokkal történt. Emögött az intuíció, hogy bizonyos szavak helyett könnyen lehetne használni sok esetben egy másik szót, egy szinonimát, például „kutya” helyett „eb”, és ha egy adott szövegben az író konzisztensen egy adott változatát preferálja egy ilyen szónak, az jellemezheti őt. Ennek egy problémája, hogy sok esetben nem volt elég szinonimapáros ahhoz, hogy ezt a módszert megbízhatóan alkalmazzni lehessen. Erre bizonyos mértékig megoldás volt kifejezetten a funkció szavakra fókuszálni,

mert egyrészt sok különböző módon lehet velük fogalmazni, másrészt témafüggetlenek, harmadrészt pedig nagyon gyakoriak. Magyar nyelvben funkciószó alatt főleg névutókat, névelőket, névmásokat, igekötőket stb. értünk [9]. Ez a megközelítés szorosan kapcsolódik a vetélkedőn használt „stabil és instabil szavak” módszeréhez, melyet a kutatásom során szemügyre is vettem. Ezeken kívül próbálkoztak az úgynevezett Qsum technikával is, melynek lényege, hogy valamilyen metrikát (hagyományosan a két vagy három betűs szavak százalékos előfordulását) számosítanak több dokumentumon keresztül, majd felveszik őket egy vektorba. Ezután a vektorban minden értéket lecserélnek az átlagtól vett különbségükre, majd az így kapott vektorra kiszámolják a kumulatív összegeket. Például az $\{1, 2, 3, 4, 5\}$ vektorból először a $\{-2, -1, 0, 1, 2\}$ vektor jön, végül erre kumulatív összegeket számítva kijön az alábbi vektor: $\{-2, -3, -3, -2, 0\}$. Ez alapján megfigyelhetjük, mennyire „stabil” az adott jellemző a dokumentumokon keresztül, ami utalhat szerző azonosságra. Ezt a módszert is elvetették végül; bár először sikeresnek tűnt, végül kiderült, hogy nagymértékben pontatlan. Ennek a módszernek egy változata a WQsum, ahol az értékeket súlyozzuk is, erre több bizonyíték van, hogy működik, azonban a korábbi hatalmas bukás után sok a kétely körülötte. Az eddig említett módszerek mind egyszerű metrikákkal dolgoznak, de ahogy korábban is említettem, a megfelelő teljesítményhez ennél komplexebb megközelítésekre lehet szükség. Ezen metrikák kinyerésében kulcsszerepet játszanak a modern gépi tanuló eszközök, mint például a BERT. Ezek a módszerek bár biztatóbb eredményeket produkálnak, nehezen magyarázhatóak fekete doboz jellegük miatt, ezért nagy az igény egyszerűbben értelmezhető, de szintén hatékony módszerekre is. Kutatásom alatt ezt a megközelítést is megvizsgáltam, és valóban, még a „teljes nyílt halmazos” verzióban is használható eredmények születtek, és jelenleg ez az irány tűnik a legígéretesebbnek, azonban komplexitása miatt korlátozottak a felhasználási lehetőségei.

4. Saját kísérletek, tapasztalatok a korábbi módszerekkel

4.1 Egyszerű nyelvtani jellemzőkészlettel való tapasztalatok

Kezdeti kísérleti korpuszom 964 szövegből állt, ennek 10%-át használtam tesztalmaznak. A halmaz rendkívül szűk volt, fenyegetőlevelekre fókuszáltam, amiből az elérhető anyagom nagyon korlátozott volt (későbbi kísérletek során bevezettem egy sokkal nagyobb, hétköznapi szöveges korpuszt is, erről bővebben később). Ezek az ügyek titkosak, egy projekt kapcsán folyt velük munka, azonban az eredményeik felhasználásra kerültek a kutatás során. Kísérleteimet

egyszerűbb metrikákkal kezdtem meg, annak reményében, hogy magyar nyelven esetleg jobban működhetnek, azonban hamar rá kellett jönnöm, hogy sajnos ezen a nyelven sem könnyebb a feladat. 104 különböző nyelvtani és szemantikai jellemzőt használtam, többek között a negatív szavak arányát, pozitív szavak arányát, rasszista szavak arányát, főnevek arányát, igék arányát, folyamatos/befejezett/beálló melléknévi igenevek arányát, stb. Ezek a jellemzők egy projekt keretein belül készült rendszer segítségével lettek kinyerve, melynek alapjairól a [\[29\]](#) cikkben lehet olvasni bővebben. Minden jellemző egy arányszám volt, annak érdekében, hogy a dokumentumok hossza ne befolyásolja a végeredményt. Ez a fajta megközelítés eredménytelennek bizonyult, több különböző gépi tanuló modellel próbálkozva is teljes kudarcot vallott, egyedül a befejezett melléknévi igenevekbe tudott minimálisan kapaszkodni az egyik modell (döntési fa) mint jellemző, de abban az esetben is úgy láttam, inkább csak rátanulásról volt szó, a szűk halmazunk miatt lehetett csak értékelhető eredménye. Mélytanuló modellel is kísérleteztem. A paraméterek tapasztalatok, kísérletezések alapján lettek beállítva (és a kutatás során vizsgált későbbi modelleknél is). A konfigurációk az alábbi táblázatban láthatóak:

Tanulási Ráta:	Rejtett neuron-rétegek száma:	Neuron/háló:	Regularizáció:	Epoch szám:	F1 pontszám teszt-halmazon:
0.0001	4	184	Kiejtéses (0.1-gyel)	500	0
0.0001	2	104	Kiejtéses (0.1-gyel)	100	0.013303769401330375
0.0001	2	104	Kiejtéses (0.25-tel)	100	0
0.0001	2	104	Kiejtéses (0.25-tel)	20	0
0.001	2	104	Kiejtéses (0.25-tel)	100	0
0.0001	2	184	Kiejtéses (0.25-tel)	100	0
0.0001	2	104	semmi	100	0
0.0001	2	104	Kiejtés (0.1), l1 (0.01)	100	0
0.001	2	104	Kiejtés (0.1), l1 (0.01)	100	0
0.0001	2	104	Kiejtés (0.1), l2 (0.01)	100	0
0.0001	2	104	l2 (0.01)	500	0

A rejtett réteg aktivációs függvénye minden esetben relu volt, míg a kimeneti réteg aktivációs függvénye sigmoid. Adam optimalizáló volt használva, 128-as batch mérettel, keresztentrópia veszteségfüggvénnyel.

Egyik sem vezetett eredményhez, 0%-os eredményeket (F1-pontszám) kaptam minden esetben. A konklúzió, hogy ez a jellemzőhalmaz – önmagában legalábbis – nem tűnik megfelelőnek.

4.2 Stabilitás alapú jellemzőhalmaz bemutatása

Egy másik próbálkozásom a korábban említett stabil-instabil szavakkal, kifejezésekkel történt. A vetélkedőn ez volt a legjobb módszer, ezért ez tűnt a legígéretesebbnek. Moshe Koppel és Jonathan Schler alkalmazták ezt a módszert, és bizonyos korpuszokon kiváló eredményeket értek el vele. Az intuíció a módszer mögött, hogy vannak a nyelvben úgynevezett „instabil” szavak, kifejezések. Attól instabil egy kifejezés (mely lehet egy szó, egy név, vagy akár egy nagyobb mondatrész is), ha az könnyen helyettesíthető valami más, ugyanolyan jelentésű kifejezéssel. [10] Például „A barátom neve Ákos.” könnyen fogalmazható át arra, hogy „A barátomat Ákosnak hívják.”, így ez a mondat instabil. Egy probléma ezzel a jellemzőhalmazzal, hogy könnyen lehet, hogy nincs benne kellő mennyiségű instabil kifejezés, ezért másodlagos jellemzőhalmaznak használtak funkciósavakat, a tanulóhalmazból az 500 leggyakoribb szót, és POS (part-of-speech, szófaj) tageket is. A jellemzőket a vetélkedőn egy lineáris kernelt alkalmazó SVM-nek (Support Vector Machine) adták át. Fontos kiemelni, hogy a vetélkedő 2008-as volt, akkoriban még nem voltak elterjedve a mélytanuló neuronhálós modellek, napjainkra már lehetséges, hogy jobb eredményeket is el tudnánk érni, mint egy SVM-mel. Azonban a kísérletem során hamar kiderült, hogy bár intuitívan remek jellemzőhalmazról beszélünk, a gyakorlatban kevésbé alkalmazható. Meg kell keresnünk a szövegben, hogy pontosan melyik kifejezések azok, amik instabilak benne, ez azonban – főleg automatizálva – egy nemtriviális feladat. [10]-ben is találkoztak ezzel a problémával, és arra a megoldásra jutottak, hogy a szövegből alternatív verziókat kell generálni, melyek bár jelentésileg egyeznek az eredeti dokumentummal, a fogalmazás, szóhasználat mégis eltér. Ezen alternatív verziók elkészítéséhez gépi fordítókat alkalmaztak, olyan módon, hogy lefordították a szöveget az eredeti nyelvről (angol) különféle más nyelvekre, majd azokról visszafordítottak. [10] Ennek a módszernek egy nyilvánvaló hátulütője, hogy hatékonysága erősen függ a gépi fordítónak a teljesítményétől. A kísérletem során a Google fordítójával próbálkoztam, fizetős fordítók nem álltak rendelkezésemre. A fordító azonban nem teljesített elfogadhatóan jól, az

oda-vissza fordítás után (például magyar->angol->magyar fordítások után) a szöveg jelentésileg is megváltozott sok helyen, túl nagy mértékben. Ezt az útvonalat végül zsákutcának ítéltém, legjobb esetben is csak nagyon szűk, zárt halmazok esetén lehet működőképes, melyek a vetélkedőn is szerepeltek. A szerző-attribúció feladatában jellemzően kiemelt fontosságot élvez a pontosság – különösen a kriminológiai felhasználásában, ami a kutatás egyik fő motivációja is –, a gépi fordítók pedig jelenleg még nem elég megbízhatóak ahhoz, hogy átfogalmazhassunk velük szövegeket automatikusan. Ez akkor különösen igaz, ha figyelembe vesszük, hogy a módszer maga sem tökéletesen pontosan, még akkor sem, ha a jellemzők már rendelkezésünkre állnak; bár a vetélkedőn leveleken 10/10-es pontosságot értek el, hallgatók által írt esszéken már kevésbé voltak pontosak (részletes eredmények [\[4\]](#)-ben).

Ennek ellenére ez a módszer még így is ígéretesnek bizonyulhat a jövőben, csupán meg kell alapozni azt. Ha lenne egy megbízható „átfogalmazó” modellünk (akár magyar nyelvre), vagy ha már alapból rendelkezésre állnának alternatív megfogalmazások egy-egy dokumentumról (például szakértők által, kézzel írt változatok), akkor érdemes lehet megpróbálkozni ezzel a jellemzőhalmazzal ismételten.

4.3 N-gram alapú jellemzőhalmaz bemutatása

A vetélkedőn a második legjobb módszernek a CNG (Common N-Grams) bizonyult. Ez a megközelítés egy bájt n-gram jellemzőkészletet alkalmazó módszer, mellyel Vlado Keselj és Nick Cercone álltak elő. A módszert cikkükben részletesebben is bemutatták. [\[11\]](#) Ennek a módszernek a magyarázatához először azt fontos tudni, hogy pontosan az n-gram kifejezés mit jelent. Az n-gram kifejezésben az „n” egy számot takar. A lényege, hogy az egymást követő „n” elemű egységet vesszük jellemzőnek. Például ha adott az alábbi mondat: „Az alma szép és piros.”, akkor amennyiben szó trigramokat (3-gram) veszünk, az alábbi jellemzőket kapjuk: „Az alma szép”, „alma szép és”, „szép és piros”.

Az n-gramok nem csak szavakból állhatnak, hanem akár karakterekből is, például az előző példánál az alábbi karakter trigramokat kapnánk: „Az_”, „alm”, „lma”... stb (a _ a szóközt jelöli). A CNG módszer során lényegében karakter grammokkal dolgozunk, azonban előtte a szöveget bájtokkal reprezentáljuk. Egy példa erre, UTF-8-as kódolással: „Aladár elment hét órákor örölt lisztet és búzát venni.” -> „Alad\xc3\xa1r elment h\xc3\xa9t \xc3\xb3r\xc3\xa1kor \xc5\x91r\xc3\xb6lt lisztet \xc3\xa9s b\xc3\xbaz\xc3\xa1t venni.”

Az angol ábécé karakterei megmaradnak eredeti formájukban, azonban az ékezetes karakterek átalakulnak. Ehhez a transzformációhoz a Python nyelv bytes metódusa került felhasználásra, majd a kimenet vissza lett konvertálva olvasható, szöveges formába (str típusra). [12] Intuitívan ez a reprezentáció a magyar nyelv esetén azért működhet akár jól, mert abból is információ lesz, hogy ha a szerző egyáltalán ékezetes karaktert használ. Például az „ó” betűből a „\xc3\xb3” karaktersorozat lett, az „é” karakterből pedig „\xc3\xa9”. Mindkettő úgy kezdődik, hogy „\xc3”, és jól láthatóan hasonló jellegűek ezek a karakterláncok.

A módszernek egy súlyos hátránya, hogy a jellemzőkészlet az embernek nagyon keveset mond, nehezen érthető, és ennek következtében nem is tűnik hitelesnek. Rámondani két szövegre, hogy ugyanaz írta, mert mindkettő fél szövegében gyakran fordultak elő bizonyos, látszólag teljesen véletlenszerű karakterhármassok (mondjuk az „alma” szóból az „lma”), nem kifejezetten hihető, és különösen kriminológiában, ahol a félreazonosítás súlyos következményekkel járhat, ez egy kritikus hiányosság. Ha egy mélytanuló hálónak adnánk be ezeket a jellemzőket, akkor már végképp követhetetlenné válik, hogy a modell szerint pontosan miért is azonos vagy nem azonos két szövegnek az írója. A vetélkedőn a gramoknak a normalizált gyakoriságát vették jellemzőnek, majd az ezekből készült vektorokat egy k-legközelebbi szomszéd osztályozónak adták át, egy-egy szöveget az abból készült gyakoriságvektor reprezentált. A módszernek egy másik szembevetendő hátránya, hogy csak zárt halmazra működik jól; új szövegek esetén le vagyunk korlátozva azokra a jellemzőkre, azokra a gramokra, amiket már láttunk korábbi szövegeknél, ami túl kevés, különösen, ha a szövegeink eleve rövidek is, és ennek következtében kevés különböző grammal találkozunk. Oda kell figyelni arra is, hogy hány különböző n-grammot engedünk meg, könnyen felrobbanhat a jellemzőtér magasabb gram számok esetén; például hány különböző 7-8 hosszú karakter(bájt)gram fordulhat elő egy nagyobb korpuszban? Legrosszabb esetben milliós nagyságrendeket is elérhetünk. A kísérleteim során túl magas gram számok esetén már pusztán az n-gram gyakoriságvektorok előállítás sem történt sikeresen technikai korlátok miatt, a számítógép kifutott a memóriából. Ez különösebb akadályt erősen feltételezhetően nem jelent azonban, korábbi kutatások során is rendszerint csak a leggyakrabban előforduló nyelvi elemek azok, amik igazán súlyosak, amik túl ritkák azok rendszerint figyelmen kívül hagyhatók (például a vetélkedőn a legjobb módszerben a másodlagos jellemzőhalmazhoz is csak az 500 leggyakoribb szóval dolgoztak). Ennek fényében legfeljebb az 5000 leggyakoribb gram volt figyelembe véve a kísérletek többsége során, pár kivétellel.

Az n-gram gyakoriságvektorok előállításához az ingyenes Stylo csomagot használtam, mely az R nyelvhez elérhető. [13] A gyakoriságvektorok z-pontszám szerint kerültek normalizálásra, és az alábbi képet alapján számolható ki: $Z = (x - \mu) / \sigma$

A képletben az x jelöli az adott gram gyakoriságát (előfordulásainak számát), a μ jelöli az adott gramnak az átlagos gyakoriságát a korpuszban, a σ betű pedig a szórást jelöli a populációra (korpuszra) vetítve.

Az így kapott gyakoriságvektorokat ezek után vagy egymáshoz fűztem (például az (1, 2, 3) és (4, 5, 6) vektorokból az (1, 2, 3, 4, 5, 6) vektor lett), vagy pedig a két vektor abszolút különbségét vettem, elemenként (például az imént említett vektorokból az eredmény vektor (3, 3, 3) lenne). Az előbbi megközelítésről korán kiderült, hogy eredménytelen, ezért főleg az utóbbival dolgoztam. Nem csupán bájgramokkal teszteltem, hanem megvizsgáltam az eredményeket szó és egyszerű karaktergrammokra is, különböző n-ekre. Alább az így kapott eredményeket fogom részletezni.

Gépi tanuló modellnek az SGDClassifier lineáris osztályozó került választásra az sklearn Python-csomagból, alapértelmezett paraméterezéssel [14]. A halmazunk eredetileg 542 dokumentumból állt, amiből 435 volt tanulóhalmazban, 107 a tesztben. Később 595 dokumentumunk lett, és a teszthalmaz bővült 160 ügyre. 54 különböző szerző volt a tanulóhalmazban, 29 a tesztben. A bővülés után 32 különböző szerző volt a tesztben. A kutatásunk ezen pontján még csak nagyon kevés adat állt rendelkezésre. A későbbi, BERT-es kísérletek során már sokkal nagyobb halmazzal történtek a kísérletek, arról a halmazról később lesz szó ejtve.

4.3.1 A módszer kiértékelése

A kiértékeléshez az F1-pontszám metrika volt alkalmazva, ez volt a fő szempont, bár ki lett számolva sok esetben az osztályozási pontosság (angol irodalomban accuracy) is. Az osztályozási pontosság egyszerűen csak a helyes címkék százalékos aránya a címkék számához képest. Az F1-pontszám két másik pontszámból számolódik, a pontosság (angol irodalomban precision) és a fedés (angol irodalomban recall) értékekből. A **pontosság** (precision) azt írja le, hogy a modell mennyire megbízható. Pontosabban, hogy milyen arányban volt valóban azonos két szöveg szerzője azokból az alkalmakból, amikor a modell az azonos címkével látta el őket. Ha értéke magas, akkor ha a modell szerint két szöveget ugyanaz írt, akkor nagy

valószínűséggel tényleg ugyanaz írta. Értékét úgy kaphatjuk meg, hogy a helyes „igen” (azonos szerzőt mondtunk, és azonos is volt, angol irodalomban True Positive) címkék számát elosztjuk a helyes „igen” címkék és a helytelen „igen” (azonos szerzőt mondtunk, de különbözőek voltak, angol irodalomban False Positive) címkék számainak összegével. A **fedés** (recall) ezzel szemben a modell „hasznosságát” írja le, azt, hogy mekkora arányban találta el a modell az azonos szerzős címkéket. Ha értéke magas, akkor ha két szöveget ugyanaz írt, arra nagy valószínűséggel a modell rájön. Értékét majdnem ugyanúgy számoljuk, mint a pontosság értékét, annyi különbséggel, hogy az osztóban a helyes „igen” címkék számához nem a helytelen „igen” címkék számát adjuk hozzá, hanem a helytelen „nem” (különböző szerzőt mondtunk, de valójában azonosak voltak, angol irodalomban False Negative) címkék számát. Az F1-pontszám ennek a két metrikának a harmonikus átlaga, azaz úgy számolandó, hogy a pontosság és a fedés szorzatát megduplázzuk, majd az eredményt elosztjuk a pontosság és a fedés összegével.

A metrikákhoz összefoglalva a képletek a mellékelt ábrán láthatóak [\[15\]](#).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Az adathalmaz nagymértékben egyenletlen volt, jellegéből fakadóan: sokkal több olyan vektor volt, amihez a „nem azonos szerző” címke tartozott, mint olyan, amihez az „azonos szerző” címke társult. Ez várható volt, annak a következménye, hogy „mindent mindennel” vetettünk össze. Ezért látszott az F1-pontszám egy jobb metrikának, mert az egyszerű osztályozási pontosság metrika esetén elég lenne egyszerűen csak minden szövegpárosra azt mondani, hogy különbözőek a szerzők, és akkor is látszólag nagyon jó eredmények születnének, mert a párosok túlnyomó többsége valóban különböző szerzőtől van. Az F1-pontszám esetén egy ilyen stratégia nem működne jól, mert a fedés értéke nagyon alacsony lenne. Természetesen ha mindenre azt mondanánk, hogy azonos a szerző, akkor pedig a pontosság értéke lenne túl alacsony. Meg kell találni a helyes egyensúlyt, „óvatos” modell kell, ami nem mondja minden szövegpárosra, hogy azonos a szerző, de ha túl ritkán találja meg emiatt az azonos szerzős szövegpárokat, akkor a modell gyakorlati értéke drasztikusan romlik.

Ahol nem említi a kutatásom külön, hogy hány szógram/karaktergram/bájtgram lett felhasználva, ott alapértelmezetten 4901 volt. A szövegek kisbetűsítve voltak majdnem mindig (a kivételek említve lesznek), a stopszavak ki lettek szűrve. Más egyéb előfeldolgozó lépés nem történt.

4.3.2 Kísérletek szógramokkal

Kezdetben szógrammokkal történtek a kísérletek, bigramokkal (2-gram) és trigramokkal (3-gram), azonban ezekkel nagyon rossz eredmények születtek. A fentebb említett abszolút értékben vett különbségvektorokkal dolgoztam. Lehetséges, hogy nagyobb adathalmaz esetén az eredmények javulnának, az adathalmazban különösen az azonos szerzős címkékből nagyon kevés példány szerepelt.

A korai kísérletek során az adathalmazba minden szövegpáros kétszer került be, mert bekerült „A összevetve B-vel” és „B összevetve A-val” példányokban is, azonban mivel a vektorok abszolút különbségének számítása szimmetrikus, ezért ezek a vektorok minden esetben egyeztek. Későbbi kísérletek során a duplikált vektorok eltávolításra kerültek. A szógramok azonban továbbra is 0-s eredményt mutattak az adatkiegészítések nélkül.

4.3.3 Kísérletek karaktergramokkal

A karaktergramok ezzel szemben sokkal ígéretesebb eredményeket mutattak még a kis adathalmaz ellenére is. A kezdeti kísérletek során a vektorok a korábban említett módon egymáshoz lettek fűzve, azonban ilyen módon nem születtek érdemleges eredmények, az F1-pontszám ebben az esetben is nulla volt. Azonban az abszolútértékben vett különbséges vektorok már sokkal jobban teljesítettek. Trigramokkal történtek először próbálkozások, ekkor a legjobb eredmény 0.39-es F1-pontszámú volt, 0.28-as fedés értékkel és 0.62-es pontosság értékkel. Történtek kísérletek a grammok számának korlátozásával. Az alap 4901-es értékről 100-ra és 2000-re is le lettek korlátozva, viszont mindkettő rosszabb eredményt adott, mint az alapérték, az előbbi különösen, annyi grammal nem tudott a modell hatékonyan tanulni, és az F1-pontszám 0 lett. A korábban említett szimmetrikus összevetésből következő duplikált vektor probléma megoldásával („A összevetve B-vel” és „B összevetve A”-val ugyanaz, ezért egy rekordnak számít) az eredmények marginálisan javultak, 0.4-es F1-pontszámra, 0.3-as fedés és 0.6-os pontosság értékekkel. Ez a különbség azonban feltehetően csak a gépi tanuló modell véletlenszerűségéből fakad (a jellemzők kezdetleges

súlyai véletlenszerűek). 4-karaktergramot felhasználva az eredmények nagyon hasonlóak voltak, azonban 5-gram esetén a pontosság nagyon magas lett, de a fedés kárára: az F1-pontszám ebben az esetben 0.31 lett, 0.92-es pontosság és 0.19-es fedés értékkel. A modell túl kevés szöveg esetén volt képes észlelni a szerzők azonosságát, de ahol észlelte, ott rendkívül megbízhatóan. Történt kísérlet 6-grammal is, azonban ez már túl soknak bizonyult, bár a pontosság értéke 1 lett, a fedés értéke közel 0, így az F1-pontszám is közel 0 lett: a modell majdnem egyáltalán nem talált szerzőegyezt.

4.3.4 Kísérletek bájtgramokkal

Ezek után a korábban említett „bájtgramokkal” folytak kísérletek, ezúttal minden esetben a vektorok abszolút különbsége (elemenként) lett véve, és nem voltak duplikált vektorok. Kezdetben trigramokkal folytak kísérletek, a felhasznált gramok száma volt változó, 4001 és 5000 között mozgott. A legrosszabb eredmény 4001 gram esetén született, abban az esetben 0 volt az F1-pontszám, míg a legjobb eredmény 4901 gram esetén született, amelyik esetben 0.27-es volt az F1-pontszám, a pontosság 0.99 volt, a fedés eredménye pedig 0.16 lett. A fedés eredménye konzisztensen alacsony volt az összes bájtgramos kísérletnél, a pontosság viszont kifejezetten magas. Ebben az irányban lehet valamennyi potenciál, mert bár ritkán találja meg a modell a szerző egyezést, amikor megtalálja, az majdnem garantáltan tényleg egyezik. Tekintve, hogy 4001 grammal nullás eredmény született, feltételezhető hogy nagyon magas dimenziószámú vektorok kellenek ezzel a módszerrel, hogy érdemleges eredmények születhessenek. Sokkal jobb eredmények születtek azonban 4 gram esetén: ebben az esetben 0.34-es F1-pontszám volt a legjobb, 0.2-es fedés értékkel és 0.95-ös pontosság értékkel, azaz körülbelül minden ötödik egyező szerzőjű szövegpárosról sikeresen megállapította, hogy a szerző azonos, 95%-os sikerrel. Ebben az esetben semmilyen korlát nem volt adva a gramok számára, ami szintén hozzájárulhatott a jó eredményhez. Azonban a nagyon magas dimenziószám sokkal gépigényesebb adatfeldolgozást és gépi tanulást is jelent, ami problémát jelenthet, főleg az adatok számának növekedésével. Dimenziócsökkentő módszereket érdemes lehet kutatni, tesztelni ezzel a megközelítéssel. Történt még egy teszt ezt látva trigram esetén is, ezúttal a gramok számának korlátozása nélkül, viszont ebben az esetben nem javultak az eredmények, az F1-pontszám csak 0.22 lett, a fedés 0.12, a pontosság pedig 0.98.

Az adathalmazunk később bővült 53 tesztadattal (a tanulóhalmaz változatlan maradt), és történt még 2 teszt ezzel a halmazzal is, mindkét esetben 4 grammal. Technikai korlátozások miatt

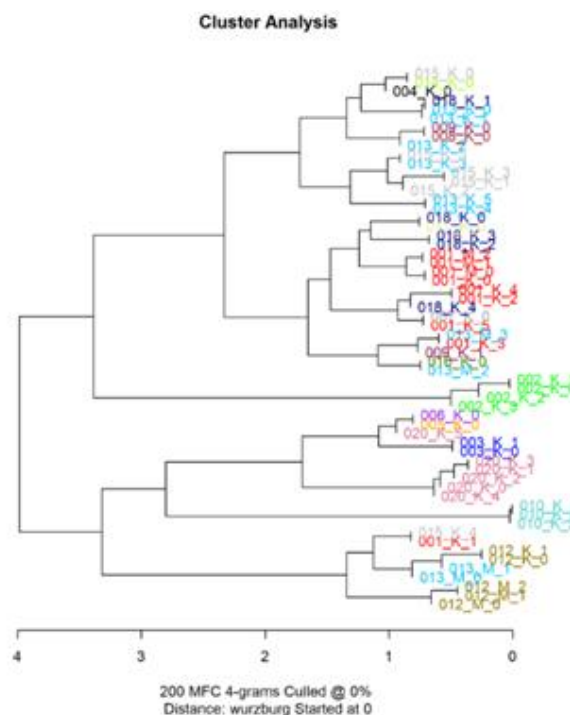
csak 35 000 illetve 40 000 grammal volt lehetőség kísérletezni, és ezekkel az eredmények romlottak, az utóbbi esetben 0.24 volt az F1-pontszám, 0.15 a fedés és 0.62 a pontosság. Ezeknél a kísérleteknél azonban a szöveg nem lett kisbetűsítve sem. Tekintve, hogy az eredmények már 35 000 gramról 40 000-re váltva is javultak, lehetséges hogy a dimenziószám korlátozása nélkül hasonló eredmények jöhetnének ki mint a kisebb halmaz esetén, azonban ezen a ponton még kritikusabbá válnak a hatékony dimenziócsökkentő módszerek.

Összességében a vetélkedőn is említett „bájtgramok” módszere ígéretesnek bizonyult, viszont erősen korlátozott a felhasználása a zárt halmaz miatt, illetve magas számítása miatt is. Érdemes lehet tovább kutatni ezt a megközelítést dimenziócsökkentő módszerekkel és nagyobb mennyiségű adattal, esetleg más modellekkel.

4.3.5 Kísérletek a Stylo csomaggal

A szerző-attribúció feladatához készült R nyelven egy ingyenesen felhasználható csomag is, a Stylo, mellyel zárt halmazon lehet dolgozni. [\[16\]](#) Ezzel a csomaggal lettek előállítva az n-gramos jellemzőkészletek is. Azonban a csomag önmagában is felhasználható sokrétűen paramétereizhető kísérletekre, amiknek eredményeit ábrákkal is képes szemléltetni. A csomag számtalan nyelvet támogat, többek között a magyart is. Alapvetően az n-gram jellemzőkészlettel dolgozik, állítható hogy karakter vagy szógramokkal akarunk-e kísérletezni, kisbetűsíthető a korpusz, meg lehet adni, hogy csak az X leggyakrabban előforduló szót használjuk, illetve azt is, hogy legalább a dokumentumok hány százalékban kell előfordulnia az adott szónak, hogy figyelembe vegyünk. Ezen kívül figyelmen kívül lehet vele hagyni a leggyakrabban előforduló szavakat is, lehetőség van stopszó szűrésre is, és egyéb funkciókat is támogat. Az elemzés nem gépi tanulás alapján történik, hanem az n-gramokból számított z-pontszám vektorok között számít távolságokat, ami szintén megadható, hogy milyen metrika szerint. Tapasztalatok alapján a koszinusz delta távolság működik a legjobban.

Ez a cikk nem megy bele mélyen a csomag működésébe és funkcióiba, csupán egy betekintést nyújt, azonban kisebb, zárt halmazos feladatokhoz kiválóan működik a gyakorlatban is ez a csomag, bár magasabb dokumentumszám esetén a teljesítménye gyorsan romlik. Egy példa a program teljesítményére, 60 dokumentum esetén az alábbi ábra.



Az azonos szín azonos szerzőtől származó szöveget jelöl. Jól láthatóan a színek jól vannak csoportosítva, egy csoportba kerülnek az azonos szerzőtől származó szövegek. Ez a kísérlet karaktergramokkal történt, 4-gramokkal, kisbetűsített szövegekkel, koszinusz távolságot felhasználva, a 200 leggyakrabban előforduló szó figyelembe vételével. Egyéb szűrés a szavakra nem történt.

5. Mélytanuló illetve BERT modellekkel folytatott kísérletek

A kísérlet során modern módszerek is kipróbálásra kerültek, a feladat teljes nyílt halmazos verziójához ebbe az irányba haladt a kutatás. Kezdetben az úttörő BERT (Bidirectional Encoder Representations from Transformers) modell tűnt ígéretesnek a feladathoz. A BERT egy transformer alapú gépi tanuló modell, mely 2018-ban eredményeivel a legkorszerűbbnek bizonyult a természetesnyelv-feldolgozás (NLP) számos területén. Újítása, hogy korábbi nyelvi reprezentációs modellekkel ellentétben képes két irányból, balról és jobbról is értelmezni a szöveget, ezzel sokkal hatékonyabban reprezentálva a szöveg kontextusát. [17] A modell egy

masszív korpuszon van előtanítva, ezt a modellt pedig tovább lehet tanítani specifikus feladatokhoz, például esetünkben a szerző-attribúcióhoz, ehhez egy kimeneti rétegre van csupán szükség [17]. A BERT felhasználása többnyelvű, azonban készült kifejezetten a magyar nyelvhez is modell, huBERT névvel, melynek teljesítménye magyar nyelvű szövegeken felülmúlja az eredeti BERT által nyújtottat, ezért a kutatás során ez a modell lett felhasználva. [18] A modell a BERT-re alapoz, azonban más korpuszokon van előtanítva, a Webcorpus 2.0, illetve a magyar Wikipédia szövegein. [18]

A kutatás során használt mélytanuló modellt (beleértve a huBERT-et felhasználó modellt is) egy projekt keretei között egy kolléga írta meg, később ezt a modellt konfigurálva, paraméterezve folytattam a kísérleteket.

A modell alapja egy beágyazó réteg, mely a szövegben előforduló szavakból vektorokat készít, azok aktuális jelentése alapján, a szöveggörnyezetet figyelembe véve. Ezt követi egy statisztikai jellemzőkből készült réteg, mely egy másik jellemzőkészletből dolgozik. Annak érdekében, hogy a dimenziószámok kompatibilisek legyenek a beágyazás dimenziójával, a kettő között található egy dimenziócsökkentésre használt réteg is. A statisztikai réteget követi szöveges jellemzők rétege, melyben a beágyazás vektorok közötti távolságok lettek felhasználva. A vektorok távolságának a vektorok elemeinek abszolútértékben vett különbségeit értettük. A szöveges jellemzők rétegét követi egy újabb statisztikai réteg a dimenziószámok egyeztetéséhez, végül egy egyszerű, 1 dimenziós, sigmoid aktivációs kimeneti réteg megadja a szerzők egyezésének valószínűségét.

A statisztikai rétegben számos jellemző lett használva, melyek különböző módokon jellemzik a szövegeket. Ezek alább kerülnek részletezésre.

Az alábbi jellemzők kerültek felhasználásra:

- leghosszabb közös részsorozat (karakter- és szószinten is)
- szó- és karaktergramok egyezési aránya (részletesebben később)
- a szövegpáros átlag BLEU (Bilingual Evaluation Understudy) értéke [19]

Illetve az alábbi jellemzők abszolútértékben vett különbsége a szövegpáros között:

- ARI (Automated Readability Index) [\[20\]](#)
- Flesch Reading Ease pontszám [\[20\]](#)
- Flesch-Kincaid Grade Level, mely a Flesch Reading Ease pontszámon alapul [\[21\]](#)
- Coleman-Liau index (más néven Coleman-Liau Readability Formula) [\[22\]](#)
- Gunning Fog Index [\[23\]](#)
- Linsear Write Readability Formula [\[24\]](#)
- SMOG (Simple Measure of Gobbledygook) pontszám [\[25\]](#)

A jellemzők nagy vonalakban a szövegpárosok olvashatóságát, bonyolultságát hasonlítja össze, ez intuitívan segíthet azok összehasonlításánál. Többek között nyomra vezethet a szerző iskolázottságát illetően is. Ezen felül ismételten előkerültek az n-gramok, mint jellemzők, illetve a BLEU metrika is felmerült. Ez a metrika gépi tanuláshoz használatos a szöveg, a fordítás minőségének megállapításához. A pontszám számítása során a szöveget összevetjük egy professzionálisan fordított szöveggel, és megnézzük, mennyire áll hozzá közel. Jelen esetben a szöveget nem annak fordításával vetjük össze, hanem azzal a szöveggel, amellyel össze akarjuk hasonlítani, hogy ugyanaz írta-e. Ez két irányba is megtörténik: először úgy vesszük, hogy az egyik szöveg az eredeti szöveg, és a másik szöveg a „fordítás”, majd ezt megvizsgáljuk fordítva is, végül az eredmény a két pontszám átlaga lesz. Intuitívan ez a metrika is segíthet feladatunkban, hiszen ez is „szöveg közelséget mér”, még ha más megközelítéssel is. Kiszámításához az NLTK csomag `sentence_bleu` függvénye volt alkalmazva [\[26\]](#), 3-gramig lettek megadva a súlyok, melyek 0.5, 0.3 és 0.2 voltak, a különböző gram értékekre.

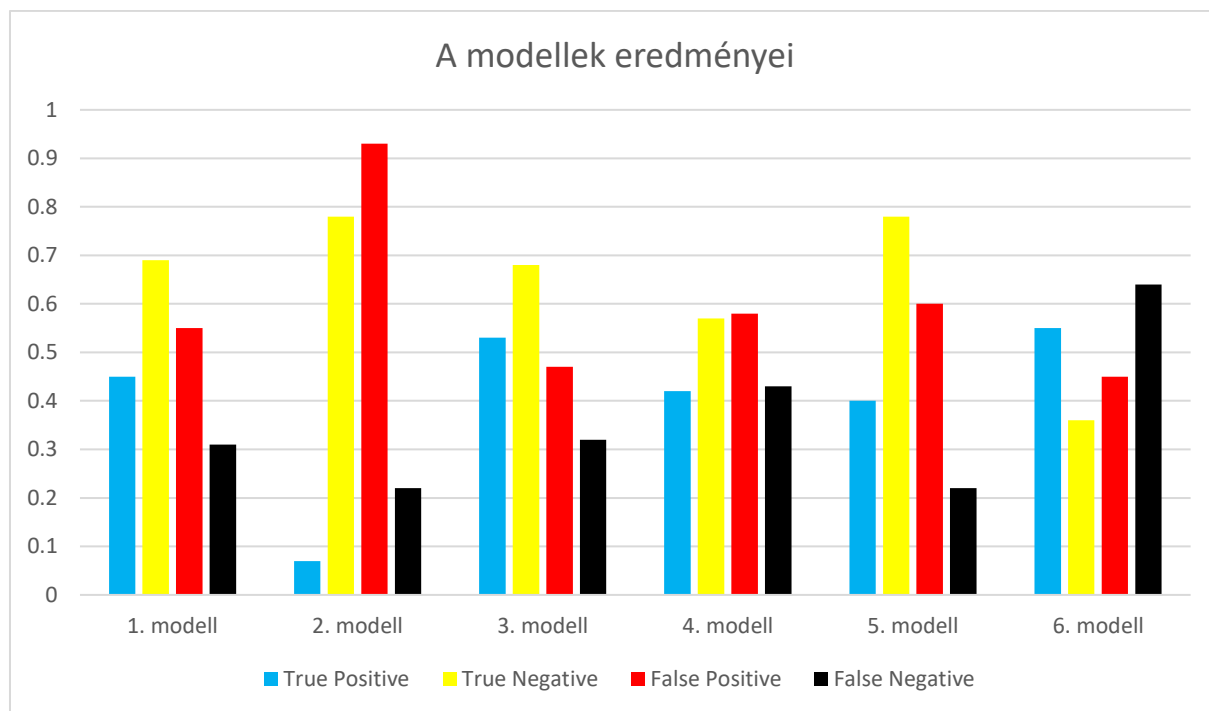
A korábban említett szógramok egyezési aránya esetén megnéztük, mekkora arányban egyeznek az 1-gramok, 2-gramok és 3-gramok a szövegpárosban, majd ezeket súlyoztuk 0.5, 0.3 és 0.2 értékekkel. Ugyanazt megtettük a karaktergramokkal, ott 4-gramig ellenőrizve, 0.3, 0.3, 0.3 és 0.1-es súlyokkal.

A modellnek egy hátránya, hogy nagyon sok ideig tart tanítani, a kutatás során szervergépen folytak a tanulások, és ennek ellenére is napokig tartott minden kísérlet. Ezen okból kifolyólag kevés kísérleti eredmény is született, az idő hiánya miatt, így érdemes lehet a jövőben kísérletezgetni a paraméterek további finomhangolásával is. Az adathalmazunk a kutatás ezen fázisára szignifikánsan megnőtt.

Tanítóhalmaznak 77018 szöveg lett használva, míg teszhalmaznak a teljes korábbi adathalmazunk, azaz 964 szöveg. Tanítóhalmaznak a Hunglish 1.0 korpusz egy részét használtuk, melyben magyar nyelvű, hétköznapi szövegek vannak az index.hu oldaláról. A korpusz elérhető a [\[27\]](#) oldalon található FTP linken keresztül [\[28\]](#). A Hunglish 2.0 korpusz, ami szintén megtalálható az oldalon, bár nagyobb, nem tartalmaz ilyen jellegű, hétköznapi szövegeket, ezért lett a 1.0-ás verzió használva. A pers fájlban található korpusz lett használva. A teszhalmaz a kísérletek többségében ki lett egyenlítve, olyan értelemben, hogy az azonos szerzős adatok száma egyezzen az eltérő szerzős adatokkal. Ez olyan módon történt, hogy az eltérő szerzős adatok mennyiségét korlátoztuk az azonos szerzős adatok számára. Ennek hátránya azonban, hogy kevésbé tükrözi az adatok reális, valóéletbeli eloszlását, ezért a legjobbnak ítélt próbálkozás tesztelve lett az eredeti, kiegyenlített teszhalmazon is.

Az eredmények az alábbi táblázatban illetve ábrán láthatóak:

Használva van BERT modell	Össze-fűzött állapotok számai:	TP	TN	FP	FN	Teszt-halmaz:
igen	-2, -1	0.45	0.69	0.55	0.31	Ügyek, kiegyenlítő.
igen	-12, -11, -10, -9, -8, -7, -6, -5, -4, -3, -2, -1	0.07	0.78	0.93	0.22	Ügyek, kiegyenlítés nélkül.
nem	-2, -1	0.53	0.68	0.47	0.32	Ügyek, kiegyenlítő.
nem	-4, -3, -2, -1	0.42	0.57	0.58	0.43	Ügyek, kiegyenlítő.
igen	-4, -3, -2, -1	0.4	0.78	0.6	0.22	Ügyek, kiegyenlítő.
nem	-3, -2, -1	0.55	0.36	0.45	0.64	Ügyek, kiegyenlítő.



A legjobb eredményt a 2. modell nyújtotta. Hátránya, hogy a nagyon magas False Positive (helytelenül azonos szerzőnek vélt) érték miatt a modell igenleges válaszai megbízhatatlanok,

azonban ha nemleges választ ad, az megbízható. A magas True Negative (különböző a szerző, és különbözőt is mondott a modell) értékből láthatóan az ilyen eseteket gyakran meg is találja a modell. Ennél a kísérletnél az előtanított BERT modell volt alkalmazva, ebből látszik, hogy a BERT ezen a feladaton is jól teljesít. Azonban a többi eredményből az is látszik, hogy bár általánosságban rosszabbul teljesítenek az egyszerű, nem előtanított beágyazási réteget alkalmazó modellek, a teljesítmény romlása nem drasztikus összességében. A False Negative (különböző szerzőt mondott a modell, pedig azonos volt) értékek nem BERT-es modellek esetén viszont konzisztensen magasabbak, ami például kriminológiai felhasználás esetén kritikus hiba.

A korábban említett legjobb modell a gyakorlatban is felhasználható eredményeket produkált, mely ígéretes a feladat jövőjét tekintve. További kísérletek szükségesek azonban, a paraméterek további finomhangolásával, több tanítóadattal, illetve a statisztikai jellemzők súlyának módosításával. A modellek konfigurációi a következő oldalakon találhatóak.

			A beágyazásban található Feed Forward Net dimenziója	Beágyazást követő dimenzió- csökkentő réteg dimenziója		
Beágyazás dimenziója	Beágyazás rétegeinek száma	A beágyazásban használt Attention Headek száma	FFD Dimenzió	Dimenzió- csökkentő réteg mérete (dimenziója)	Statisztikai jellemzők rejtett rétegének dimenziója	Statisztikai jellemzők rejtett rétegének aktivációja
768	12	12	3072	384	0	-
768	12	12	3072	384	384	gelu
768	12	12	3072	768	128	tanh
768	6	16	1536	384	384	gelu
768	6	16	1536	384	0	-
768	12	12	3072	384	384	gelu
768	8	16	1536	384	384	gelu

					A beágyazás megadott rejtett állapotait használjuk, nem csak az utolsót (-1 az utolsó állapot)					
Szöveges jellemzők köztes rétegének dimenziója	Szöveges jellemzők köztes rétegének aktivációja	Statisztikai jellemzők köztes rétegének dimenziója	Statisztikai jellemzők köztes rétegének aktivációja	Használva van BERT modell	Összefűzött állapotok számai:	TP	TN	FP	FN	Teszt-halmaz:
0	-	0	-	igen	-4 -3 -2 -1	0.57	0.61	0.43	0.39	Ügyek, kiegyenlítve.
64	gelu	32	gelu	igen	-2, -1	0.45	0.69	0.55	0.31	Ügyek, kiegyenlítve.
32	gelu	16	tanh	igen	-12, -11, -10, -9, -8, -7, -6, -5, -4, -3, -2, -1	0.07	0.78	0.93	0.22	Ügyek, kiegyenlítés nélkül.
64	gelu	32	gelu	nem	-2, -1	0.53	0.68	0.47	0.32	Ügyek, kiegyenlítve.
0	-	0	-	nem	-4, -3, -2, -1	0.42	0.57	0.58	0.43	Ügyek, kiegyenlítve.
128	gelu	64	gelu	igen	-4, -3, -2, -1	0.40	0.78	0.60	0.22	Ügyek, kiegyenlítve.
128	gelu	64	gelu	nem	-3, -2, -1	0.55	0.36	0.45	0.64	Ügyek, kiegyenlítve.

6. Összefoglalás, konklúzió

Az angol nyelven kipróbált módszerek az eredmények fényében jól teljesítenek magyar nyelven is, sok bennük a potenciál és ígéretesnek mutatkoznak jövőbeli kísérletezések velük. Zárt halmazon több módszer is tesztelve lett, különböző jellemzőkészletekkel. A jellemzőkészletbe kerültek szógramok, karaktergramok, bájtgramok, illetve egy elméleti módszer során felmerültek az instabil kifejezések, melyek olyan nyelvi elemek, amiket könnyen helyettesíthetünk a szövegben a jelentés megváltozása nélkül. Ezek közül a két legígéretesebbnek a karaktergramok, illetve a bájtgramok bizonyultak, melyek közül az utóbbihoz kapcsolódó módszer legjobb esetben 0.2-es fedést és 0.95-ös pontosságot adott, míg az előbbihez legjobb esetben 0.3-as fedést és 0.6-ot pontosságot. A stabil-instabil kifejezéseken alapuló módszer intuitívan hatékornak hangzik, de a gyakorlatban nehezen tesztelhető, alkalmazható, amennyiben viszont jövőbeli kutatások során a tudományos világ talál egy megfelelő módot dokumentumok alternatív verziójának elkészítéséhez, abban az esetben ezt a módszert érdemes lehet kipróbálni. Teljes nyílt halmazon a transzformációs modellek ígéretesnek bizonyulnak. Ezen a ponton még nem megbízhatóak akkor, amikor azt az eredményt adják, hogy a két szöveg szerzője azonos, viszont abban az esetben, ha a modell szerint a két szerző eltérő, a konklúzió magas valószínűséggel helyes. Ennek köszönhetően a modell felhasználható lehet egy-egy író kizárására, mint potenciális szerzője a szövegnek. A rendelkezésünkre álló korpusz még a kutatás későbbi pontjain is kicsi volt, és érdemes lehet nagyobb korpuszon megismételni a kísérleteket, azonban egy ilyen korpusz begyűjtése jelentős kihívásnak bizonyulhat. Egy másik komoly nehézség, melyen felül kell kerekedni az egy olyan módszer találása, mely nem csupán hatékony és megbízható, hanem ezen felül ember által könnyen átlátható és hitelesíthető is. A jelenlegi állás szerint a módszerek megbízhatóságát nehéz hitelesíteni, sok esetben nem elég bizonyíték a rendszer megfelelő működésére az, hogy a teszhalmazon jól teljesített. Ilyen esetek közé tartozik többek között a kriminológiai felhasználás, ahol például fenyegetőlevelek íróit keressük, és ezért ha egy szöveghez hozzátársítunk egy szerzőt, azt megalapozottan kell tennünk a következmények fényében. Az említett instabil-stabil kifejezéseket alkalmazó módszer akár egy lépés lehet ebbe az irányba, ezzel szemben mind az n-gramos módszerek, mind a transzformációs modellek egy másik irányba haladtak el, az ezekből származó döntések nehezen hitelesíthetők. Különösen kriminológiai szempontból kritikus fontosságot élvez egy megfelelően hitelesíthető módszer találása, mely a gyakorlatban is megbízhatóan felhasználható és magyarázható.

7. További irányok, kitekintés

A centrális kérdése a szerző-attribúció feladatának, hogy pontosan mi is határoz meg egy szerzőt, milyen nyomokat hagy magáról, a személyéről egy-egy szövegben. Gyakorlatilag a jellemzőkészlet meghatározása tekinthető kulcslépésnek. Új, egyéni jellemzőkészletekkel akár jobb eredményeket is el lehetne érni. Ügyelni kell arra, hogy a jellemzőkészlet kellően komplex legyen ahhoz, hogy nehéz legyen szándékosan manipulálni (például kriminológiai felhasználás esetén ez kritikus fontosságú), viszont ha túl bonyolult, akkor nehezen értelmezhetővé válik, és ha nem tudja az ember megmagyarázni a modell döntését emiatt, akkor annak hitelessége is kétségbe lesz vonva. Például csábító lehet egy olyan ötlet, ahol helyesírási hibák alapján állítjuk össze a jellemzőkészletet, azonban ez könnyen manipulálható lenne. Egy megfelelő jellemzőkészletet találni, ami a gyakorlatban is felhasználható, vitathatóan az egyik legnehezebb része a feladatnak.

Amennyiben azonban megfelelő jellemzőkészletet találnánk, abból lehetséges, hogy több információt is le lehetne szűrni, mint azt, hogy két szöveget ugyanaz írt-e. Egy aktív kutatási irány például a szerző nemének, korának, nemzetiségének és egyéb információk megállapítása az írása alapján, és eddigi kutatások alapján a szerző-attribúciós technikák ilyen feladatokhoz is alkalmasak. [4] Lehetne ezáltal akár általánosítani is a feladatot, például megállapítani, hogy két szerző ugyanabba a korosztályba tartozik-e, ugyanolyan nemű-e, akár hasonló kultúrában nevelkedtek-e. Annak megállapításához, hogy milyen információk utalnak a szövegben arra, hogy egy adott csoportba (például korosztályba) tartozik egy szerző, érdemes lehet kutatni, hogy milyen megkülönböztető jegyei vannak az adott csoport íróitól származó szövegeknek. A kutatás során ehhez a feladathoz is folytak keresések magyar nyelvű korpuszok után, azonban ez még nehezebb feladatnak bizonyult, mint a korábban taglalt, klasszikus szerző-attribúció feladathoz találni megfelelő korpuszt. Ilyen jellegű korpuszok alkotása illetve összegyűjtése nagy jelentőséggel bírna a kutatási irány továbbvitelében. A szerző-attribúció feladatához elégséges lenne olyan korpusz is, ahol a szövegekhez akár felhasználónevek társulnak, még a szerző információi nem is elérhetőek, vagy nem oszthatók meg.

Szintén érdemes lehet más jellegű írott szövegeken is tesztelni a módszereket, például beszédleiratokon. Ez az irány még kevésbé van jelen az irodalomban, mint a klasszikus szerző-attribúció, ugyanakkor erre is szükség van.

Irodalomjegyzék

- [1] Juola, Patrick: The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. Digital Scholarship in the Humanities. 30. fqv040. 10.1093/llc/fqv040. (2015)
- [2] Sally McGrane: Diary of the Hitler Diary Hoax | The New Yorker (2013)
- [3] Eric Pace: Konrad Kujau, 62, 'Hitler Diaries' Swindler (2000)
- [4] Juola, Patrick: Authorship attribution. Foundations and Trends® in Information Retrieval. 1. 233-334. 10.1561/15000000005. (2008)
- [5] Chen Qian, Tianchang He, Rao Zhang: Deep Learning based Authorship Identification] (2017)
- [6] Sarwar, Raheem & Yu, Chenyun & Tungare, Ninad & Chitavisutthivong, Kanatip & Sriratanawilai, Sukrit & Xu, Yaohai & Chow, Dickson & Rakthanmanon, Thanawin & Nutanong, Sarana: An Effective and Scalable Framework for Authorship Attribution Query Processing. IEEE Access. 6. 10.1109/ACCESS.2018.2869198. (2018)
- [7] Russell-Rose, Tony & Stevenson, Mark & Whitehead, Miles: The {Reuters Corpus Volume 1} -- from Yesterday's News to Tomorrow's Language Resources. (2002)
- [8] Project Gutenberg: <https://www.gutenberg.org/> Utolsó látogatás: 2022. 04. 06.
- [9] Gervain Judit: Magyar Tudomány 172. évf. 8. sz. A csecsemőkori beszédészlelés mechanizmusai és a nyelvelsajátítás (2011)
- [10] Koppel, M., Akiva, N. and Dagan, I.: Feature instability as a criterion for selecting potential style markers. J. Am. Soc. Inf. Sci., 57: 1519-1525. <https://doi.org/10.1002/asi.20428> (2006)
- [11] V. Keselj, F. Peng, N. Cercone, és C. Thomas: "N-gram-based author profiles for authorship attribution," Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING03, 255–264, Dalhousie Egyetem, Halifax, NS, (2003. augusztus.)
- [12] <https://docs.python.org/3/library/functions.html#> Utolsó látogatás: 2022. 04. 06.
- [13] <https://cran.r-project.org/web/packages/stylo/index.html> Utolsó látogatás: 2022. 04. 06.
- [14] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html Utolsó látogatás: 2022. 04. 06.
- [15] F1-pontszám és metrikái, <https://www.tutorialexample.com/an-introduction-to-accuracy-precision-recall-f1-score-in-machine-learning-machine-learning-tutorial/> Utolsó látogatás: 2022. 04. 06.
- [16] Eder M, Rybicki J, Kestemont M: "Stylometry with R: a package for computational text analysis." R Journal, 8(1), 107–121. <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>. (2016)
- [17] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018. október 11.)
- [18] Nemeskey, Dávid Márk: "Natural Language Processing methods for Language Modeling". PhD thesis. Eötvös Loránd University. (2020)

- [19] Papineni, Kishore, et al: "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics (311-318. oldal). (2002)
- [20] Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch; (1975. Február 1.)
- [21] A Flesch-Kincaid Grade Level, <https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/> Utolsó látogatás: 2022. 04. 09.
- [22] Coleman, Meri; and Liau, T. L.: A computer readability formula designed for machine scoring, Journal of Applied Psychology, Vol. 60, (283-284. oldal) (1975)
- [23] A Gunning Fog Index, <https://readabilityformulas.com/gunning-fog-readability-formula.php> Utolsó látogatás: (2022. 04. 09.)
- [24] A Linsear Write Readability Formula, <https://readabilityformulas.com/linsear-write-readability-formula.php> Utolsó látogatás: (2022. 04. 09.)
- [25] A SMOG, <https://readabilityformulas.com/smog-readability-formula.php> Utolsó látogatás: (2022. 04. 09.)
- [26] BLEU számítása az NLTK csomaggal, https://www.nltk.org/_modules/nltk/translate/bleu_score.html Utolsó látogatás: (2022. 04. 10.)
- [27] A Hunglish korpusz, <http://mokk.bme.hu/resources/hunglishcorpus/> Utolsó látogatás: (2022. 04. 10.)
- [28] FTP link a Hunglish korpuszhoz, <ftp://mokk.bme.hu/Hunglish> Utoljára használva: (2022. 04. 10.)
- [29] Vincze, Veronika, et al.: "A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei." 275-288. (2021)