A mesterséges intelligencia alkalmazása görög irodalmi szövegek elemzésére

The application of artificial intelligence for the analysis of Greek literary texts

Dr. TÓTH Erzsébet[1], Dr. GÁL Zoltán[2]

[1]Debreceni Egyetem, Informatikai Kar, 4028 Debrecen, Kassai út 26., toth.erzsebet@inf.unideb.hu
[2]Debreceni Egyetem, Informatikai Kar, 4028 Debrecen, Kassai út 26., gal.zoltan@inf.unideb.hu

Abstract

In our paper we have elaborated a classification model in which several hundred different old Greek subtexts were used for supervised learning with the purpose of subtext class recognition. We could determine a triplet of a, b, c values for describing a power function which fits precisely to a curve determined by the word frequencies in the selected texts. Concerning 200 subtexts the triplet of a, b, c values, the classes of the subtexts and their 16 dimensional feature vectors were learnt for the Recurrent Neural Network (RNN). We concluded that the Long-Short Term Memory RNN predicted efficiently which class a chosen subtext could be categorized into.

Keywords: deep learning; old Greek literary texts; text classification; Recurrent Neural Network (RNN); Long-Short Term Memory

Kivonat

Dolgozatunkban egy olyan osztályozási modellt fejlesztettünk ki, amiben több száz különböző ókori görög szövegentitást használtunk ellenőrzött tanulásra abból a célból, hogy az felismerje a szövegentitások osztályát. Meghatároztuk az (a, b, c) hármas értékeit egy olyan hatványfüggvény leírására, amely pontosan illeszkedik a kiválasztott szövegekben lévő szavak relatív gyakorisága által megadott görbére. A 200 darab szövegentitással kapcsolatban az (a, b, c) hármas értékeinek becsléséhez a szövegentitások osztály azonosítóját és a 16 dimenziós tulajdonság („feature”) vektorokat használtuk fel a Visszacsatolásos Neurális Hálózat (RNN – Recurrent Neural Network) betanításához. Arra a következtetésre jutottunk, hogy az LSTM (Long-Short Term Memory) RNN hálózat hatékonyan előrejelezte számunkra, hogy a kiválasztott szövegentitás melyik osztályba sorolható.

Kulcsszavak: mély tanulás; ókori görög irodalmi szövegek; szövegosztályozás; Visszacsatolásos Neurális Hálózat (RNN); Long-Short Term Memory hálózat

1. BEVEZETÉS

A híres Alexandriai Könyvtár az eredeti, hiteles görög irodalmi szövegek összes felkutatható példányát gyűjtötte az ókori időszakban. Ilyen értelemben egyetemes könyvtárnak tekinthető, amely az emberi tudás gyakran idézett szimbólumát testesíti meg egészen napjainkig. Annak gyűjteményében az ókori szövegek a szerzői nevek betűrendjébe voltak sorolva, valamint különböző irodalmi műfajok szerint csoportosítva a könyvtár híres katalógusa, az úgynevezett Pinakes alapján. Ezt a neves katalógust az ókori görög tudós, Kallimakhosz állította össze a Kr. e. 3. században. Kutatásunkban az ókori szövegek elemzéséhez a Kallimakhosz-féle kidolgozott osztályozási rendszert használtuk.

Kapcsolódó kutatásként megemlíthetjük az ókori Alexandriai Könyvtár 3D virtuális könyvtár modelljét (3DVLM – 3D Virtual Library Model) megvalósító fejlesztést, amely a Kognitív Infokommunikációk (Cognitive InfoCommunications) kutatási keretrendszer [1, 2] részeként jött létre 9 évvel ezelőtt [3]. A jelenlegi modell a MaxWhere Szemináriumi rendszer [4] 3D megjelenítési és navigációs lehetőségeit aknázza ki és a magyar diákok idegennyelv tanulását támogatja angol és magyar nyelvű hiperszöveges tananyagával haladó és középhaladó nyelvi szinten [5].

2. ELEMZÉSI MÓDSZERTAN ÉS EREDMÉNYEK

Az ókori szövegek különböző események sorozatát írják le adott kronológiai sorrendben, akár több idősíkban mozogva. A szöveg szerzője mindig a saját stílusát érvényesítve fogalmazza meg a történetét. Mindez pedig a szöveg kohézióját eredményezi, ami előnyösen felhasználható minden egyes szöveghez tartozó szövegrészlet automatikus leírására, jellemzésére. A dolgozatban adott szöveg egymás utáni mondataiból képezett részletet szövegentitásnak nevezzük, így a szöveg szövegentitások sorozataként fogható fel. A továbbiakban bemutatott elemzési módszer szoros korrelációt mutat a mondatrészek relatív gyakoriságának mintázata és a különböző szövegosztályok típusa között.

2.1. A feldolgozott szövegek alap jellemzői

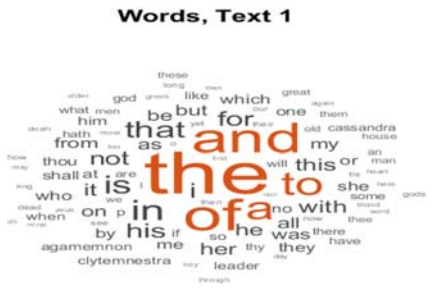
A Gutenberg projektből összesen 37 darab szöveget töltöttünk le, amelyek a költők és prózaírók két fő csoportjába tartoztak. Kallimakhosz a szövegeket további 6 alkategóriába sorolta azok irodalmi műfaja szerint (lásd az 1. táblázatot).

A vizsgált ókori szövegek kategóriái 1. táblázat

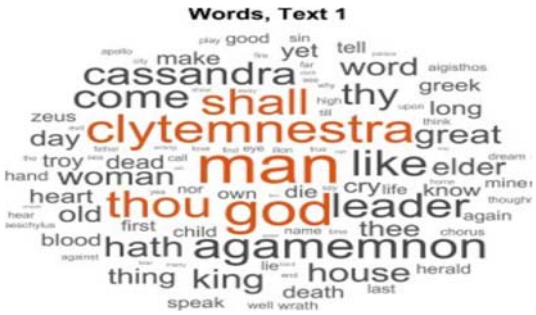
	Kallimakhosz fő- és alkategóriái az ókori irodalmi művek esetén	Szövegek száma
1.	Költők – tragédia írók	8
2.	Költők – vígjáték írók	8
3.	Költők - epikusok	11
4.	Költők -lírikusok	2
5.	Prózaírók - filozófusok	5
6.	Prózaírók - szónokok	3

2.2. Szövegek letisztítása és előfeldolgozása

A különböző megjelenésű írásokat az automatizált feldolgozás céljából egyenszilárdságú elemek sorozatára szükséges módosítani. Így a szövegekből eltávolítottuk a stopszavakat, rövid szavakat ($|w| \leq 2$), hosszú szavakat ($|w| > 15$) és figyelmen kívül hagytuk az archaikus angol nyelvezetre jellemző szavakat, mint például: „thou”, „hath”, „thy”, stb. Adott „w” szó esetén $|w|$ a karakterszámban kifejezett szóhosszt jelenti. Ezek a letisztított szövegek témaspecifikus szavakat eredményeznek a szó-rangsor legelején (lásd a 4. ábrát). Ebben a kontextusban a szavak rangja azok előfordulási gyakoriságának a sorrendjét jelzi a vizsgált szövegben. Szófelhők segítségével jeleníthetjük meg a tokenek (szózsetonok) gyakoriságának rangsorát. Nyilvánvaló ok miatt minél nagyobb a karakterek száma, annál nagyobb a tokenek előfordulási gyakorisága a szövegben.



1. ábra. Eredeti szöveg$_1$



2. ábra. Letisztított szöveg$_1$

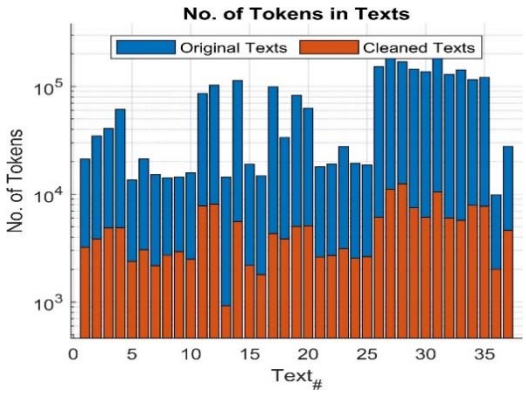
Az N-Gram egy olyan N darab tokenből ($N = 1, 2, \ldots$) álló csoportot jelöl, amelyek közvetlenül egymás mellett helyezkednek el a szövegben. A 2-Gram-ok olyan egymás melletti szópárokat jelentenek, amelyeknek kötött az előfordulási sorrendje. A felesleges tokenek eltávolítására a Matlab rendszer beágyazott programcsomagjait használtuk („Text Analysis Toolbox”, „Machine Learning and Deep Learning Toolbox”). A modern angol nyelvezetre jellemző stopszavak listáját alkalmaztuk és integráltuk a szoftver aktuális verziójába. Az eredeti szöveg$_1$-re és ugyanezen letisztított szövegre kapott szófelhők a 1. és a 2. ábrán láthatók (a szöveg$_1$ eredeti szerzője: Aeschylus, annak fordítója: Murray, Gilbert címe: Agamemnon, Kallimakhosz fő- és alkategóriái: költők – tragédia írók).

A stopszavak torzító hatásának kiküszöbölésére és a szó-ranghoz illeszkedő paraméterek előállítására mind a $k = 37$ darab szöveg esetében kifejlesztettük adott tetszőleges szöveg letisztító és illesztő algoritmusát. A mindegyik függvény által végrehajtott feladatok magyarázatát a 2. táblázat tartalmazza.

A szövegfeldolgozás függvényei 2. táblázat

	Függvény neve	Függvény tevékenysége, hatása
1.	Import()	Szöveg importálása állományból.
2.	Lower()	Sztring konvertálása kisbetűkre.
3.	TokenizeDoc()	Sztring konvertálása tokenekre.
4.	AddPartOfSpeechDetails()	Mondatrész jellemzők hozzáadása.
5.	NormalizeWords()	Mindegyik token szótőre redukálása.
6.	RemoveStopWords()	A modern angol nyelv stopszavainak törlése.
7.	ErasePunctuation()	Központozás tokenek törlése.
8.	RemoveShortWords()	Rövid tokenek törlése. ($\|w\| \leq 2$)
9.	RemoveLongWords()	Hosszú tokenek törlése. ($\|w\| \geq 16$)
10.	BagOfWords()	Szócsoportok generálása.
11.	TopkWords()	Toplistás szavak generálása.
12.	FitCurve()	Hiperbola görbe illesztése a toplistára.

A 3. ábra jól tükrözi az eredeti és a letisztított szövegek hossza közötti viszonyt. Megfigyelhető, hogy a letisztítás után megmaradt tokenek aránya a teljes eredeti szómennyiségnek körülbelül 10%-a, amely mennyiség viszont tartalom specifikus. Tehát a letisztítási feladatok eredményeként a szövegek hossza egy nagyságrenddel csökkent.



3. ábra. A szöveghossz csökkenése az előfeldolgozás után

2.3. Különböző angol nyelvű szövegek szavainak tulajdonságai

A Kaggle adatbázis [6] 333 333 a weben leggyakrabban használt angol szót tartalmazza, amelynek forrása a Google Web Trillion Word Corpus. A szavak relatív gyakorisága és azok rangja kerül megjelenítésre a 4. ábrán. A szó intenzitás értékeire illeszkedő görbe egyenlete a következő:

$$y(x) = \exp\left(\alpha \cdot x^{\beta} + \gamma\right) \tag{1}$$

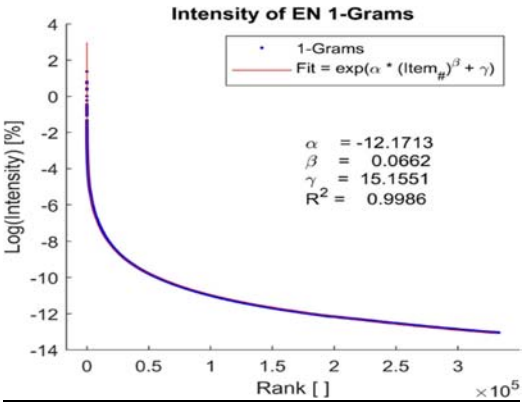
ahol y a szavak számát jelöli, x pedig a 4. ábrának megfelelő szavak rangját. Az (α, β, γ) paraméter hármas az Interneten használt modern angol nyelvezetre jellemző. Megjegyezzük, hogy ez a lista az angol nyelvben használatos stopszavakat is tartalmazza, amelyek erőteljesen befolyásolják a görbét a szó-rangok legelején. A stopszavak, rövid szavak ($|w| \leq 2$) és a hosszú szavak ($|w| > 15$) eltávolításának hatására az angol nyelvű szövegekből a fennmaradó szavak relatív gyakorisága megváltozik, és az nagymértékben függ a szöveg kontextusától. Jelen esetben is $|w|$ a „w” szó hosszát jelenti karakterszámban megadva.

A modern angol nyelvezetre jellemző szógyakoriságot összevethetjük az elő-feldolgozott ókori görög irodalmi szövegekre jellemző szógyakorisággal, azaz, hogy az milyen szóintenzitás görbét eredményez

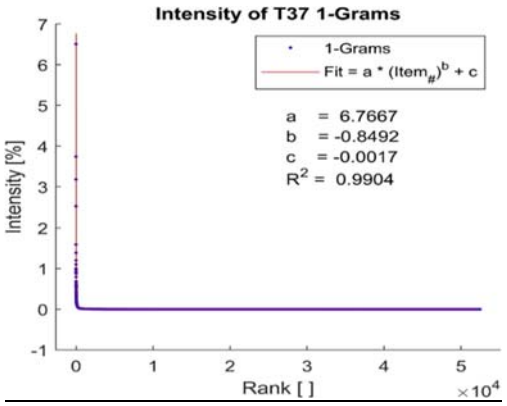
számunkra. Ezeket az irodalmi szövegeket letisztítottuk a lényeges tartalom leszűrése céljából. A szóintenzitás a 37 szövegnek egyetlen aggregált formára összevont szövegváltozata esetén az 5. ábrán jelenik meg. A szó intenzitás értékekre illeszkedő görbe egyenlete az alábbi:

$$y(x) = a \cdot x^b + c \tag{2}$$

ahol *y* a szavak számát jelöli, *x* pedig a 5. ábrának megfelelő szavak rangját. Az 1. és a 2. egyenlet nagyon eltér egymástól. Az előbbi egyenlet egy hatványfüggvény exponenciálisa, míg az utóbbi csupán hatványfüggvény. Az (*a, b, c*) paraméter hármas az elemzett ókori angol nyelvezet speciális jellemzőit mennyiségi formában fejezi ki.



4. ábra. *Angol szavak gyakorisága az Interneten*



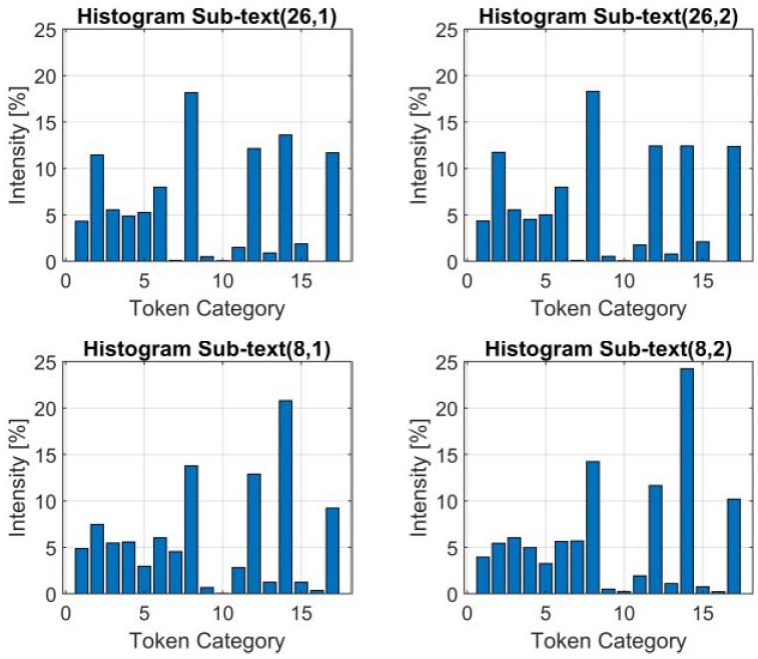
5. ábra. *Angol szavak gyakorisága az ókori szövegek fordításában*

A 4. és az 5. ábrák alapján azt a következtetést vonhatjuk le, hogy a modern angol nyelv hajlamos intenzívebben használni az általános szavakat. Ezeket a felesleges szavakat hívjuk stopszavaknak a szövegek gépi tanulás alapú elemzései esetén. A Matlab *AddPartOfSpeechDetails()* függvénye segítségével (lásd a 2. táblázatot) előállítottuk mindegyik letisztított szövegentitás *tokenDetails* objektumát. Ez az objektum lehetővé tette számunkra, hogy mindegyik token kategóriával rendelkezzünk (lásd a 3. táblázatot).

Token kategóriák felsorolása — 3. táblázat

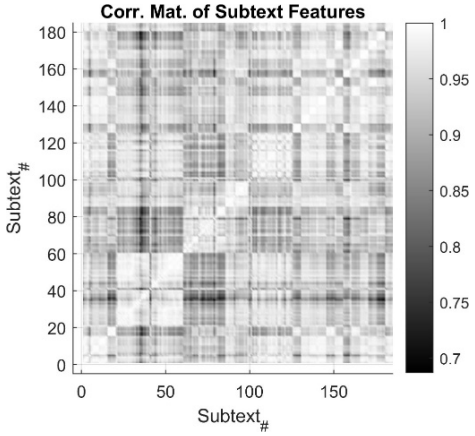
ID	Token kategória	ID	Token kategória	ID	Token kategória	ID	Token kategória
1	melléknév	5	kötőszó	9	számnév	13	írásjel
2	értelmező	6	elválasztószó	10	viszonyszó	14	alárendelőszó
3	határozószó	7	indulatszó	11	névmás	15	szimbólum
4	segédige	8	főnév	12	tulajdonnév	16	ige

Ezután a token kategóriák relatív számát előállítottuk minden egyes szövegentitásra vonatkozóan, majd ez alapján azok hisztogramjait hoztuk létre. A négy különböző szövegentitásra vonatkozó token kategóriák hisztogramjai a 6. ábrán láthatók. Az ábrázolt intenzitás értékek normalizáltak és összesen 100%-ot adnak. Mindegyik token kategória értéke 0,25%-os értéktartományon belül fordult elő, függetlenül a szövegentitás osztályoktól. A 17. „egyéb” token kategóriát nem vettük figyelembe, mert az a 16 token kategória intenzitásának lineáris kombinációja. Megjegyezzük, hogy néhány token kategória (pl. a szimbólum, a számnév és a viszonyszó) nagyon alacsony intenzitással rendelkezik, míg más kategóriák magasabb gyakoriság értékekkel rendelkeznek, mint például a főnév és az alárendelőszó. A többi kategória (pl. az értelmező, az indulatszó) gyakoriságánál megfigyelhető, hogy azok erőteljesen ingadoznak mindez pedig eltéréseket eredményez a különböző szövegentitások számszerűsített jellemzői („feature”-ök) között.

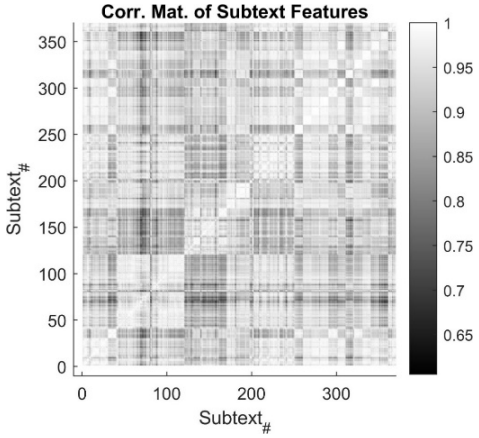


6. ábra. *A szöveg$_{26}$ és a szöveg$_{8}$ token kategóriáinak a hisztogramjai. A felső sorban a szöveg$_{26}$ látható, az alsó sorban a szöveg$_{8}$. Az ábra bal oldali oszlopában a szövegentitás$_{1}$, annak a jobb oldali oszlopában a szövegentitás$_{2}$ található.*

Megjegyezzük, hogy a Normalizált Token Kategória Gyakoriság (NTCF) a szöveg$_{26}$ és szöveg$_{8}$ szövegentitásai esetén került bemutatásra, ahol i = 1, 2, … m, és m = 10. A különböző szövegentitások számszerűsített jellemzőit azok vizsgált token kategóriáinak relatív hisztogramjai tükrözik.



7. ábra. S*zövegentitás jellemzők korrelációs mátrixa (*m = 5)

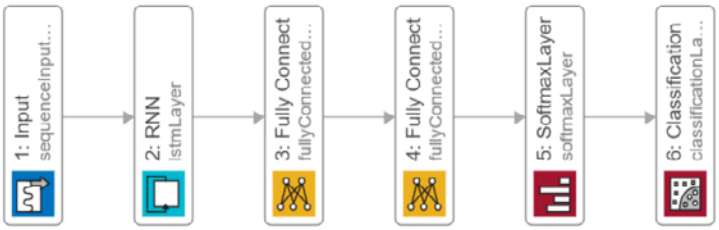


8. ábra. S*zövegentitás jellemzők korrelációs mátrixa (*m = 10)

A 7. ábrán lévő korrelációs mátrixon egyértelműen látható, hogy az 10*10 pixel méretű cellákból áll, amelyek homogén színű négyzet és hosszú vonal mintázatokat alkotnak, amelyek sorokként és oszlopokként jelennek meg a mátrixban. A főátló közvetlen környezetében található fehér színű cellák értéke közel 1, mivel ugyanannak a szövegnek a szövegentitásai erős korrelációt mutatnak egymással. Ez erős szövegkohéziót tükröz. A sötétebb tónusú téglalapok és vonalterületek olyan szövegpárokhoz tartoznak, amelyek már eltérő Kallimakhosz-féle osztályba sorolhatók (v. ö. a 7. és a 8. ábrán kapott korrelációs mátrixokat). A 8. ábrán megfigyelhető, hogy a korrelációs mátrix értéktartománya [0,6 … 1], valamint az elemek átlagértéke 0,9, ami a szövegentitások szavai gyakoriságának erős korrelációját igazolja. A korreláció alacsonyabb értékét az összehasonlított szövegentitások kisebb számú tokenjei okozzák. Evidens számunkra, hogy a túl rövid szövegentitások nem fognak egymásra hasonlítani. Szélsőséges esetekben a szövegentitások csak egyetlenegy tokent tartalmaznak és az eltérő tokenjeik különböző osztályokba tartoznak, mindez pedig alacsony korrelációra enged következtetni.

2.4. Szövegek Kallimakhosz-féle osztályokba sorolása LSTM RNN-nel

A vizsgált irodalmi szövegek viszonylag kis száma miatt azokat szövegentitásokra daraboltuk. Ezeknek a szövegeknek egy fontos számszerűsíthető jellemzőjét, nevezetesen a szavak gyakoriság megoszlását használtuk fel a szövegek sajátosságainak leírására. Megfigyelhető, hogy adott szöveghez tartozó szövegentitások hasonlóan viselkednek, vagyis kvantitatív mértéket képeznek e jellemzőnek megfelelően. Ebben a megközelítésben $37x10 = 370$ darab szövegentitással rendelkezünk és azok mindegyike leírható egy 16 dimenziós tulajdonság („feature") vektorral. Ezek a szövegentitások 6 Kallimakhosz szerinti kategóriába sorolhatók az irodalmi műfaj szerint. Külön-külön 200, 85 és másik 85 darab eltérő szövegentitást használtunk a felügyelt tanulás betanítás, validálás és tesztelés feladatainak elvégzése során. A Rekurrens Neurális Hálózat (RNN) architektúrája a 9. ábrán látható. Hat szintből áll, középen két azonos típusú réteggel.



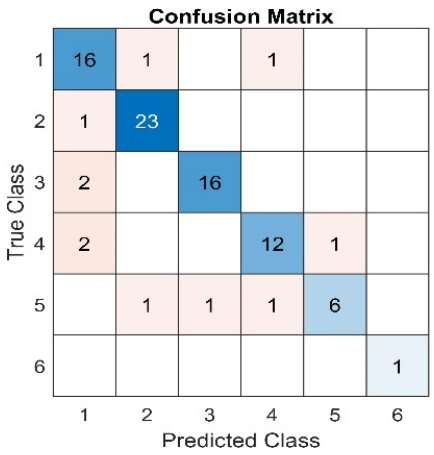
9. ábra. Az *LSTM RNN felépítése*

Az RNN típusa Long-Short Term Memory (LSTM) volt, aminek paraméterei az alábbiak:

- Solver: ADAM
- Gradient Decay Factor = 0.90
- Squared Gradient Decay Factor = 0.99
- Initial Learn Rate = 0.02
- Gradient Threshold = 1
- Max Epochs = 1000
- Mini Batch Size = 200
- Hidden units# on L2 = 100
- Classes# on L3 = 100
- Classes# on L4 = 6.

Az RNN betanítási folyamata a veszteség és pontosság jellemzők időbeni alakulása a 10. ábrán látható.

10. ábra. *A betanítás vesztesége és pontossága*



11. ábra. *A tesztelés tévesztési mátrixa*

A bal oldali és jobb oldali függőleges tengelyek a veszteséget és a pontosságot mutatják külön, külön. A veszteség exponenciális trendjét a logaritmikus skálájú bal oldali tengelyen a lineáris görbe igazolja. Megjegyezzük, hogy az RNN tanulási folyamata egy asztali számítógépen futott és viszonylag rövid ideig, csupán 50,02 mp ideig tartott. A maradék 85 szövegentitás osztályba sorolásának tesztje során megállapítottuk, hogy a szövegentitások osztályozásának pontossága 87,06% volt, a fellépő veszteség pedig 0,42%. A 11. ábra szemlélteti számunkra a 85 tesztelt szövegentitás besorolásának tévesztési mátrixát. A kapott eredmény alapján megállapítható, hogy csupán a szavak számának gyakorisága alapján végzett feldolgozás segítségével általgosan kevesebb, mint minden 5. szövegentitás azonosítható be hibásan a megfelelő Kallimakhosz-féle osztállyal a szövegentitás jelentésének mélyebb értelmezése nélkül [7].

3. ÖSSZEFOGLALÁS ÉS KÖVETKEZTETÉSEK

A dolgozatban angol nyelvre fordított ókori görög szövegeket elemeztünk. Az ókori szövegek osztályozásának meghatározó személyisége, Kallimakhosz kategóriáit és alkategóriáit alkalmaztuk az elemzés során. Megfelelő előfeldolgozás, valamint az angol nyelvben létező általános, mondatkialakítási segédszavak, ún. stopszavak kiszűrése után a tartalom jellemzéséhez csupán mennyiségi mutatókat alkalmaztunk. Az angol nyelv időbeni változásának érzékelése céljából a weben használatos mai szövegek szavainak gyakoriságát összevetettük az ókori görög szövegek hasonló jellemzőivel. Vizsgálatunkkal kapcsolatban a következő összegző megállapításokat tehetjük: a letisztítás után megmaradt tokenek aránya ~ 10%. A letisztítás után maradt tokenek tartalom specifikusak, azaz kellő biztonsággal a tulajdonság („feature") összetevőjeként használhatók fel a szövegentitás kategóriába sorolásához. A normalizált token kategória gyakoriság (NTCF – Normalized Token Category Feature) homogén négyzetekben jellemzi a szövegentitásokat, vagyis ugyanahhoz a szöveghez tartozó szövegentitások NTCF vektora közeli, míg eltérő szövegek esetén e vektorok lényegesen különböznek. A mennyiségi NTCF vektor tehát minőségileg jellemzi a szövegentitást az osztályba sorolásnál. Az ókori görög szövegek 170 elemzett szövegentitásának osztályozási jellemzői a következők: pontosság 87,06%, tanulási idő: 50,02 mp. Jelen mennyiségi elemzési módszert aktuálisan további, más témájú szövegek tartalmi jellemzésénél is alkalmazzuk. Nyitott kérdés jelenleg a veszteséges hálózati átviteli mechanizmusok hatásának mértéke a különböző kategóriájú szövegekre vonatkozóan.

KÖSZÖNETNYILVÁNÍTÁS

Ezt a kutatást a QoS-HPC-IoT Laboratórium és a Debreceni Egyetem TKP2021-NKTA projektje támogatta. A TKP2021-NKTA-34 projektet a Magyarországi Nemzeti Kutatási, Fejlesztési és Innovációs alap támogatta a TKP2021-NKTA finanszírozási formának megfelelően.

IRODALMI HIVATKOZÁSOK

[1] Baranyi P., Csapó Á. *Definiton and synergies of Cognitive InfoCommunications*. Acta Polytechnica Hungarica, 2012, 9(1), 67–83.

[2] Baranyi P., Csapó Á., Sallai Gy. *Cognitive Infocommunication (CogInfoCom)*, Springer International, Heidelberg [etc.], cop. 2015.

[3] Boda I., Bényei M., Tóth E. *New dimensions of an ancient Library: the Library of Alexandria*. In: CogInfoCom 2013. Proc. of the 4th IEEE International Conference on Cognitive Infocommunications, (Budapest, Hungary, December 2-5, 2013,) 537–542.

[4] *MaxWhere VR Even More*: https://www.maxwhere.com/ (Utolsó letöltés: 2022. 09.10).

[5] Boda I. K, Tóth E, T. Nagy L.: *Improving a bilingual learning material in the three-dimensional space using Google Translate*. In: CogInfoCom 2022. Proc. of the 13th IEEE International Conference on Cognitive Infocommunications, (Online on 3D MaxWhere, September 21-23, 2022,) 6 p. (accepted for presentation)

[6] *Kaggle Database: 1/3 Million Most Frequent English Words on the Web*: https://www.kaggle.com/rtatman/english-word-frequency (Utolsó letöltés: 2022.09.13)

[7] Gál, Z., Tóth, E.: *Deep learning-based analysis of ancient Greek literary texts: A statistical model based on word frequency for the classification of texts*. In: 12th IEEE International Conference on Cognitive Infocommunications: CogInfoCom 2021: Proceedings. Ed.: Jan Nikodem, Ryszard Klempous, IEEE-INST Electrical Electronics Engineers INC, Piscataway, 529-535, 2021. ISBN: 9781665424950