

A NEMZETI ALAPTANTERV ELEMZÉSE

Nyéki Lajos, nyeki@sze.hu
Széchenyi István Egyetem

1. Bevezetés

Az előadás a hatályos 110/2012. (VI. 4.) Korm. rendelet és melléklete alapján elemzi a Nemzeti Alaptantervet.

A rendelet mellékletét tartalmi egységekre bontja, és a rendeletből és a melléklet tartalmi egységeiből álló korpuszt elemzi szövegbányászati eszközökkel.

Az elemzés az R szöveg- és adatbányászati csomagjaira épül.

2. A szövegbányászat folyamata

A szövegbányászat folyamatának lépései a következők:

- Nyers, strukturálatlan információ gyűjtése (web oldalak, e-mail-ek, dokumentumok).
- Előfeldolgozás (a nyers, strukturálatlan információ strukturált formátumúvá alakítása).
- Adatbányászati technikák (felügyelt, félig felügyelt és nem felügyelt) alkalmazása.

3. A dokumentumok beszerzése

A dokumentumok forrása a Jogtárban megjelent 110/2012. (VI. 4.) Korm. rendelet a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról.¹⁸

A Jogtárban megjelent web alapú jogszabályt és mellékletét azok szövegének kijelölése, másolása, majd az EditPad 7 Lite szövegszerkesztőbe való beillesztés után tudtuk elmenteni utf-8 kódtáblájú szöveggént. Az általunk használt szövegbányászati csomagok esetén ugyanis az utf-8 kódtábla az alapértelmezés.

A bemeneti, txt formátumú adatállományok a következők:

- A 110/2012. (VI. 4.) Korm. rendelet.

¹⁸ 110/2012. (VI. 4.) Korm. rendelet a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról, <https://net.jogtar.hu/jogszabaly?docid=a1200110.kor>, (2019-10-28)

- Az iskolai nevelőmunka tartalmi szabályozása és a szabályozás szintjei.
- A kulcskompetenciák.
- A Nat műveltségi területeinek felépítése, ajánlás a Nat műveltségi területek százalékos arányaira.
- A műveltségi területek anyagai.
- A glosszárium.

4. A Nat szerkezete, tartalmi észrevételek

A Nat szerkezete:

I. Az iskolai nevelőmunka tartalmi szabályozása és a szabályozás szintjei

II. Kompetenciafejlesztés, műveltségközvetítés, tudásépítés

II.1 A kulcskompetenciák

II.2 Műveltségi terület

II.2.1. A Nat műveltségi területeinek felépítése

II.2.2. Ajánlás a Nat műveltségi területek százalékos arányaira

II.2.3. Ajánlás a nemzeti iskolákban a Nat műveltségi területek százalékos arányaira

II.3 A műveltségi területek anyagai

III. Glosszárium

Tartalmi észrevételek:

A glosszárium rész néhány „érdekes” megfogalmazást tartalmaz a képesség, a követelmények és a tanulói tevékenység címszavakban.

A képességek a glosszárium szerint veleszületettek és fejleszthetők¹⁹. A megállapítás első része téves.

A magyar pedagógiai terminológia szerint az adottságok velünk született diszpozíciók, a képességfejlődés kiindulási alapjai. A Nat szerzői feltehetően az angolszász „aptitude” fogalmat fordították helytelenül képességgnek. Az angolszász szóhasználatban az „aptitude” veleszületett, az „ability” tanulás révén alakul ki.²⁰

¹⁹ Képesség, Nat, <https://net.jogtar.hu/jogszabaly?docid=a1200110.kor>, (2019-10-28)

²⁰ Aptitude, Wikipedia, <https://en.wikipedia.org/wiki/Aptitude>, (2019-10-28)

A követelmények fogalom a glosszárúum szerint többféle értelemben használatos. Jelenti egyrészt a fejlesztő feladatok összességét, másrészt a tantervi követelményeket.

Emellett a glosszárúum külön definiálja a vizsgakövetelmények fogalmát. Ráadásul a glosszárúum a tanulás eredményeit kimeneti követelményként értelmezi.

A Nat szövegében a követelmények fejlesztési feladatok formájában jelennek meg.

A tanulói tevékenység a glosszárúum szerint azon tevékenységek rendszere, melyek során - tanítói, tanári tanácsadás, irányítás és értékelés mellett - a tanuló feldolgozza, elsajátítja a tantervben definiált iskolai tananyaghoz rendelt (iskolában elsajátítható, elsajátítandó) követelményeket.

A tanulónak véleményünk szerint nem a követelményeket kell elsajátítania, hanem a követelmények alapján összeállított tananyagot.

5. A szóalakok problémája

A magyar mondatokban a szavakhoz végződés (jelek, képzők, ragok) társulnak, illetve a szavaknak lehetnek elő- és utótagjai. A szövegbányászat során alapvető feladat a szavak szótári alapalakjának meghatározása. Ennek egyik módja a szótövezés, a másik a lemmatizálás.

A szótövezés (stemming) a jelek, a képzők vagy a ragok leválasztásával próbálja meghatározni a szótári alapalakot. Ez sokszor nem sikerül. A magyar nyelvhez a Snowball stemmer használható. Pl.: végzett -> végz, ablakból -> ablak

A lemmatizálás a szótári alapalak meghatározása szabályok vagy szótár használatával. A szótár alapú lassabb, de pontosabb, a szabály alapú gyorsabb, viszont kevésbé pontos. Pl.: mesterszakon -> mesterszak, idegen -> ideg.

6. Az alkalmazott szoftverek

A szövegbányászati feladat megoldásához az udpipe és a textrank R alapú csomagokat használtuk.

A lemmatizáláshoz a szegedi egyetem nyelvtechnológiai csoportja által kifejlesztett hungarian-szeged-ud-2.4-190531.udpipe²¹ nyelvi modellt használtuk.

²¹ UD 2.4 Models for UDPipe, <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2998>, (2019-10-28)

A nyelvi modell hibrid megközelítést alkalmaz, szabályokat és szótárt is használ a lemmák előállításához.

Az udpipe csomag elsődleges feladata az annotáció elvégzése. Ennek során a korpusz dokumentumait bekezdésekre, mondatokra és tokenekre bontja. A tokenhez meghatározza a lemmát és a Universal Dependencies²² terminológia szerinti szófajt (upos), és szintaktikai reláció²³ elemzést is végez (dependency parsing). Az annotáció kimeneti táblázatának első nyolc oszlopát az 1. ábra mutatja.

doc_id	paragraph_id	sentence_id	sentence	token_id	token	lemma	upos
<chr>	<int>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>
doc1	1	1	nemzeti alaptanterv i.	1	nemzeti	nemzeti	ADJ
doc1	1	1	nemzeti alaptanterv i.	2	alaptanterv	alaptanterv	NOUN
doc1	1	1	nemzeti alaptanterv i.	3	i	i	ADJ
doc1	1	1	nemzeti alaptanterv i.	4	.	.	PUNCT
doc1	1	2	rész az iskolai nevelő-oktató munka tartalmi szabályozása és szabályozási szintjei i. 1. a köznevelés feladata és értékei	1	rész	rész	ADJ
doc1	1	2	rész az iskolai nevelő-oktató munka tartalmi szabályozása és szabályozási szintjei i. 1. a köznevelés feladata és értékei	2	az	a	DET
doc1	1	2	rész az iskolai nevelő-oktató munka tartalmi szabályozása és szabályozási szintjei i. 1. a köznevelés feladata és értékei	3	iskolai	iskolai	ADJ
doc1	1	2	rész az iskolai nevelő-oktató munka tartalmi szabályozása és szabályozási szintjei i. 1. a köznevelés feladata és értékei	4	nevelő-oktató	nevelő-oktató	ADJ
doc1	1	2	rész az iskolai nevelő-oktató munka tartalmi szabályozása és szabályozási szintjei i. 1. a köznevelés feladata és értékei	5	munka	munka	NOUN
doc1	1	2	rész az iskolai nevelő-oktató munka tartalmi szabályozása és szabályozási szintjei i. 1. a köznevelés feladata és értékei	6	tartalmi	tartalmi	ADJ

9. ábra

A nyelvi modell lemmatizálási és egyértelműsítési hibáinak kiküszöbölésére saját szótárt kellett építenünk.

A lemmatizálási hibák javítására szolgáló R feltételes utasításokra néhány példa:

```
x$lemma <- ifelse(x$token=="antik", "antik", x$lemma)
x$lemma <- ifelse(x$token=="anyagok", "anyag", x$lemma)
x$lemma <- ifelse(x$token=="atom", "atom", x$lemma)
```

A nyelvi modell az antik szót ant-ra, az atom szót at-ra szótövezte volna.

A szótár a jelenlegi állapotában 953 lemmatizálási hibát javít, ez ugyanennyi feltételes utasítást jelent.

Az egyértelműsítési hibák javítására szolgáló R feltételes utasításokra néhány példa:

```
x$upos <- ifelse(x$lemma=="antik", "ADJ", x$upos)
x$upos <- ifelse(x$lemma=="anyag", "NOUN", x$upos)
```

²² Universal POS tags, <https://universaldependencies.org/docs/u/pos/>, (2019-10-28)

²³ Universal Dep. Relations, <https://universaldependencies.org/u/dep/>, (2019-10-28)

```
x$upos <- ifelse(x$lemma == "atom", „NOUN”, x$upos)
```

A nyelvi modell az antik és az atom szavakat néha igének (VERB), az anyag szót néha melléknévnek (ADJ) értelmezte.

A szótár a jelenlegi állapotában 341 egyértelműsítési hibát javít, ez ugyanennyi feltételes utasítást jelent.

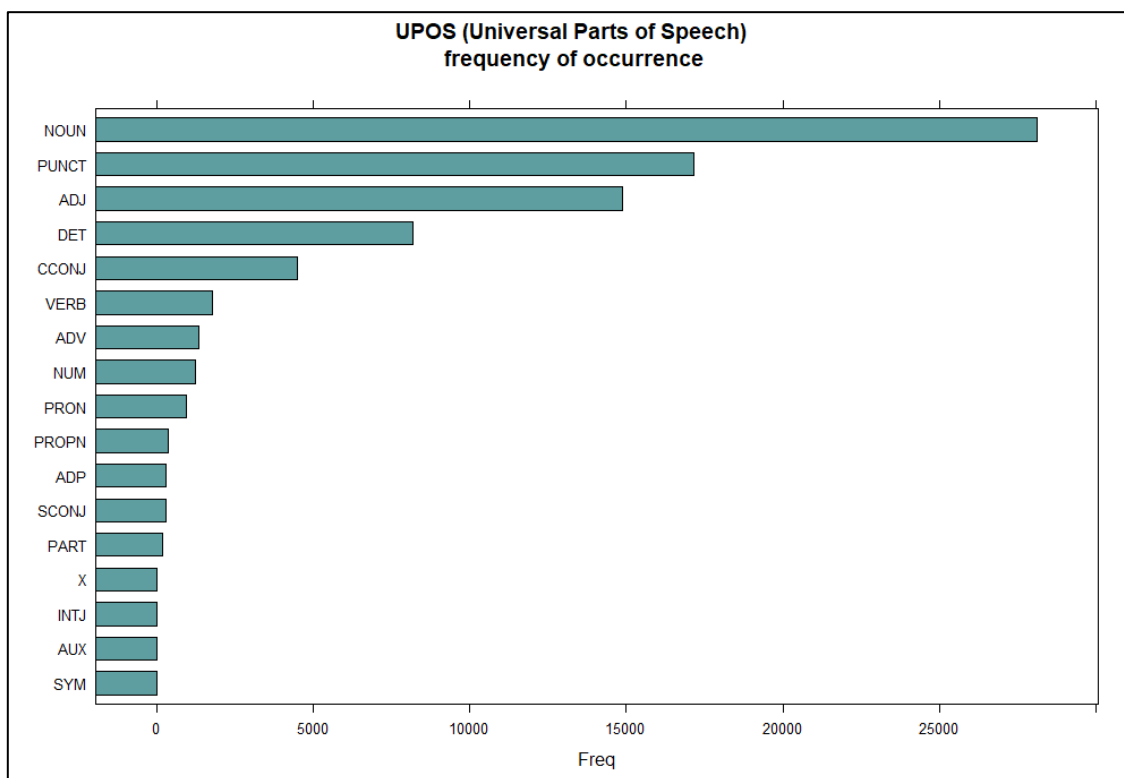
Az udpipe csomag az annotáció alapján (helyenként a textrank csomaggal együtt) az alábbi feladatok elvégzésére képes:

- Alapvető gyakorisági statisztikák készítése
- Kulcsszavak keresése
- Szöveg összegzés

7. Alapvető gyakorisági statisztikák készítése

A legtöbb nyelven a főnév (NOUN) és az ige (VERB) a leggyakoribb szófaj, ezek a legfontosabbak analitikai szempontból a melléknév (ADJ) és a tulajdonnév (PROPN) mellett. A szófaj címkék nagyon jól használhatók a szó diagramok készítéséhez. Nincs szükségünk stopszavakra, csak ki kell választanunk a főnevet, az igét vagy a melléknevet, és már el tudjuk végezni az alapvető gyakorisági elemzést.

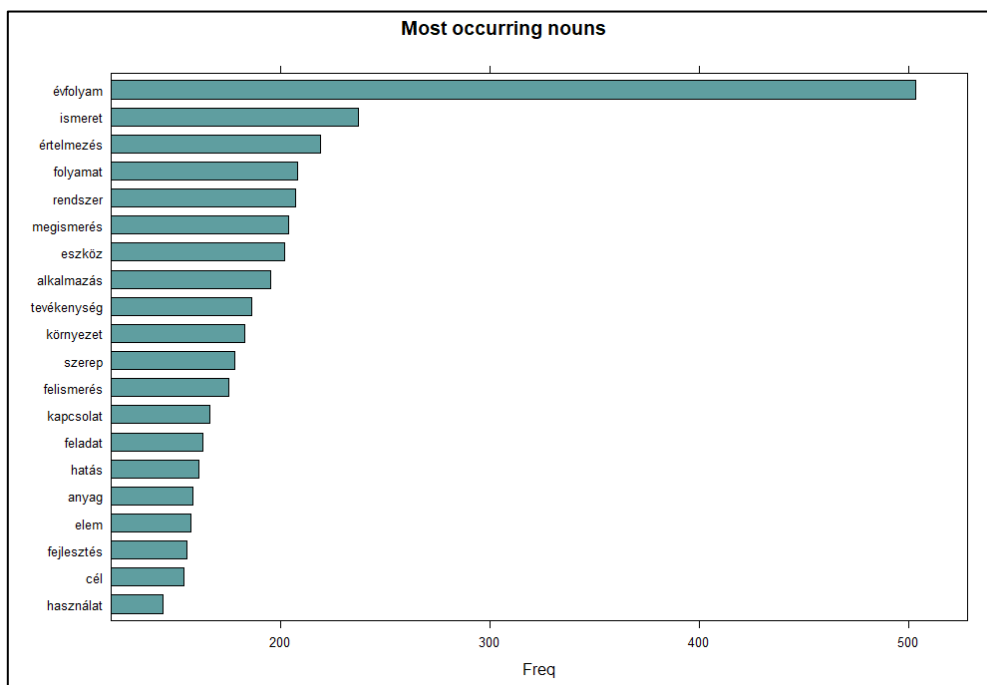
A szófajok előfordulási gyakoriságait a 2. ábra mutatja.



10. ábra

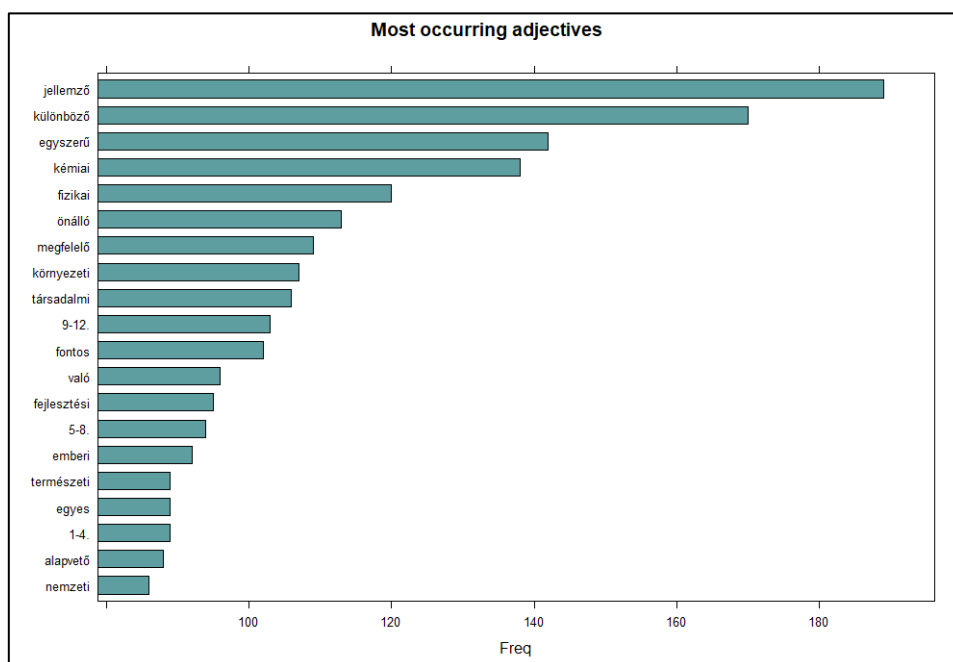
Láthatóan a korpuszban a leggyakoribb szófaj a főnév (NOUN), ezt követi a központosítás (PUNCT), majd a melléknév (ADJ). Az igék száma ezekhez képest alacsonynak mondható.

A leggyakoribb főneveket a 3. ábra mutatja.



11. ábra

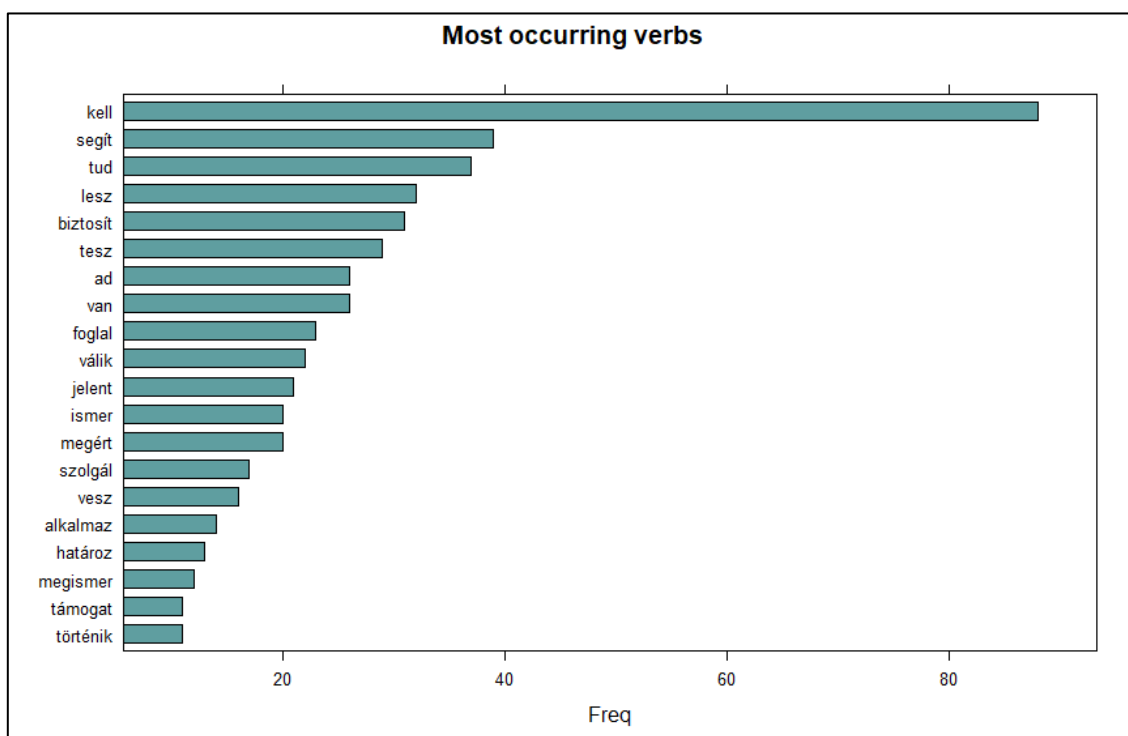
A leggyakoribb mellékneveket a 4. ábra mutatja.



12. ábra

A korpuszban a leggyakoribb főnév az évfolyam, ezt követi fele akkora gyakorisággal az ismeret. A korpusz leggyakoribb melléknevei a jellemző, a különböző és az egyszerű.

A leggyakoribb igéket az 5. ábra mutatja.



13. ábra

A korpusz leggyakoribb igéje a kell, ezt követi egyharmad akkora gyakorisággal a segít.

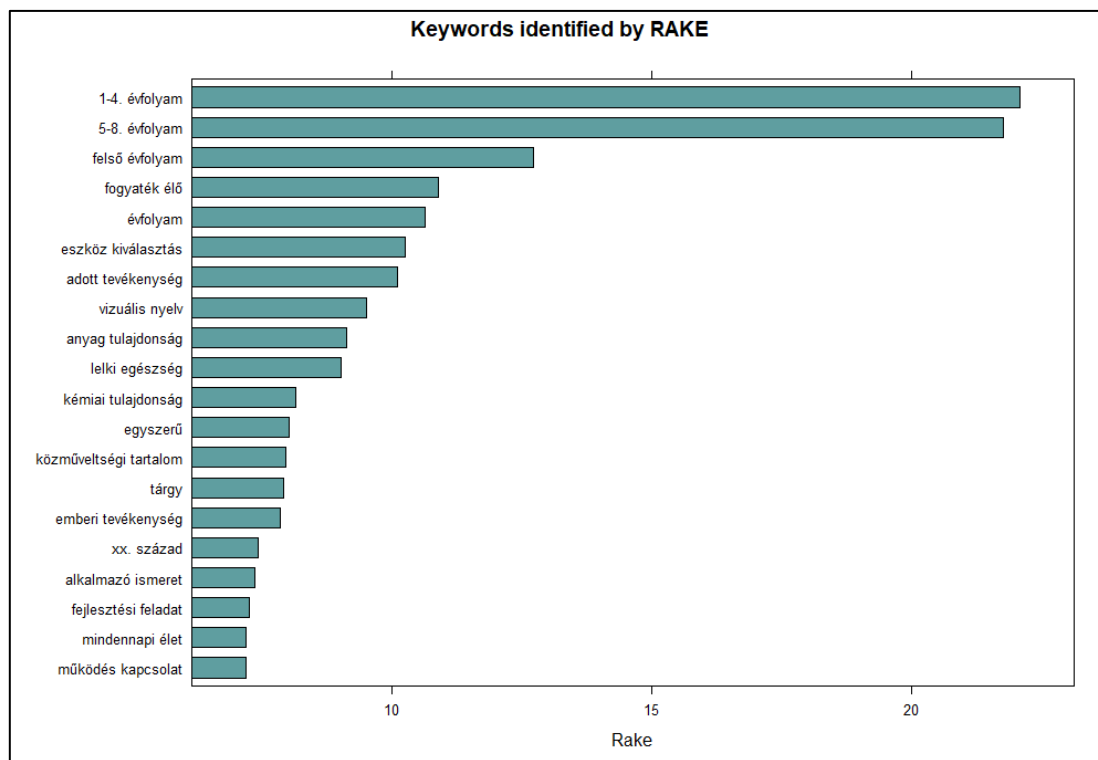
8. Kulcsszavak keresése

Az udpipe²⁴ csomag a szó gyakoriságok meghatározásán túl kulcsszavak keresésére is képes (helyenként a textrank csomaggal együtt). A legfontosabb kulcsszó keresési technikák a következők:

- RAKE (Rapid Automatic Keyword Extraction) algoritmus
- PMI (Pointwise Mutual Information) Collocation
- Egyszerű főneves kifejezések
- TextRank algoritmus

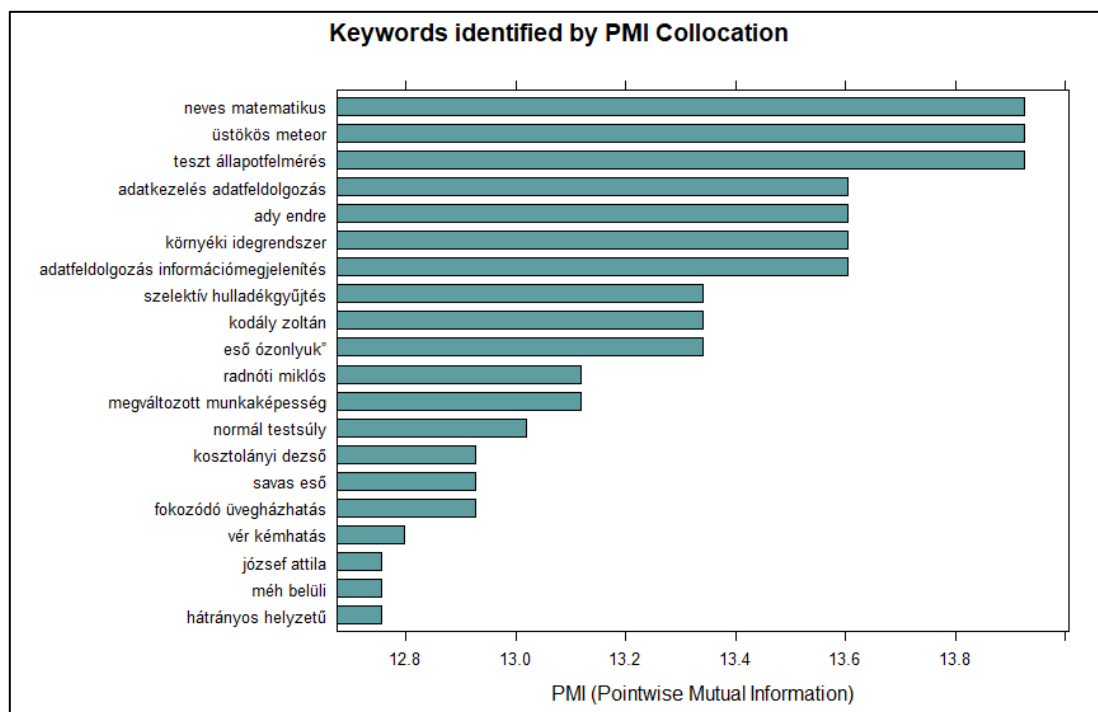
A RAKE algoritmus legmagasabb pontszámú kulcsszavait a 6. ábra mutatja.

²⁴ udpipe, <https://cran.r-project.org/package=udpipe>, (2019-10-28)



14. ábra

A PMI algoritmus legmagasabb pontszámú kulcsszavait a 7. ábra mutatja.

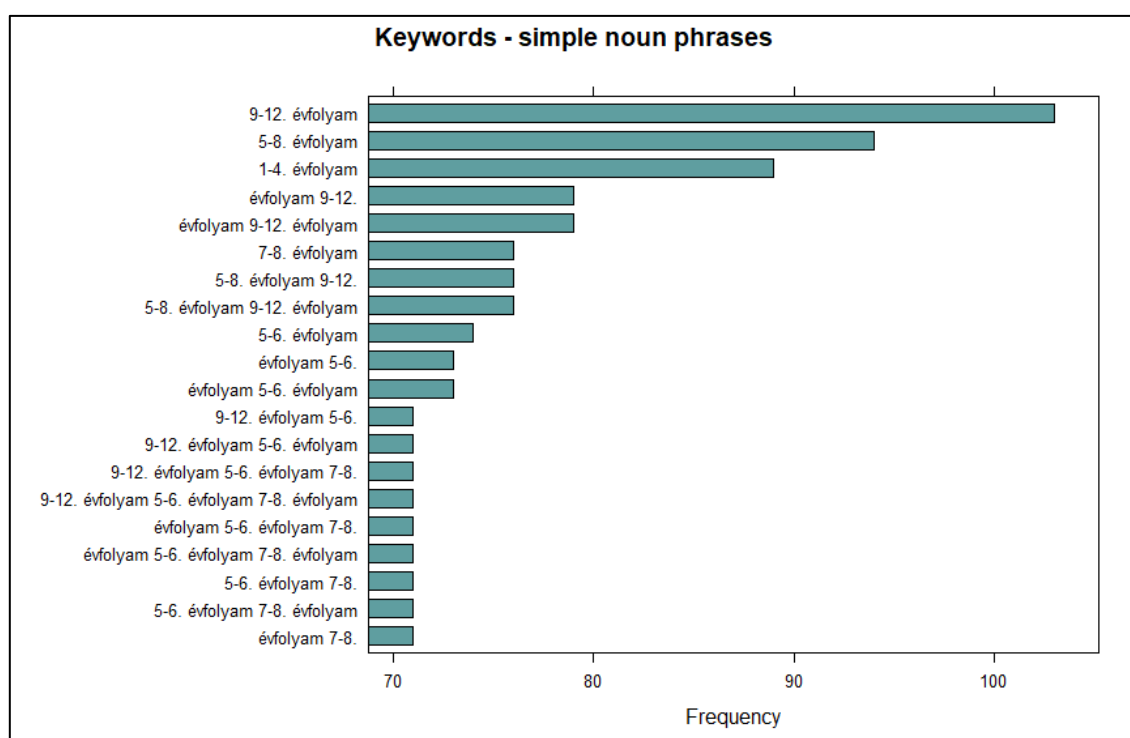


15. ábra

A következő lehetőség az egyszerű főneves kifejezések keresése. Az udpipe a szófajokat a reguláris kifejezésekben a következő betűjelekkel kódolja: A melléknév (ADJ), C mellérendelő kötőszó (CCONJ), D határozószó (DET), M az ige módosítója, N főnév vagy tulajdonnév, P prepozíció.

E betűjelekkel meghatározhatunk egy, a szófajok kívánt sorozatát leíró reguláris kifejezést: pl. (A|N)*(A|N)*. Itt a | az OR (vagy) művelet jele, a + a legalább egyszeri előfordulás jele, a * a nulla vagy többszöri előfordulás jele, a zárójelek a rész-kifejezések csoportosítására szolgálnak.

A leggyakoribb, legfeljebb négytagú egyszerű főneves kifejezéseket a 8. ábra mutatja.



16. ábra

Láthatóan a leggyakoribb egyszerű főneves kifejezések az évfolyamokkal kapcsolatosak. Ennek az a magyarázata, hogy a képzési szakaszok megnevezése sűrűn előfordul a Nat szövegében, és a tíz műveltségi terület táblázatainak fejlécében.

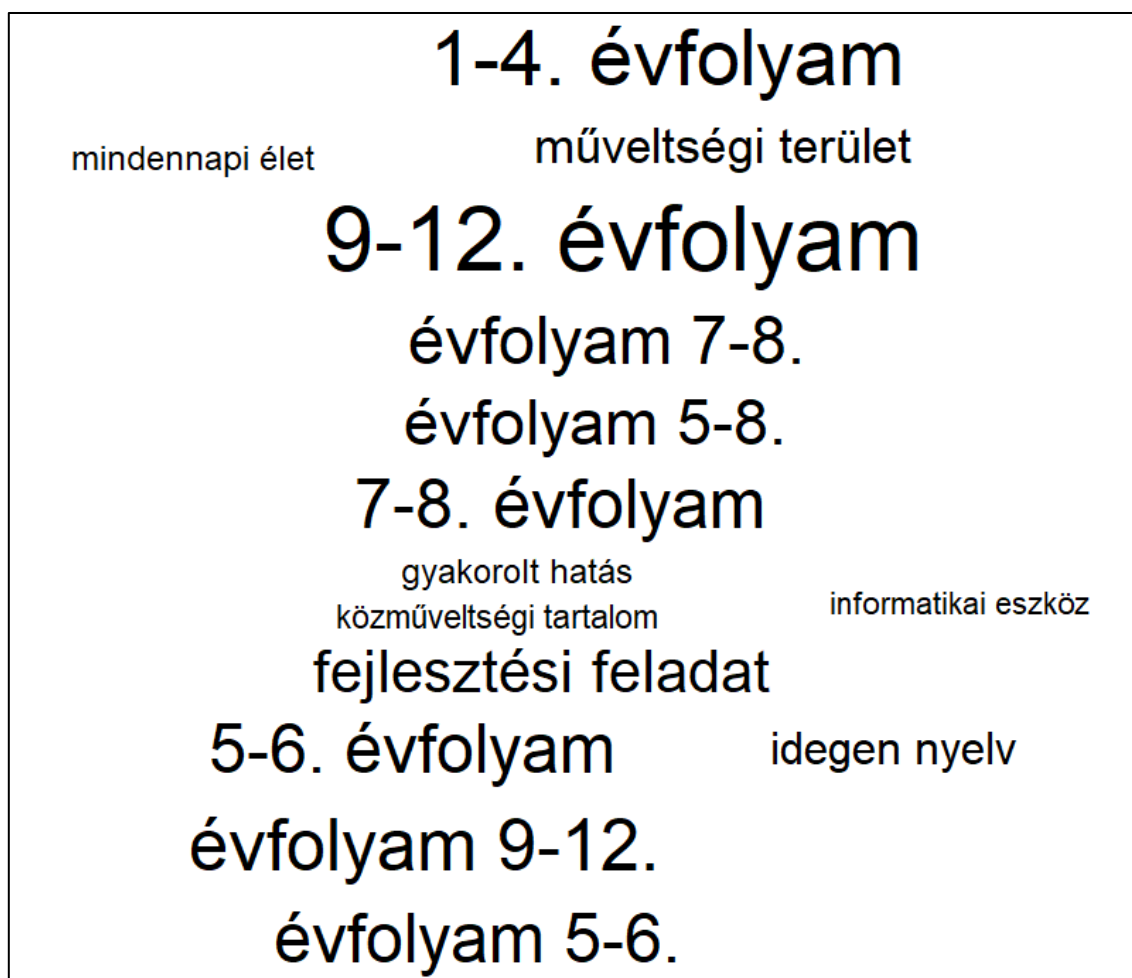
A TextRank algoritmust a textrank²⁵ R csomaggal használhatjuk. Az algoritmus kulcsszavak keresését és szövegek összegzését teszi lehetővé.

²⁵ textrank, <https://cran.r-project.org/package=textrank>, (2018-10-28)

A kulcsszavak keresése úgy történik, hogy az egymást követő szavakból felépít egy szó hálózatot, a TextRank algoritmussal meghatározza a releváns szavakat és az egymást követő releváns szavakból állítja elő a kulcsszavakat.

Esetünkben főnevekből vagy mellénevekből álló kulcsszavakat állítunk elő a TextRank algoritmussal.

A TextRank algoritmussal kapott leggyakoribb kulcsszavak szófelhőjét a 9. ábra mutatja. A szófelhő érdekessége, hogy az egyszerű szófelhőtől eltérően nem szavakat, hanem kifejezéseket ábrázol.



17. ábra

Várakozásunknak megfelelően a leggyakoribb kifejezések a különböző évfolyamokkal kapcsolatosak, de gyakori kifejezés a fejlesztési feladat, a műveltségi terület és az idegen nyelv is.

9. Szöveg összegzése (textrank)

A textrank R csomag lehetővé teszi egy dokumentum szövegének összegzését. Jól integrálódik az udpipe R csomaghoz, az udpipe annotáció kimenetét dolgozza fel.

A textrank csomag alkalmas egy dokumentum releváns mondatainak páros összehasonlítással való meghatározására a főnevekből és melléknévekből összeállított releváns kulcsszavak alapján.

A mondatpárok száma n elem k -ad osztályú kombinációjaként számítható ki.

$$C_n^k = \frac{n!}{k! (n - k)!}$$

Mivel itt mondatpárokról van szó, ezért $k=2$.

Nagyszámú mondatból álló dokumentum esetén a feladat számítási szempontból kezelhetetlenné válik.

A Nat mondatainak száma a kimeneti tábla mondat (sentence) oszlopában 4978.

$$C_{4978}^2 = \frac{4978!}{2! (4978 - 2)!}$$

A 4978 másodosztályú ismétlés nélküli kombinációja 12387753.

A mondatpár kombinációk száma a minhash algoritmussal vagy mintavétellel csökkenthető.

Alapértelmezés szerint a mondatokat a TextRank fontosság szerint adja meg. Alapbeállításaként az első 5 legfontosabb mondatot adja meg.

Megadhatjuk, hogy hány legfontosabb mondatot írjon ki a számítógép, vagyis megválaszthatjuk n értékét.

Kiírathatjuk fontossági sorrendben az n legfontosabb mondatot, de meg is tarthatjuk a mondatok eredeti sorrendjét a dokumentumban.

A mondatkombinációk számának csökkentése mintavétellel, max = 10000 beállítás mellett a következő négy releváns mondatot eredményezte:

[1] "rövidebb szövegek alkotása, személyes és olvasmányélmények megfogalmazása különböző szövegtípusokban és műfajokban."

[2] "szaporodás, egyedfejlődés, szexualitás a gyermekvárással, születéssel, szoptatással kapcsolatos elképzelések megbeszélése, az emberi termékenység befolyásoló folyamatok megbeszélése, a káros hatások elkerülésére alkalmas módszerek megismerése."

[3] "a nyelvhasználat társadalmi jelenségként való (szociolingvisztikai) szemlélete, néhány alapvető helyesírási szabály megismerése és

alkalmazása (a mondathatárok jelölése, nevek írása, az egyszerű szavak elválasztásának szabályai)."

[4] "a hazai társadalmi-gazdasági élet földrajzi jellegzetességeinek összefüggéseikben való felismerése a mindennapi élet eseményeiben, folyamataiban."

Láthatóan értelmes mondatokat kaptunk. A mondatokat körülvevő idézőjelek az angol billentyűzet szerintiek, hiszen ezeket a mondatokat a textrank csomag állítja elő.

Tapasztalataink szerint egyszerű szövegeket könnyebb feldolgozni, mint félig strukturált, címeket, alcímeket, listákat, decimális számozást és táblázatokat tartalmazó szövegeket. A félig strukturált szövegeknél nem egyszerű feladat a mondathatárok felismerése. A mondatok helyes azonosításától függ ugyanis a szöveg összegzésének eredményessége.

Irodalomjegyzék

110/2012. (VI. 4.) Korm. rendelet a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról,
<https://net.jogtar.hu/jogszabaly?docid=a1200110.kor>, (2019-10-28)

Anandarajan, M. - Hill, C. - Nolan, T. (2019): Practical Text Analytics, Springer, Cham

Aptitude, Wikipedia, <https://en.wikipedia.org/wiki/Aptitude>, (2019-10-28)

Képesség, Nat, <https://net.jogtar.hu/jogszabaly?docid=a1200110.kor>, (2019-10-28)

Jockers, M. L. (2014): Text Analysis with R for Students of Literature, Springer, Cham

Kwartler, T. (2017): Text Mining in Practice with R, Wiley, Hoboken, NJ

textrank, <https://cran.r-project.org/package=textrank>, (2018-10-28)

UD 2.4 Models for UDPipe,
<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2998>, (2019-10-28)

udpipe, <https://cran.r-project.org/package=udpipe>, (2019-10-28)

Universal POS tags, <https://universaldependencies.org/docs/u/pos/>, (2019-10-28)

Universal Dep. Relations, <https://universaldependencies.org/u/dep/>, (2019-10-28)

Wachsmuth, H. (2015): Text Analysis Pipelines, Springer, Cham