

Používateľská príručka

Obsah používateľskej príručky

[Úvod](#)

[Domovská obrazovka](#)

[Všeobecné textové informácie](#)

[Graf korpusov a početnosti ich článkov](#)

[Graf rozloženia veľkosti korpusov](#)

[Tabuľka najrelevantnejších slov podľa korpusov](#)

[Živá analýza](#)

[Záložka Text](#)

[Záložka Odtlačok](#)

[Záložka Porovnanie článku](#)

[Články](#)

[Zoznam článkov](#)

[Detail článku](#)

[Záložka Index](#)

[Záložka Histogram](#)

[Záložka Odtlačok článku](#)

[Záložka Porovnanie článku](#)

[Invertovaný index](#)

[Hra s vetami](#)

Úvod

Táto používateľská príručka má za cieľ priblížiť a objasniť jednotlivé funkcionality nástroja na inteligentnú analýzu slovenských textov s názvom TxtEnv. Tento nástroj je vyvíjaný v rámci predmetu Tímový projekt na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave a vyvíja ho tím číslo 5 - TextMania.

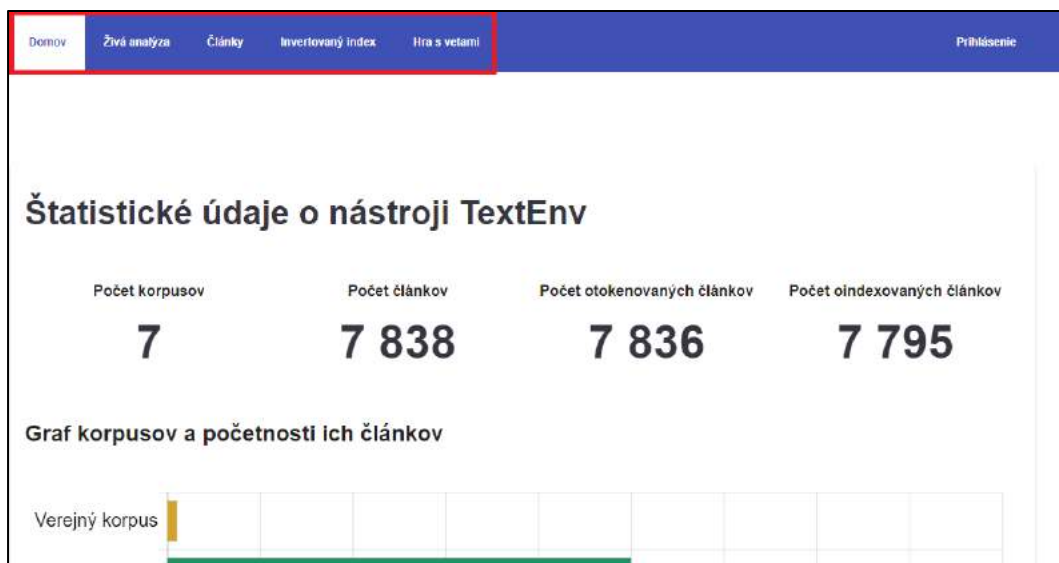
Adresa, na ktorej je spomínaný nástroj dostupný, je <http://bit.ly/2U6YFYa>.

Štruktúra používateľskej príručky je rozdelená do častí podľa jednotlivých častí (modulov) nástroja.

Jednotlivé časti (moduly) aplikácie sú sprístupnené pomocou navigačného menu umiestneného na vrchnej časti stránky a do žiadanej časti aplikácie sa možno dostať jednoduchým kliknutím na názov časti.

Dostupné časti aplikácie sú:

- [Domovská obrazovka](#)
- [Živá analýza](#)
- [Články](#)
- [Invertovaný index](#)
- [Hra s vetami](#)



Domovská obrazovka

Domovská obrazovka sa zobrazí ako hlavná obrazovka pri spustení webovej aplikácie a zobrazuje štatistické údaje o nástroji.

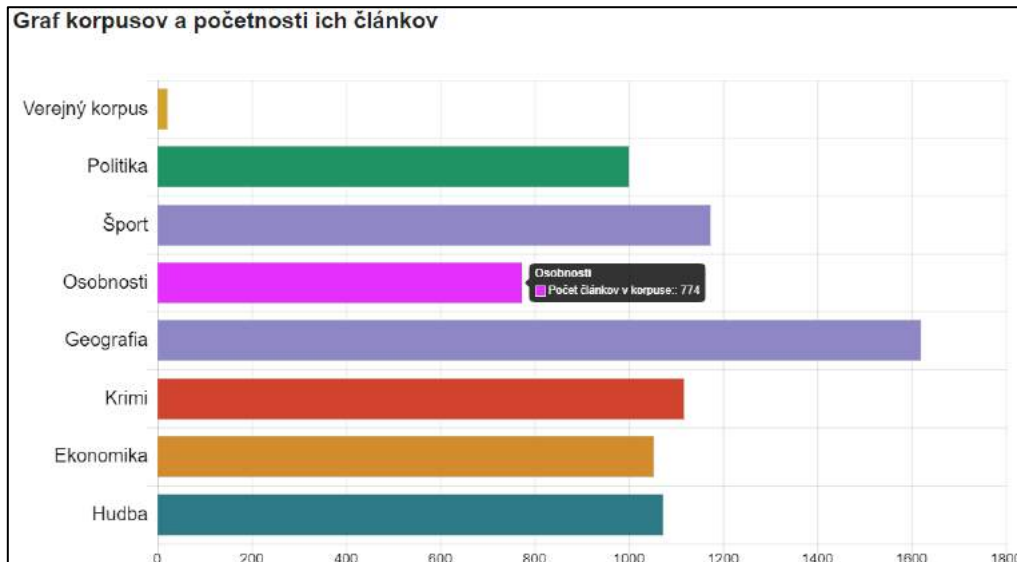
Všeobecné textové informácie

Na vrchnej časti obrazovky sú textovou formou zobrazené nasledovné údaje:

- Počet dostupných korpusov s článkami
- Celkový počet dostupných článkov
- Počet otokenovaných článkov
- Počet oindexovaných článkov

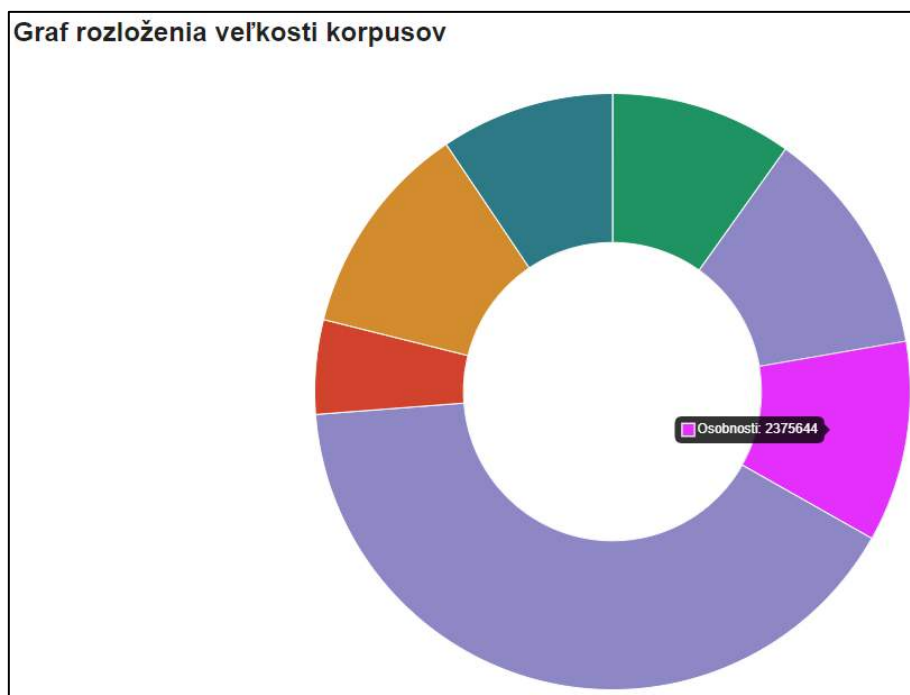
Graf korpusev a početnosti ich článkov

Pod textovými informáciami o počte dostupných korpusev a článkov sa nachádza stĺpcový graf znázorňujúci počty článkov v jednotlivých korpusech. Pri ukázaní myšou na jednotlivé korpuse sa v textovej forme zobrazí informácia o presnom počte článkov v danom korpuse.



Graf rozloženia veľkosti korpusev

Ďalším zobrazeným grafom je koláčový graf, ktorý vyjadruje veľkosť jednotlivých korpusev s ohľadom na celkový počet znakov, ktoré sa v korpuse nachádzajú. Vyjadruje teda súčet dĺžok jednotlivých článkov v danom korpuse s ohľadom na počet znakov. Podobne, ako v predchádzajúcom grafe, aj v tomto grafe je možné zobrazit' presnejšiu informáciu o počte znakov ukázaním na daný korpus pomocou myši.



Tabuľka najrelevantnejších slov podľa korpusov

V spodnej časti domovskej obrazovky je vyobrazená tabuľka s najrelevantnejšími slovami vo všetkých dostupných korpusoch. Riadky v spomínanej tabuľke prislúchajú k jednotlivým korpusom a stĺpce reprezentujú poradie relevantnosti slov (prvý stĺpec zobrazuje najčastejšie vyskytujúce sa slovo v korpuse atď.). Nad tabuľkou sa nachádza prepínač, pomocou ktorého je možné si z tabuľky odfiltrovať tzv. stopslová (*stopwords*) tak, aby sa v tabuľke zobrazovali čo najrelevantnejšie výrazy. Kliknutím na názov korpusu v tabuľke je možné presmerovať sa na [invertovaný index](#) slov pre daný korpus.

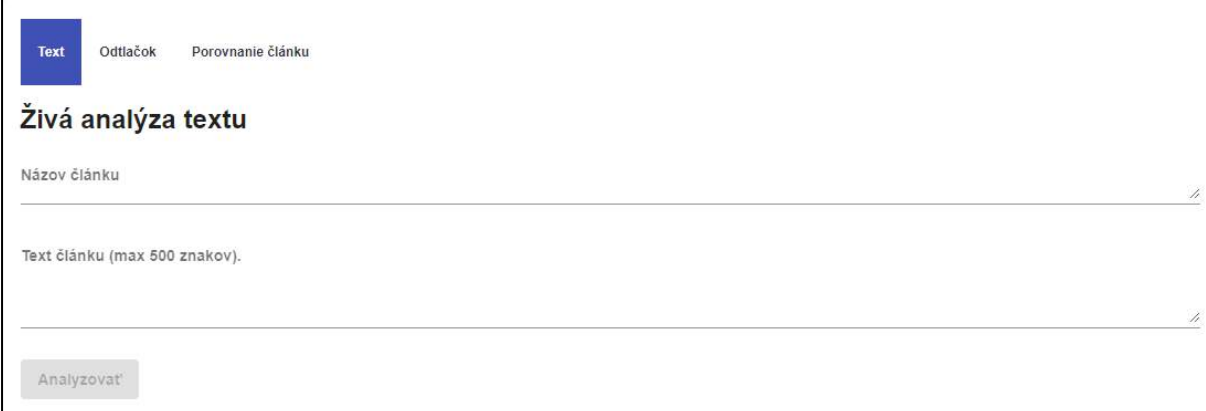
Živá analýza

Časť aplikácie s názvom “Živá analýza” ponúka používateľovi možnosť analyzovať ľubovoľný vlastný text, zobrazí jeho tokeny, porovnať ho s článkami v ostatnom korpuse a tým ho percentuálne zaradiť do jednotlivých korpusov. Samotná časť so živou analýzou obsahuje 3 záložky, v ktorých sa možno navigovať obdobným spôsobom ako v hlavnom navigačnom menu aplikácie vo vrchnej časti obrazovky:

- [Text](#)
- [Odtlačok](#)
- [Porovnanie článku](#)

Záložka Text

V tejto záložke sa nachádzajú dve textové polia. Menšie textové pole slúži pre zadanie nadpisu analyzovaného článku a väčšie textové pole slúži na samotný text článku. Dĺžka textu článku je obmedzená, a to na dĺžku 500 znakov. Na spustenie analýzy článku zadaného používateľom je potrebné, aby obe zmienené polia obsahovali nejaký text a neboli prázdne.



The screenshot shows the 'Text' tab selected in the 'Živá analýza textu' application. The interface includes three tabs: 'Text', 'Odtlačok', and 'Porovnanie článku'. Below the tabs, the title 'Živá analýza textu' is displayed. There are two input fields: 'Názov článku' and 'Text článku (max 500 znakov)'. At the bottom left, there is a button labeled 'Analyzovať'.

Po vyplnení oboch textových polí je umožnené používateľovi kliknúť na tlačidlo s textom “Analyzovať” a tým spustiť analýzu vloženého článku. Po úspešnej analýze článku sa pod dvoma textovými poliami zobrazia tokeny článku. Používateľ má možnosť ukázať myšou na jednotlivé tokeny, čím sa mu zobrazia dodatočné informácie o danom tokene (tagy, POS, léma, NER atď.). Pod tokenmi článku sú zobrazené percentuálne podobnosti článku s dostupnými korpusmi, resp. pravdepodobnosti príslušnosti analyzovaného článku do korpusov.

Text Odtlačok Porovnanie článku

Živá analýza textu

Názov článku
Článok aktuality

Text článku (max 500 znakov).
Čo to môže znamenať? Napríklad to, že pokiaľ Mario Hoffmann v tomto prípade mohol pokojne spávať, tak sa to môže zmeniť. Aj keď je tam silný moment – „špeciálny zametač Kováčik“. Táto nelichotivá prezývka sprevádza špeciálneho prokurátora Kováčika dlhšiu dobu.
Stručný zoznam káuz, ktoré skončili štýlom „skutok sa nestal“, nájdete na konci komentára.

Analyzovať

Lemma: spávať
POS: Vie

Tokeny textu

Čo to môže znamenať? Napríklad to, že pokiaľ Mario Hoffmann v tomto prípade mohol pokojne spávať, tak sa to môže zmeniť. Aj keď je tam silný moment – „špeciálny zametač Kováčik“. Táto nelichotivá prezývka sprevádza špeciálneho prokurátora Kováčika dlhšiu dobu. Stručný zoznam káuz, ktoré skončili štýlom „skutok sa nestal“, nájdete na konci komentára.

Category	Percentage
Osobnosti	66 %
Hudba	58 %
Krimi	44 %
Ekonomika	59 %
Šport	62 %
Geografia	58 %
Politika	60 %

Záložka Odtlačok

V záložke “Odtlačok” sa zobrazuje tzv. odtlačok článku a nachádzajú sa tu celkovo 3 tabuľky, ktoré budú opísané v takom poradí, v akom sa zobrazujú na obrazovke (v smere zvrchu nadol). Prvá tabuľka zobrazuje celkový počet tokenov v analyzovanom článku spolu s počtom plnovýznamových a počtom neplnovýznamových slov.

Text **Odtlačok** Porovnanie článku

Odtlačok článku

Počet tokenov	66
Počet plnovýznamových slov	39
Počet neplnovýznamových slov	9

Druhá tabuľka reprezentuje histogram POS (part of speech) pre jednotlivé tokeny. V prvom stĺpci tabuľky sa zobrazuje slovný druh, v druhom stĺpci počet slov s daným slovným druhom a v treťom stĺpci sa zobrazuje percentuálne zastúpenie daného slovného druhu v článku.

POS histogram		
Slovný druh	Počet	%
neidentifikovane	2	3.03
Spojka O	6	9.09
Zámená P	7	10.61
Slovesá V	11	16.67
Interpunkcia Z	14	21.21
Častice T	1	1.52
Predložky E	2	3.03
Podstatné mená S	14	21.21
Príslovky D	1	1.52
R	2	3.03
Prídavné mená A	6	9.09

Tretia tabuľka má obdobný výzor ako predchádzajúca tabuľka zobrazujúca slovné druhy. Rozdielom však je, že táto tabuľka zobrazuje počty jednotlivých typov identifikovaných pomenovaných entít v texte. V prvom stĺpci tabuľky sa nachádza typ identifikovanej pomenovanej entity a nasledujúce dva stĺpce sú rovnaké ako v predchádzajúcej tabuľke.

Záložka Porovnanie článku

Záložka Porovnanie článku slúži na porovnanie článku s dostupnými korpusmi pri iničiálnom zobrazení obsahuje stĺpce so slovom v článku, jeho lemov, hodnotou TF-IDF a poradovým číslom slova podľa hodnoty TF-IDF.

Text Odtlačok **Porovnanie článku**

Porovnanie textu

Ekonomika
 Geografia
 Hudba

Krimi
 Osobnosti
 Politika

Šport

Článok			
Slovo	Lema	TF-IDF	Poradie
.		60.61	1
.		60.61	2
to	to	45.45	3
môže	môct	30.30	4
sa	sa	30.30	5
"		30.30	6
"		30.30	7
čo	čo	15.15	8
znamenaf	znamenaf	15.15	9
?		15.15	10
napríklad	napríklad	15.15	11
že	že	15.15	12
pokiaľ	pokiaľ	15.15	13
mario		15.15	14
hoffmann		15.15	15
v	v	15.15	16
tomto	tento	15.15	17
pripade	pripad	15.15	18
mohol	môct	15.15	19

Nad tabuľkou sa nachádzajú zaškrtačacie políčka s názvami korpusov. Pri zaškrtnutí korpusu sa do tabuľky pridajú dva stĺpce pre každý zaškrtnutý korpus - hodnota TF-IDF daného slova v príslušnou korpusu spolu s jeho poradím na základe hodnoty TF-IDF vrámci zvoleného korpusu.

Text

Odtlačok

Porovnanie článku

Porovnanie textu

 Ekonomika

 Geografia

 Hudba

 Krimi

 Osobnosti

 Politika

 Šport

Článok				Geografia		Osobnosti		Politika	
Slovo	Lema	TF-IDF	Poradie	TF-IDF	Poradie	TF-IDF	Poradie	TF-IDF	Poradie
,		60.61	1	1.46	19	0.38	705	12.37	2
.		60.61	2	0.51	239	0.42	610	4.66	5
to	to	45.45	3	0.82	95	1.04	78	2.45	26
môže	môcť	30.30	4	0.22	1128	0.17	1295	1.25	162
sa	sa	30.30	5	3.55	5	neexistuje		2.22	36
„		30.30	6	1.14	43	2.83	7	2.26	33
“		30.30	7	1.12	46	2.82	8	2.13	41
čo	čo	15.15	8	0.87	81	0.90	114	1.52	119
znamenáť	znamenat'	15.15	9	0.02	3099	0.03	1431	0.11	1343
?		15.15	10	0.24	1000	0.62	290	0.71	383
napríklad	například	15.15	11	0.53	215	0.43	594	0.93	251
že	že	15.15	12	1.05	54	1.78	21	2.99	17
pokiaľ	pokiaľ	15.15	13	0.07	2883	0.08	1411	0.67	418
mario		15.15	14	0.02		0.06	1419	neexistuje	
hoffmann		15.15	15	0.00		neexistuje		neexistuje	
v	v	15.15	16	0.76	109	1.29	50	3.19	12
tomto	tento	15.15	17	0.48	271	0.39	689	0.46	714
prípade	prípad	15.15	18	0.15	1897	0.10	1400	1.36	143

Pod tabuľkou sa taktiež zobrazia percentuálne podobnosti analyzovaného textu so zvolenými korpunami. Tieto percentuálne podobnosti majú rovnaký význam ako v časti [Záložka Text](#).

stylom	styl	15.15	48	0.05		0.06	1421	neexistuje	
skutok	skutok	15.15	49	neexistuje		neexistuje		0.03	1365
nestal	nestať	15.15	50	neexistuje		neexistuje		0.02	1366
nájdete	nájsť	15.15	51	0.05	2996	0.02	1432	0.02	1366
na	na	15.15	52	0.58	176	1.56	33	1.87	71
konci	koniec	15.15	53	0.33	545	0.18	1274	0.16	1304
komentára	komentár	15.15	54	neexistuje		neexistuje		neexistuje	



Články

Modul aplikácie s názvom Články možno rozdeliť do dvoch pomyselných častí, ktorými sú [Zoznam článkov](#) a [Detail článku](#), ktorý obsahuje niekoľko záložiek. Po kliknutí na záložku “Články” v hornom navigačnom menu sa používateľovi zobrazí už spomínaný zoznam článkov.

Zoznam článkov

V tejto časti sa nachádza zoznam všetkých dostupných článkov v systéme. Každý riadok v zozname zodpovedá jednému článku a nachádzajú sa v ňom nadpis článku a korpus, v ktorom sa článok nachádza. Na konci riadku sa nachádza tlačidlo “Detail”, ktoré umožňuje presmerovanie na [Detail](#) zvoleného článku.

Zoznam článkov je možné abecedne zoradiť, a to kliknutím na hlavičku príslušného stĺpca v tabuľke, podľa ktorého chceme zoznam zoradiť. Opakovaným kliknutím je možné prepínať medzi zoradením vzostupne a zostupne.

Nad zoznamom sa nachádzajú dve filtračné polia:

- **Názov** - Textové pole pre filtrovanie článkov podľa nadpisu. Na filtrovanie podľa názvu nie je potrebné zadať celý nadpis článku, keďže filter filtruje podľa začiatkových znakov (napr. pri vložení písmena “A” sa zobrazia články, ktorých názov sa začína na písmeno “A”).
Je potrebné uviesť, že filter je citlivý na veľkosť písmen (tzv. case-sensitive), to znamená že rozlišuje malé a veľké písmená v názvoch.
- **Korpus** - filtrovať na základe korpusu je možné rozkliknutím rozbaľovacieho menu a následným výberom žiadaného korpusu. Alternatívne je možné kliknúť na názov korpusu v akomkoľvek riadku z tabuľky a tým filter nastaviť na daný korpus.

Filtrovanie zoznamu sa spustí kliknutím na tlačidlo “Filtrovať”, alebo stlačením klávesy Enter.

Názov	Korpus	Detail
Agentúra Moody's zhoršila rating automobilky Jaguar Land Rover a jeho výhľad je negatívny.	Ekonomika	Detail
Agentúry potvrdili Slovensku rating A+ so stabilným výhľadom, má solídny hospodársky rast.	Ekonomika	Detail
Agrárna komora žiada ministerku Matečnú, aby v Rade SPF bol aj aktívny farmár.	Ekonomika	Detail

Detail článku

Záložka Index

V záložke Index sa vo vrchnej časti zobrazuje nadpis vybraného článku. Pod nadpisom sa zobrazujú informácie o dátume vytvorenia, anotovania a indexovania článku. Na pravej strane pod nadpisom sa nachádza odkaz na [Invertovaný index](#) slov pre tento článok. Pod nadpisom a dátumami sa nachádza samotný text vybraného článku.

Index Histogram Odtlačok článku Porovnanie článku

Agentúra Moody's zhoršila rating automobilky Jaguar Land Rover a jeho výhľad je negatívny.

Dátum vytvorenia: 11.12.2018 Dátum anotovania: 11.12.2018 Dátum indexovania: 12.03.2019 [Odkaz na invertovaný index článku](#)

Medzinárodná ratingová agentúra Moody's zhoršila automobilke Jaguar Land Rover Automotive Plc (JLR) takzvaný Corporate Family Rating (CFR) na stupeň Ba3 z úrovne Ba2 . Výhľad ratingu je negatívny . Pokračujúci slabý prevádzkový výkon . . . Zníženie na stupeň Ba3 odzrkadľuje pokračujúci slabý prevádzkový výkon JLR v prvom a druhom kvartáli účtovného roka 2019 , ktorý sa skončí v marci , " uviedol Falk Frey , hlavný analytik agentúry Moody's pre Jaguar Land Rover . Agentúra Moody's s pozitívne hodnotí to , že automobilka JLR ohlásila plán znižovania nákladov a efektívnosti , ktorý by v nasledujúcich osemnástich mesiacoch mal viesť k zlepšeniu financií firmy o 2,5 mld . libier (GBP) . Zvýšené riziká na trhu . Moody's však dodáva , že perspektíva rýchleho obratu je podľa nej problematická vzhľadom na zvýšené riziká na trhu , vrátane neistôt súvisiacich s brexitom a s tým spojenými nákladmi , ako aj vzhľadom na oslabujúci dopyt na automobilovom trhu v Číne , rastúce náklady pre vyššie ceny surovín a zvyšujúce sa ceny pohonných látok . Automobilka Jaguar Land Rover v októbri otvorila svoj závod v Nitre . Jeho ročná produkcia bude 150000 áut ročne . Investícia dosiahla 1,4 miliardy eur . (1 EUR = 0,86945 GBP)

Po ukázaní myšou na ľubovoľné slovo sa nad ním zobrazia dodatočné informácie, akými sú POS tagy, léma, NER atď (ak sú tieto informácie dostupné).

Medzinárodná ratingová agentúra Moody's zhoršila automobilke Jaguar Land Rover Automotive Plc (JLR) takzvaný Corporate Family Rating (CFR) na stupeň Ba3 z úrovne Ba2 . Výhľad ratingu je negatívny . Pokračujúci slabý prevádzkový výkon . . . Zníženie na stupeň Ba3 odzrkadľuje pokračujúci slabý prevádzkový výkon JLR v prvom a druhom kvartáli účtovného roka 2019 , ktorý sa skončí v marci , " uviedol Falk Frey , hlavný analytik agentúry Moody's pre Jaguar Land Rover . Agentúra Moody's s pozitívne hodnotí to , že automobilka JLR ohlásila plán znižovania nákladov a efektívnosti , ktorý by v nasledujúcich osemnástich mesiacoch mal viesť k zlepšeniu financií firmy o 2,5 mld . libier (GBP) . Zvýšené riziká na trhu . Moody's však dodáva , že perspektíva rýchleho obratu je podľa nej problematická vzhľadom na zvýšené riziká na trhu , vrátane neistôt súvisiacich s brexitom a s tým spojenými nákladmi , ako aj vzhľadom na oslabujúci dopyt na automobilovom trhu v Číne , rastúce náklady pre vyššie ceny surovín a zvyšujúce sa ceny pohonných látok . Automobilka Jaguar Land Rover v októbri otvorila svoj závod v Nitre . Jeho ročná produkcia bude 150000 áut ročne . Investícia dosiahla 1,4 miliardy eur . (1 EUR = 0,86945 GBP)

Lemma: skončiť
POS: VKdsc

Záložka Histogram

V záložke histogram má možnosť vidieť tabuľku početností jednotlivých POS tagov všetkých slov v článku. Tabuľka obsahuje iba dva stĺpce, a to konkrétny POS tag a počet, koľkokrát sa určitý POS tag v článku vyskytuje.

Index	Histogram	Odtlačok článku	Porovnanie článku
Histogram			
POS			Počet
--Neznáme--			28
0			6
AAfp4y			1
AAfs1x			4
AAip7x			1
AAis1x			1
AAis2x			1
AAis4x			5
AAis6x			1
AAmp2x			1
AAms2x			1

Záložka Odtlačok článku

Záložka Odtlačok článku v článkovom module funguje rovnako ako [Záložka Odtlačok v časti Živá analýza](#).

Záložka Porovnanie článku

Záložka Porovnanie článku v článkovom module funguje rovnako ako [Záložka Porovnanie článku v časti Živá analýza](#).

Invertovaný index

Obrazovka Invertovaný index zobrazuje tabuľku so všetkými dostupnými relevantnými informáciami o slovách, akými sú:

- Léma
- Korpus/Článok, v ktorom sa slovo nachádza - *zobrazenie závislé od zvoleného nastavenia rozhl'adu*
- NER
- POS
- Hodnota TF (term frequency)
- Celok/Korpus/Článok TF-IDF - *zobrazenie závislé od zvoleného nastavenia rozhl'adu*
- Poradie - poradie slova na základe hodnoty TF-IDF - *zobrazenie závislé od zvoleného nastavenia rozhl'adu*
- Odkaz na článok, v ktorom sa slovo nachádza - *zobrazenie závislé od zvoleného nastavenia rozhl'adu*

Záznamy v tabuľke je možné filtrovať pomocou filtračných polí umiestnených nad tabuľkou. Patria medzi ne:

- **Rozhľad** - rozhoduje o tom, podľa akého kritéria sú záznamy v tabuľke zoskupené, pričom používateľ má v rozbaľovacom menu na výber 3 možnosti: článok, korpus alebo celý dataset. Nastavenie filtra tiež ovplyvňuje zobrazenie relevantných stĺpcov v tabuľke.
- **Léma** - textové pole slúžiace na filtrovanie na základe lémy slova, pri čom je potrebné zadať celú lému slova (nie iba časť)
Tento filter je možné nastaviť aj kliknutím na lému jedného zo záznamov v tabuľke.
- **Slovo** - textové pole slúžiace na filtrovanie na základe slova, pri čom je potrebné zadať celé slovo (nie iba časť)
- **Článok** - textové pole slúžiace na filtrovanie na základe nadpisu článku. V tomto filtri je postačujúce zadať iba začiatok nadpisu a filter následne zobrazí záznamy pre všetky články, ktoré sa začínajú zadaným textovým reťazcom.
Tento filter je možné nastaviť aj kliknutím na nadpis článku jedného zo záznamov v tabuľke.
- **Korpus** - rozbaľovacie menu slúžiace na filtrovanie podľa korpusu
- **POS** - textové pole slúžiace na filtrovanie podľa POS tagu slova. V tomto filtri je postačujúce zadať iba začiatok POS tagu a filter následne zobrazí všetky slová, ktorých hodnota POS sa začína zadaným textovým reťazcom.
- **NER** - textové pole slúžiace na filtrovanie podľa NER daného slova, pričom je potrebné zadať celé a presné znenie hodnoty NER (nie iba časť slova)

Filtrovanie zoznamu sa spustí kliknutím na tlačidlo "Filtrovať", alebo stlačením klávesy Enter.

Slovo	Článok	Léma	Poradie	Článok Tfidf	Tf	POS	NER	Odkaz
obce	Abovce	obec	110	2.327	0.004	SSfS2		Detail
11.6	Abovce	11.6	6	25.1687	0.004	0	VALUE	Detail
kotliny	Abovce	kotlina	56	11.564	0.004	SSfS2		Detail
roku	Abovce	rok	10	22.6438	0.024	SSfS6		Detail
.	Abovce		121	1.2108	0.044	Z		Detail
leži	Abovce	ležiť	103	3.1258	0.004	VfKesc		Detail
maďarskej	Abovce	maďarský	52	11.7526	0.004	AAfS2x		Detail

Hra s vetami

Aplikácia obsahuje aj interaktívnu hru, v ktorej má používateľ za úlohu identifikovať do ktorého z korpusov patrí zobrazená veta. Veta je vybraná náhodne zo všetkých článkov. Hra prebieha tak, že používateľ začína s iniciálnym rozpočtom 100 virtuálnych peňazí. Z jeho rozpočtu môže následne rozdeliť medzi korpusy jednotlivé čiastky podľa svojho uváženia tak, aby trafil korpus, v ktorom sa zobrazená veta nachádza. V prípade úspešného tipu sa používateľovi prirába čiastka do rozpočtu a zobrazí sa nová veta. Pri neúspešnom pokuse sa naopak čiastka odráta. Hra sa končí, keď hráč už nemá dostupné žiadne prostriedky.



Na vrchnej časti pomyselného hracieho stola je zobrazený aktuálny rozpočet hráča a pod ním je zobrazená aktuálna veta. Čiastky môže používateľ prerozdeľovať pomocou znakov plus (+) a mínus (-), ktoré sú umiestnené pod názvom jednotlivých dostupných korpusov. Svoj tip a stávkou používateľ potvrdí stlačením tlačidla s textom "Vyhodnoť !"