Jak se v médiích referuje o koronaviru: wordcloudy nejčastěji se vyskytujících slov ve zpravodajství a publicistice o onemocnění COVID-19 (1. 12. 2019 – 31. 10. 2020)

Alice Němcová Tejkalová, Radek Mařík, Václav Moravec, Veronika Macková, Victoria Nainová

Úvod

V projektu COVID-19 infodemie chceme za pomoci strojového učení a umělé inteligence vytvořit platformu, která by měla pomoci snížit míru infodemie, tedy nadměrného šíření zavádějících informací, které činí pandemie a epidemie nepřehlednými a znesnadňují dosažení správného řešení. Pro začátek jsme se rozhodli s pomocí nástrojů strojového učení zmapovat výskyt slov ve zpravodajství o onemocnění COVID-19, obecněji pak o novém typu koronaviru, s rozlišením různých typů českých médií, a to jak technologicky, tak obsahově. Právě identifikace odlišných použitých výrazů a pojmů, které mohou přispívat k potenciálu zprávy více se šířit a případně i podněcovat paniku, je klíčová pro další kroky našeho výzkumu směřujícího k vytvoření platformy zaměřené na rozpoznávání dezinformací souvisejících s jakoukoliv epidemií.

Korpus mediálního monitoringu společnosti Newton Media týkající se problematiky COVID-19 obsahuje k 31. 10. 2020 1 153 999 zpráv. Dotazem na označení mediálních zdrojů lze určit, že zprávy pocházejí z 3392 různých mediálních zdrojů. Analyzovali jsme jak celek, tak jsme ze všech zdrojů vybrali 358 a vytvořili z nich 7 shluků se specifickou charakteristikou mediálního typu:

a) celostátní a regionální deníky včetně jejich webů (bez bulvárních deníků)

b) hlavní bulvární tištěná média a hlavní bulvární weby

c) komerční televize a jejich weby

d) hlavní ekonomická média a weby

e) komerční rádia a jejich weby

f) zpravodajské časopisy a jejich weby

g) veřejnoprávní média a jejich weby

Rozdělení jsme provedli na základě předvýzkumu a generování wordcloudů pro jednotlivá média z korpusu mediálního monitoringu Newton Media, kdy se ukázalo, že tituly obsažené v jednotlivých skupinách představených výše, vykazují velmi podobné použití slovníku, a pro zjednodušení je tedy prezentujeme společně. Z wordcloudů jsme vyřadili všechny tvary slova koronavirus, které přirozeně

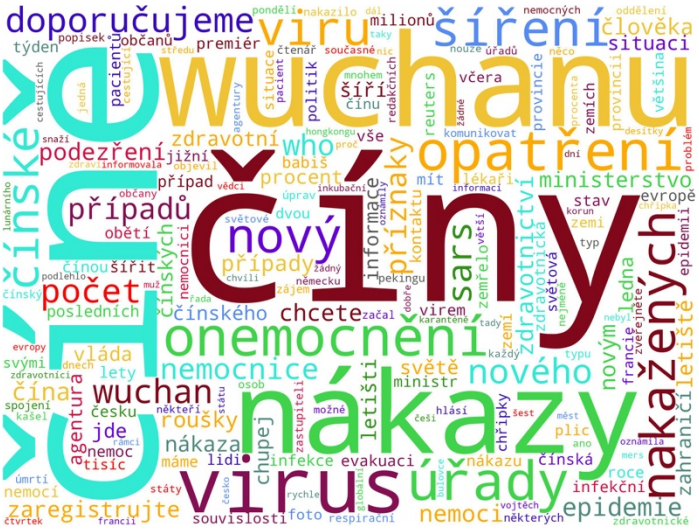
cloudům vzhledem k tématu zpráv dominovaly a ubíraly prostor slovům s tímto zpravodajstvím souvisejícím, která nás zajímala především (více o metodě viz druhá část textu).

Jak je zřejmé z obrázku 1, zpravodajství o onemocnění COVID-19 v celkovém analyzovaném zpravodajském vzorku dominovala v období od ledna do října 2020 slova *opatření*, *nakažených*, *případů*, *zdravotnictví*, *počet*, *vláda, nákazy, tisíc*. Slova se v čase měnila podle vývoje pandemie i typu médií. Vzhledem k tomu, že nejzásadnější z hlediska běžného života obyvatel byla medializace různých typů opatření (s nimiž přicházela vláda, zejména ministři zdravotnictví, zároveň zdravotnictví a kapacita nemocnic byly dalším často zmiňovaným tématem), která se neustále připravovala i dynamicky proměňovala, takže byla dlouhodobě středem pozornosti, stejně jako počty nakažených. Významné místo zaujímá zejména na jaře, kdy byly symbolem solidarity, slovo *roušky*. Diskuze o dopadu na ekonomiku ilustrují číslovky a slovo *korun*.

Obrázek 1. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v období 1. 12. 2019 – 31. 10. 2020

Velmi zajímavé je srovnání vývoje wordcloudů v čase podle témat, která byla mediálně frekventovanější obrázky 2–9 ukazují, že prosinci 2019 jednoznačně dominuje srovnání se SARS a země původu, tedy *Čína*, v lednu 2020 se zvýznamňuje i *Wuchan*, slovní zásobu února 2020 poznamenalo mediální „čekání na koronavirus“, tedy stále ještě *Čína*, ale už také *nákaza* a její *šíření*. V březnu dochází v reakci na první onemocněníné v Česku k výraznému posílení slov souvisejících se snahou o potlačení epidemie, mediálnímu pokrytí tématu proto dominují *opatření*, *vláda*, *zdravotnictví*, *situaci*, *případy*, *šíření* atp. Podobné je to i v dubnu, v květnu se už do popředí mediální pozornosti dostávají ekonomické dopady předchozích opatření a posilují ekonomická slova, jako *korun* či *procent*. Pokračuje to i v červnu, kdy se do popředí dostávají už i slova *milionů* a *miliard* v souvislosti s predikovanou ekonomickou krizí. Obdobně vypadají i cloudy z července a srpna, v září se přidává (zatím poměrně slabě) slovo nemocnice, protože se začíná s nástupem druhé vlny řešit jejich připravenost i kapacita, opět se objevují hojně diskutované *roušky* a silně pak *opatření*. V říjnu téma *nemocnice* hodně posiluje a v souvislosti s politickým rozměrem a situací na ministerstvu zdravotnictví je pak jedním z dominantních slov v září a říjnu právě *zdravotnictví.*

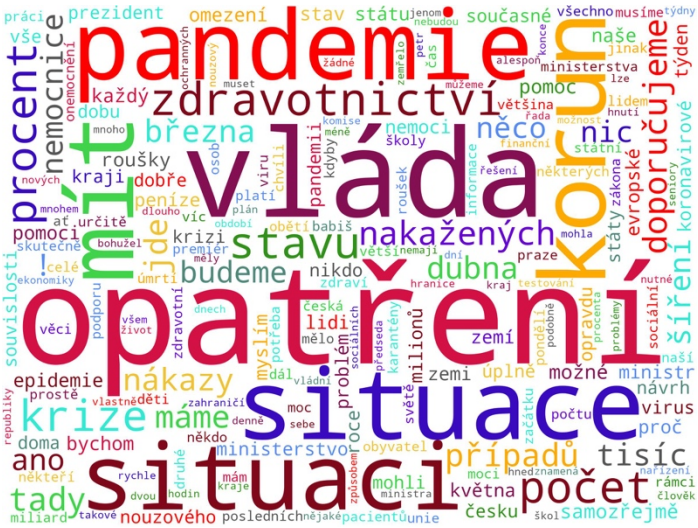
Obrázek 2. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v prosinci 2019



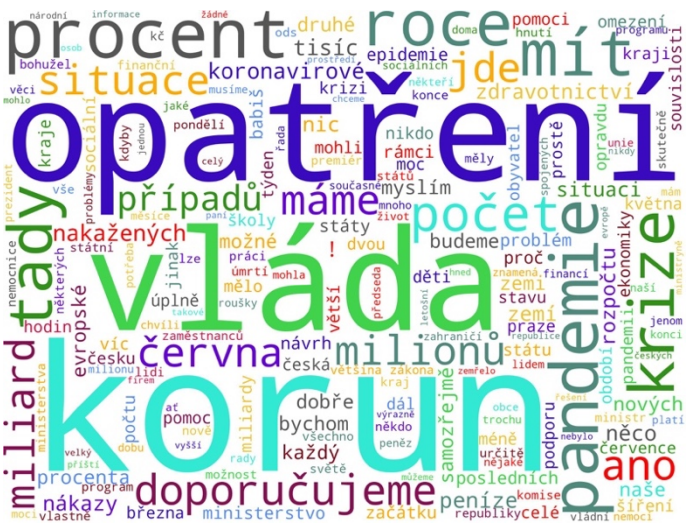
Obrázek 3. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v lednu 2020

Obrázek 4. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v únoru 2020

Obrázek 5. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v březnu 2020

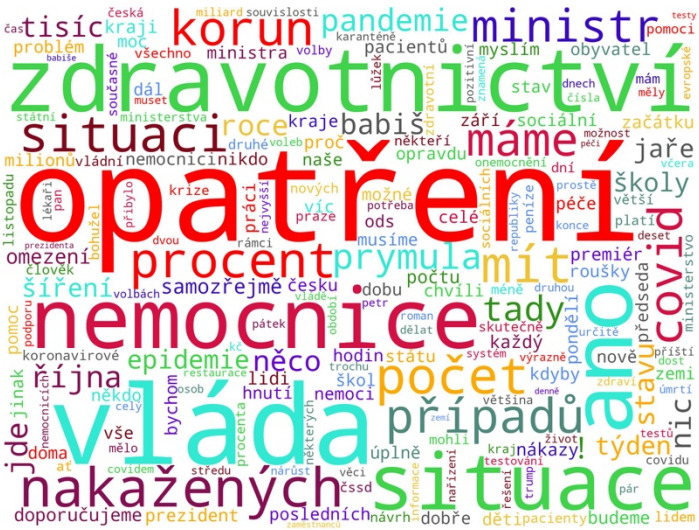


Obrázek 6. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v dubnu 2020



Obrázek 7. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v červnu 2020

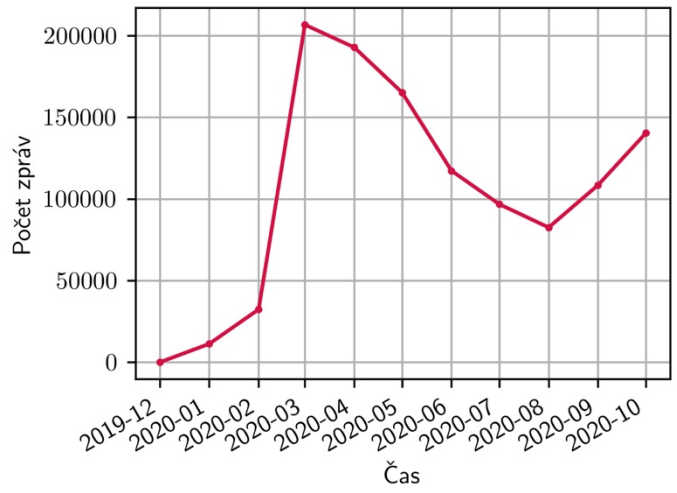
Obrázek 8. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v září 2020



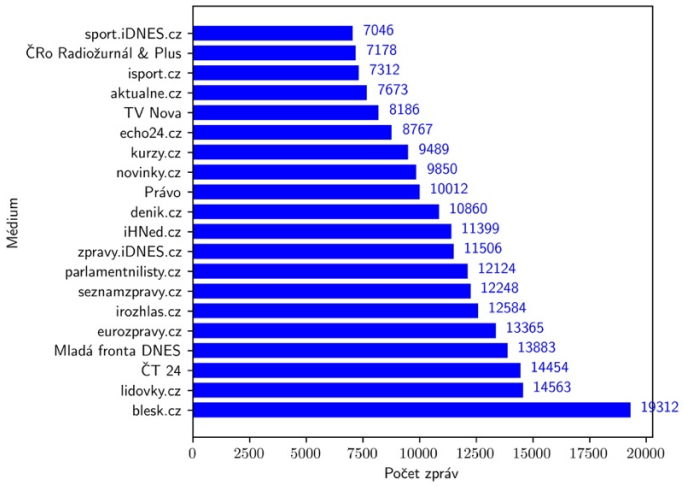
Obrázek 9. Wordcloud ze slov použitých ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v říjnu 2020

Způsob generování wordcloudů a použitá data

S cílem zajistit objektivnost hodnocení mediálních zpráv zaměřených na problematiku koronaviru jsme metody projektu postavili na předpokladu zpracování velkého množství zpráv pokrývajícího široké spektrum mediálních domů a vydavatelství. Společnost Newton Media shromáždila jenom za prvních deset měsíců roku 2020 více než milion mediálních zpráv monitorováním více než tří tisíc vydavatelů a jejich různých zpravodajských kanálů (vývoj zpráv v čase zobrazuje obrázek 10, média s nejvyšším zastoupením zpráv týkajících se nového typu koronaviru obrázek 11). V první fázi jsme proto využili metod dolování dat aplikovaných na texty zpráv v kombinaci se základními postupy zpracování přirozeného jazyka se zaměřením na specifiku českého jazyka.



Obrázek 10. Vývoj počtu zpráv o koronaviru v období 1. 12. 2019 – 31. 10. 2020



Obrázek 11. Média s nejvyšším počtem zpráv o onemocnění COVID-19 v období 1. 12. 2019 – 31. 10. 2020

V následujících odstavcích popisujeme zvolený metodický přístup konkrétní přípravy a zpracování dat. Mediální zprávy jsou firmou Newton Media získávány z mnoha zpravodajských zdrojů. Jsou monitorovány jak tištěné formy deníků a časopisů, které se z jejich různých, např. elektronických, forem převádí manuálně, poloautomatizovaně, či plně automatizovaně do normalizovaného tvaru obsahující jak vlastní text, tak i jeho metadata. Velké množství zpráv se získává přímo z online webových stránek vydavatelů, v nichž se textové a obrazové části obsahu zprávy identifikují technikami tzv. data scraping a poté kopírují do speciálního úložiště. Podobně se monitorují i řečové kanály rozhlasových a televizních stanic, které se technikami rozpoznávání řeči převádí opět na text.

Úložiště umožňuje efektivně ukládat násobné výskyty zpráv, které jsou se stejným obsahem publikovány v různých variantách zpravodajských kanálů příslušného vydavatele. Například některé regionální deníky přebírají základní centrálně vytvořený vzor a doplní na vybraná místa speciální zprávy týkající se daného konkrétního regionu. Tento fakt je potřeba zohlednit při vytváření statistik, proto jsou např. v současném exportu zpráv jsou takové zprávy uvedeny pouze jednou. Je potřeba i podotknout, že ačkoliv firma Newton Media řadu dalších zpravodajských kanálů monitoruje, tak ne všechny takové zdroje či relace jsou archivovány.

Mezi analyzované texty byla zpráva zařazena, pokud obsahovala alespoň části slov s problematikou onemocnění COVID-19 související. Vzhledem k různým inflexím českých slov se vyhledávají všechny možné varianty přípon. Technicky se takový dotaz řeší pomocí tzv. regulárních výrazů, kdy znak '*'

zastupuje libovolný řetězec znaků, v našem případě nějakou příponu slova. Výsledný dotaz je poměrně komplikovaný, ale pro jednoduchost můžeme uvést, že se hledají v principu následující slova: covi*, koron*, coron*, epidem*, pandem*, vir*, onemocnění*, zápal*, plic*, netopý*, sars*, karantén*, wuchan*, infek*, infik*, lékař*, zdravot*, šest*, nákaz*, nakaž*, vakcín*, čín*, lockdown*. Při exportu dat se obsah zpráv doplňuje o metadata a ve formě několika souborů komprimovaných pomocí ZIP techniky tvoří tak vstupní data do výpočetního procesu provádějící vlastní analýzu zaměřenou na problematiku projektu Infodemie.

Analýza dedikovaná hodnocení zaměření obsahu zpráv a jejich průběžných změn během vývoje pandemie pomocí wordcloudů je zahájena vytvořením přehledu, jak příslušný mediální zdroj používal jednotlivá slova v určitém měsíci. Tento přehled tvoří tabulku o 41 785 131 řádcích, kdy v každém řádku pro zvolený mediální zdroj, slovo a období určené příslušným měsícem roku je uveden celkový počet výskytů daného slova ve zprávách mediálního zdroje tohoto období. Před vytvořením přehledu se text zprávy upraví. Eliminují se všechny speciální znaky, které netvoří slova. Rovněž se v přehledu neuvažují čísla. Do tohoto přehledu nejsou zahrnuta slova zařazená mezi tzv. stopwords. Tato hojně se vyskytující slova typicky přispívají pouze do vytvoření syntaktické struktury věty tím, že vytváří vazby s dalšími slovy, ale hodnocení sémantického obsahu založeného pouze na slovech ovlivňují už mnohem méně. V našem případě používáme množinu prvních 300 slov, která se vyskytla vedle největšího počtu různých slov (technicky se jedná o stupeň uzlu příslušného slova v komplexní síti všech sousedních slov). Tato slova byla určena na základě analýzy téměř čtyř miliónu českých zpráv produkovaných ČTK.

Stopwords slova jsou navíc doplněna vybranými zájmeny a dalšími slovy, které se pouze odkazují na další slova, která však mohou být od takových zájmen poměrně již hodně v textu vzdálená. V rámci analýzy pomocí wordcloud diagramů jsme vyloučili i slova 'koronavirus', 'koronavir', neboť zprávy byly vybrány s ohledem na tato slova, a názvy vybraných médií, např. 'irozhlas', 'blesk', 'seznam', 'právo', 'právu', 'nova', 'tn', 'hn', 'televize', 'já', 'cz'. Tato zřejmá častá slova jsou jinak v diagramech uvedena jako nejčastější a tudíž největší a zbytečně proto zabírají místo ostatním slovům. Všechna slova se normalizují na použití pouze malých písmen. Další normalizace českého textu v analýze v současné době neprovádíme.

Podle cíleného diagramu se pak příslušným příkazem z databáze vybere frekvenční tabulka výskytů slov pro zvolenou množinu mediálních zdrojů a analyzované období a konstruují se wordcloudy. V diagramu se přibližně dvěma stovkám nejfrekventovanějších slov podle počtu výskytů slov nebo podle počtu zpráv obsahující takové slovo přiřadí velikost fontu a barva podle vybrané škály použité barevné palety. Posledním krokem je umístění obarvených slov s přiřazenou velikosti fontu do diagramu. Existuje řada metod, které se liší i kvalitou výsledků. V našem projektu používáme knihovnu Wordcloud implementovanou v jazyce Python.

Další výsledky

V následující části naší zprávy se věnujeme dvěma vybraným skupinám médií, u kterých docházelo z hlediska použitých jazykových prostředků k nejvýrazněji odlišným situacím oproti celku, a to hlavním bulvárním a hlavním ekonomickým médiím. Vygenerované wordcloudy pro vybrané typy médií pak naleznete v hypertextových odkazech v poslední části našeho textu.

Hlavní bulvární tištěná média a hlavní bulvární weby

Slovník bulvárních médií je jednodušší než např. slovník médií veřejnoprávních, zároveň je podstatně expresivnější, neboť cílem textů v bulváru je maximálně zjednodušit sdělení a připoutat s ním co největší míru pozornosti. To se potvrdilo i v případě zpravodajství o onemocnění COVID-19. Jak ukazuje wordcloud vytvořený ze všech nejfrekventovanějších slov v hlavních bulvárních médiích a na bulvárních webech (obrázek 12), na rozdíl od celkového zmapování všech médií není dominantním slovem *opatření*, ale *nakažených*, vykřičník zvýznamňuje apelativní a expresivní titulky v těchto médiích, výrazně více je zastoupené než v celku například slovo *mrtvých*. Větší výskyt příjmení hlavních politických aktérů pak ukazuje další rys bulvárního zpravodajství, a sice výraznější míru personifikace zpráv.

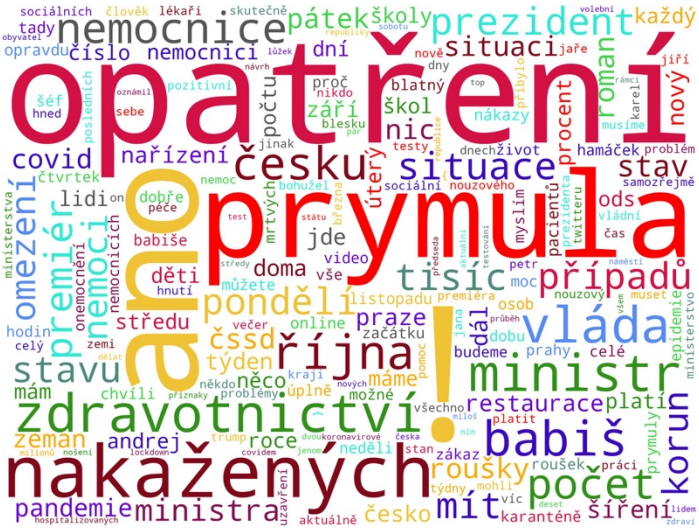
Obrázek 12. Wordcloud ze slov použitých ve zpravodajství hlavních bulvárních médií a bulvárních webů (např. Blesk, ahanonline.cz, extra.cz atp.) o onemocnění COVID-19 v období 1. 12. 2019 – 31. 10. 2020

V případě vývoje výskytu slov a interpunkce se vykřičník poprvé ve wordcloudu v bulvárních médiích objevuje v březnu 2020 (obrázek 13). Největší apelativnosti pak nabývají texty o koronaviru v květnu 2020 (obrázek 14), kde je daleko výraznější než v případě celku také slovo *vláda* a název hnutí *ANO*, zejména v souvislosti s vládními ekonomickými opatřeními. Názvy měsíců mají ve wordcloudech takové zastoupení proto, že časové údaje jsou spojovány většinou s opatřeními (nebo jejich očekáváním), která se v tomto období velmi dynamicky proměňovala a upravovala.

Obrázek 13. Wordcloud ze slov použitých ve zpravodajství hlavních bulvárních médií a bulvárních webů (např. Blesk, ahanonline.cz, extra.cz atp.) o onemocnění COVID-19 v březnu 2020

Obrázek 14. Wordcloud ze slov použitých ve zpravodajství hlavních bulvárních médií a bulvárních webů (např. Blesk, ahanonline.cz, extra.cz atp.) o onemocnění COVID-19 v květnu 2020

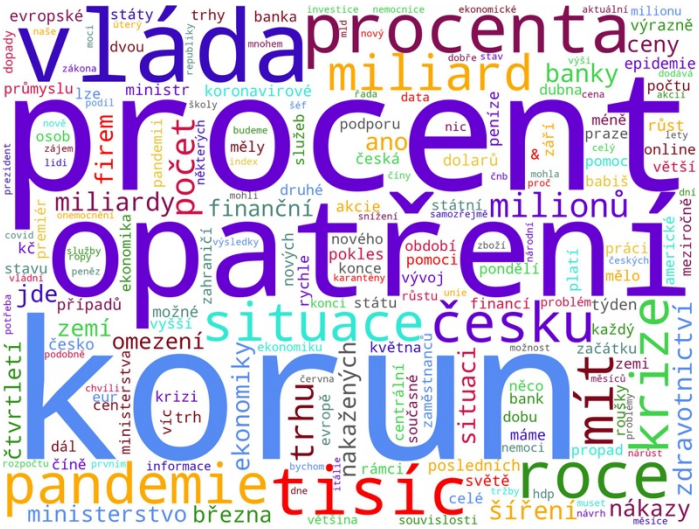
Díky porovnání obrázku 9 a obrázku 15 vidíme, jak se liší v říjnu zpracování textů o COVID-19 v bulvárních médiích od celkového mediálního pokrytí tématu. Byl to deník Blesk, který přinesl informace o schůzce tehdejšího ministra zdravotnictví Romana Prymuly s předsedou poslaneckého klubu hnutí ANO Jaroslavem Faltýnkem. Právě příjmení *Prymula* nebo příjmení premiéra *Babiš* pak zaujímají ve wordcloudu bulváru daleko více prostoru než ve wordcloudu celkovém.



Obrázek 15. Wordcloud ze slov použitých ve zpravodajství hlavních bulvárních médií a bulvárních webů (např. Blesk, ahanonline.cz, extra.cz atp.) o onemocnění COVID-19 v říjnu 2020

Hlavní ekonomická média a weby

Další z celku se vydělující entitou jsou hlavní ekonomická média a weby. Jejich slovník se z podstaty zaměřuje na ekonomiku a je logicky výrazně odbornější než u bulvárních médií nebo celkového zkoumaného korpusu. Jak ukazuje obrázek 16, wordcloudu vytvořenému ze všech nejfrekventovanějších slov v hlavních ekonomických médiích a webech dominují *koruny* a *procenta*, opakuje se silné zastoupení opatření a vlády jako ve všech skupinách médií, ale více a ve větším spektru se objevují číslovky nebo oborově specifická slova jako *banky*, *ekonomiky*, *firem* či *krize* (ve smyslu krize finanční a ekonomické).



Obrázek 16. Wordcloud ze slov použitých ve zpravodajství hlavních ekonomických médií a webů (např. Hospodářské noviny, Ekonom, Forbes atp.) o onemocnění COVID-19 v období 1. 12. 2019 – 31. 10. 2020

Období května (obrázek 17) až srpna 2020 vykazují v případě ekonomických médií ještě prohloubenější důraz na ekonomické dopady vládních opatření. V květnu se objevuje výrazné slovo *krize* či *miliard* a tento trend pokračuje i dále, s nejsilnějšími projevy v červenci a srpnu (obrázek 18).

Obrázek 17. Wordcloud ze slov použitých ve zpravodajství hlavních ekonomických médií a webů (např. Hospodářské noviny, Ekonom, Forbes atp.) o onemocnění COVID-19 v květnu 2020

Obrázek 18. Wordcloud ze slov použitých ve zpravodajství hlavních ekonomických médií a webů (např. Hospodářské noviny, Ekonom, Forbes atp.) o onemocnění COVID-19 v srpnu 2020

Související odkazy

- https://www.infomore.cz/cs/wordcloudy-fotogalerie/wordcloudy-kompletni-medialni-monitoring-newton-media/
- https://www.infomore.cz/cs/wordcloudy-fotogalerie/wordcloudy-celostatni-regionalni-deniky/
- https://www.infomore.cz/cs/wordcloudy-fotogalerie/wordcloudy-hlavni-bulvarni-media-weby/
- https://www.infomore.cz/cs/wordcloudy-fotogalerie/wordcloudy-hlavni-ekonomicka-media-weby/
- https://www.infomore.cz/cs/wordcloudy-fotogalerie/wordcloudy-zpravodajske-casopisy-weby/

T A
Č R

Projekt TL04000176 *Covid-19 infodemie: AI komunikační platforma potlačující infodemii ve vazbě na novinářskou a mediální etiku* je spolufinancován se státní podporou Technologické agentury ČR v rámci Programu Éta