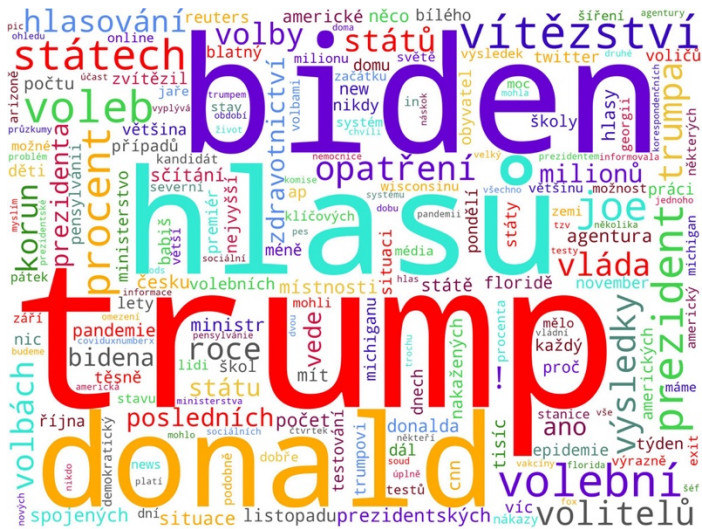


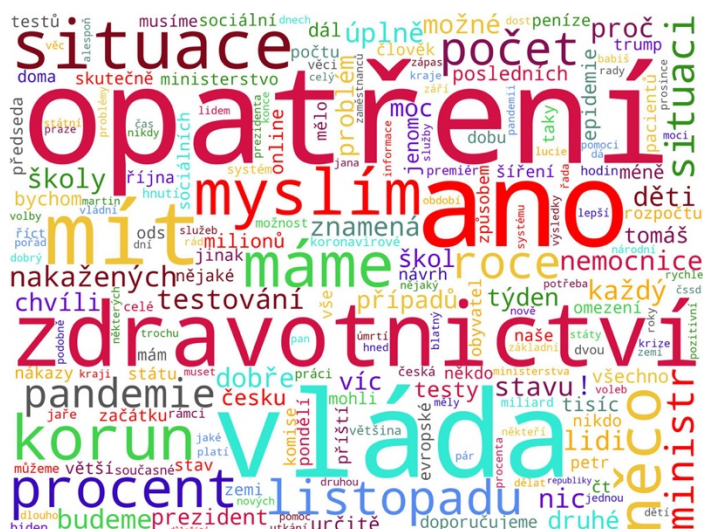
Americké prezidentské volby ve zpravodajství o onemocnění COVID-19 v českých médiích a další témata konce roku 2020

Alice Němcová Tejkalová, Radek Mařík

Nejen klasickému politickému zpravodajství dominují v posledních měsících americké volby. Velmi zajímavá situace nastala i v jednom z listopadových wordcloudů skupin českých médií, která kontinuálně sledujeme a analyzujeme. Poprvé za celý rok v něm nepřevládlo žádné slovo s pandemií nebo jejím potlačováním přímo spojené – např. opatření, vláda atp. Jak je zřejmé z obrázku 1, v textech o COVID-19 ve zpravodajských časopisech a na jejich webech ([seznam médií zařazených do této skupiny](#)) byla nejfrekventovanějšími slovy jména amerických prezidentských kandidátů Joea Bidena, Donalda Trumpa a slovo „hlasů“. Rozdíl mezi typy médií vynikne ještě více při srovnání s obrázkem 2, kde je zachycen výskyt nejčastějších slov v celém korpusu mediálního monitoringu společnosti Newton Media s 3392 různými zdroji. Příjmení Biden se v něm krčí vlevo dole a Trump diagonálně od Bidena vpravo nahoře, jejich významnost v celkovém korpusu, který vypadá tak, jak jsme zvyklí z předchozích měsíců roku 2020, je minimální. O tom, že ve zpravodajských časopisech a na jejich webech jednoznačně převládlo téma nemoci COVID-19 ve spojení s americkými prezidentskými volbami, svědčí i další výrazná slova jako „volitelů“, „volební“, „vítězství“, dále různé tvary „států“ či názvy Michigan, Florida, Pensylvánie atd. Zajímavostí je, že podobná situace nenastala například ve skupině veřejnoprávních médií, která celé volby extenzivně pokrývala díky zpravodajům přímo na místě. Pravděpodobnou příčinou je to, že volby nespojovala v takové míře s tématem přístupu kandidátů k pandemii COVID-19. Tuto tezi by ale případně bylo třeba potvrdit kvalitativním výzkumem.

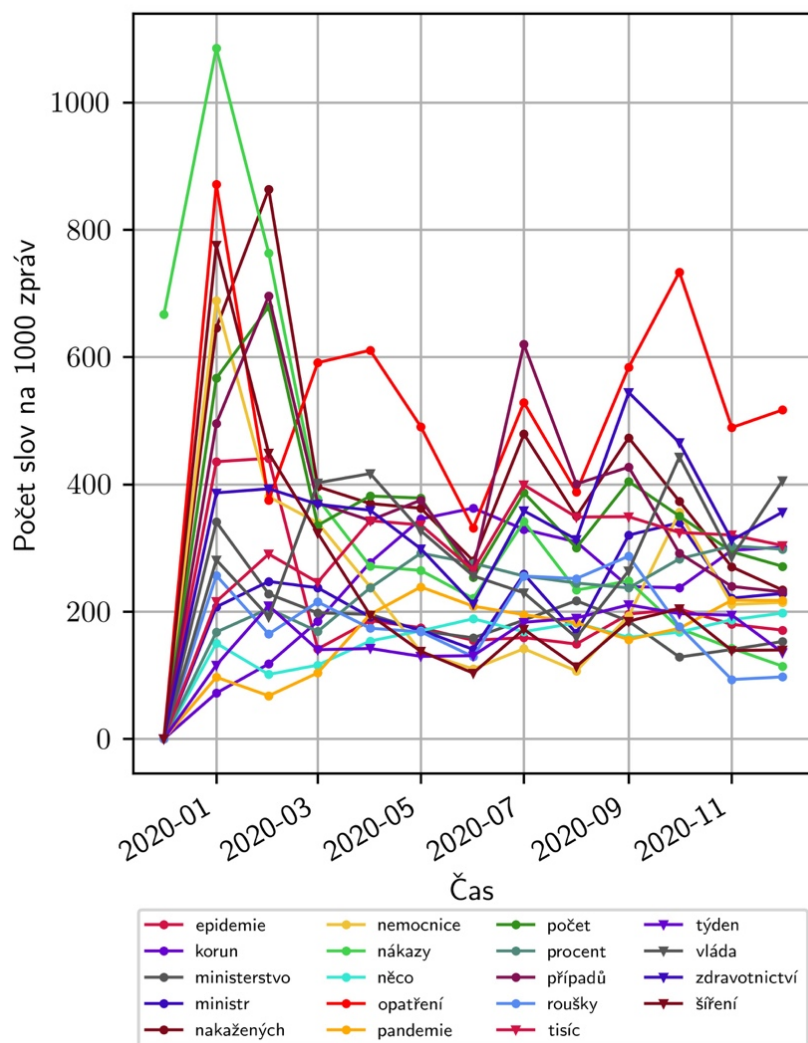


Obrázek 1. Wordcloud ze slov použitých ve zpravodajských časopisech a na jejich webech o onemocnění COVID-19 v listopadu 2020



Významným tématem, poprvé i s velmi zvýrazněným prostorovým zastoupením ve wordcloudu, se v prosinci stává očkování, s postupně stále více do popředí vystupujícími tvary slova vakcína (obrázek 3).

V grafu 1 vidíme, jakým způsobem se v čase vyvíjí výskyt slov nejčastěji se objevujících ve zpravodajství o COVID-19 ve všech mediálních zdrojích. Jsou slova, která kopírují jednotlivé vlny onemocnění, např. „opatření“, „případů“ nebo „nakažených“, zatímco některá mají tendenci s proměnou novinářského slovníku v referování o situaci ustupovat (nejlépe je to vidět na slově „nákazy“), případně se naopak zvolna od začátku roku jejich výskyt zvyšuje („pandemie“), či po úvodním prudkém růstu poklesnou, aby se v období obnoveného soustředění na lékařskou péči zase vrátila do popředí („nemocnice“).



Graf 1. Vývoj výskytu nejpoužívanějších slov ve zpravodajství všech monitorovaných médií o onemocnění COVID-19 v období 1. 1. 2020 – 31. 12. 2020

Metodologická poznámka: 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 pokrývajících 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. 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.

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 obsahujícího 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 takové zprávy uvedeny pouze jednou. Je potřeba i podotknout, že ačkoliv firma Newton Media řadu dalších zpravodajských kanálů monitoruje, tak ne všechny 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 tak tvoří 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. 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.

Tzv. stopwords jsou navíc doplněna vybranými zájmeny a dalšími slovy, která 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 zřejmě 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 velikostí fontu do diagramu. Existuje řada metod, které se liší i kvalitou výsledků. V našem projektu používáme knihovnu Wordcloud implementovanou v jazyce Python.

Související odkazy

- <https://www.infomore.cz/cs/wordcloudy-fotogalerie/wordcloudy-kompletni-medialni-monitoring-newton-media/>
- <https://www.infomore.cz/cs/wordcloudy-fotogalerie/wordcloudy-zpravodajske-casopisy-weby/>

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