CONTENT CLASSIFICATION AND KEYWORD EXTRACTION FOR *ici.radio-canada.ca*

A PREPRINT

David Alfonso
RALI, Université de Montréal

Vincent Barnabé-Lortie
Médias Numériques, Radio-Canada

Shivendra Bhardwaj
RALI, Université de Montréal

Ilan Elbaz
RALI, Université de Montréal

Abbass Ghaddar
RALI, Université de Montréal

Fabrizio Gotti
RALI, Université de Montréal

Philippe Langlais
RALI, Université de Montréal

Guillaume Le Berre
RALI, Université de Montréal

Vincent Letard
RALI, Université de Montréal

Peng Lu
RALI, Université de Montréal

Laura Elisa Salas
DIRO, Université de Montréal

Olivier Salaün
RALI, Université de Montréal

Jason Jiechen Wu
RALI, Université de Montréal

December 17, 2019

1 Introduction

Radio Canada publishes between 450 and 600 articles per day on the *ici.radio-canada.ca* site. To this date, their database contains 901 156 articles. These are mandatorily annotated by the authors using 26 themes roughly corresponding to traditional paper columns and optionally annotated using 464 Sub-themes.

The themes and sub-themes help both the human reader and the search engine bots find information in the correct domain of interest, browse multiple news connected by themes and thematically limit the field of research of a specific article.

The journalists and content editors who are in charge of choosing the theme might not always agree on how to classify a particular article. Moreover, depending on the date, context and personal preference, one slightly ambiguous article can be classified in a non-evident Theme. Concerning the Subthemes, they are simply too many and finding the right one for each article becomes a tiresome task.

This is why *Radio-Canada* wishes to improve the consistency of their classification and reduce the amount of effort for journalists and content editors by proposing automatically detected Themes and Subthemes.

Radio-Canada also wished to enrich the content description of each article. Because, historically, the idea of having keywords was not introduced in the journalistic domain until very recently, in their current state, *ici.radio-canada.ca* articles do not include them.

For this reason we were asked to study the problem of automatically extracting keywords from the article content. The ideal situation would allow not only to offer keyword suggestions to journalists and content editors but to automatically label the 901 156 articles of the data base.

2 Pre-workshop data preparation

In order to avoid the same problems encountered on the previous workshop (where some time was lost in getting and distributing the industrial data) to collect and prepare data in collaboration with *Radio-Canada* (in the person of Vincent Barnabé-Lortie), Fabrizio Gotti heavily worked on it during the week previous to the workshop so the rest of the team could get to work quicker. He facilitated the acquisition of the data, saved it in a convenient format on a server of RALI, developped a dedicated Application Programming Interface (API) for efficiently access the data. He also acquired some potentially usefull data from journal *Le Devoir* website (around 12 000 of news with their respective keywords).

3 Data analysis

The first team task we needed to tackle was the analysis of the data. Before we could translate the industry problems into implementable tasks and sub-tasks, we needed to look at the data and understand what were the tools and raw material at our disposal.

After exploring the data a bit, Vincent Barnabé-Lortie, the *Radio-Canada , Médias Numériques* representative, lighten us on the doubts that arose concerning both the structure of the data / meta-data and the journalists' procedures when submitting an article. The main ones being:

- Each article mainly consists of a title, a summary, a lead paragraph, the body (all the paragraphs), the theme and, optionally, the sub-theme.
- Some articles may be empty for some of the sections mentioned above, this mainly happens in the case of news tickers, which do not have the traditional structure of an article, per se.
- Some of the themes and sub-themes in the taxonomy are deprecated and no longer used, even thought they do appear in the data corpus.
- The journalists and content editors are required to specify one theme (and only one) for the article among the existing options.
- One of the themes in the taxonomy is called "Aucun thème sélectionné" ("No theme selected") and it seems to be used for old articles preceding the theme/sub-theme taxonomy that were not labelled back when they were written.
- It is not mandatory for the journalists and content editors to specify a sub-theme, but there is virtually no limit on the number of sub-themes an article may contain.
- There is no procedure that allows the journalist or content editor to propose a list of keywords for each article.
- One of the tools provided by *Radio-Canada* was a dictionary linking sub-themes to themes, but we observed that this theme/sub-theme link was not always respected (as further shown in Section 4).

4 Some statistics

Before continuing with the description of the tasks, we wish to rapidly expose some statistics on this specific data, which served us as guidance in the resolution of the exposed problems.

The observations derived from Figures 1, 2, 3, 4 and 5 lead us to formulate how we needed to clean the data.

As shown on Figure 1 some themes are not active or very little (and the sub-themes even more so). This means that the distribution is unequal and possibly biased towards certain themes more than others, as we will observe in the following Sections.

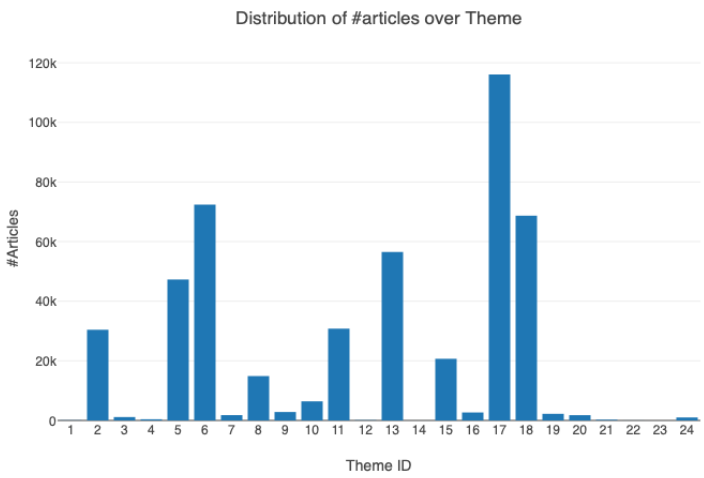


Figure 1: Distribution of the number of articles over the theme, showing some themes are not active in the whole data set.

On Figure 3 we observe that the sub-themes labeling is very imbalanced. We hypothesize that this tendency is a reflect on the need for popular sub-themes to distinguish themselves from the more general themes, which would create a motivation for the journalist to label the article with a non-mandatory sub-theme. We observe that the top 3 most frequent sub-themes are *Hockey*, *Politique provinciale* and *Éducation*.

It is worth noting that some sub-themes are not linked to one exclusive theme but to several themes (e.g.: the sub-theme *Mental health* is often linked to the themes *Health*, *Society*, *Miscelaneous news (Faits divers)*, etc.). This is further analyzed in Figures 4 and 5.

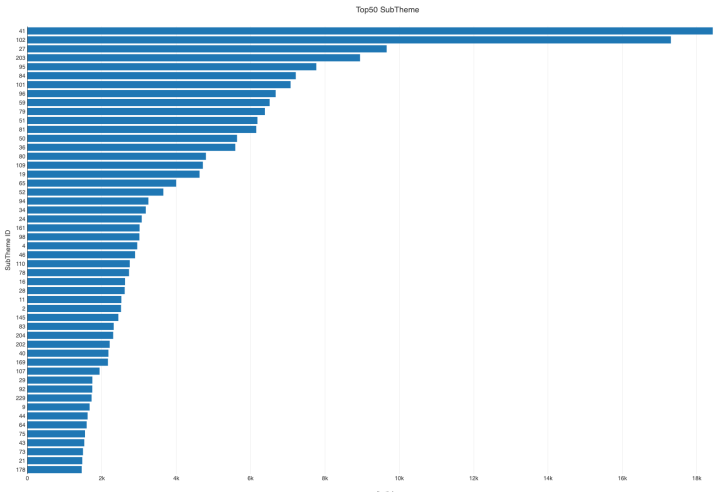


Figure 2: Excerpt of the distribution of the top 50 sub-themes over the number of articles, showing the most common sub-themes in the data set.

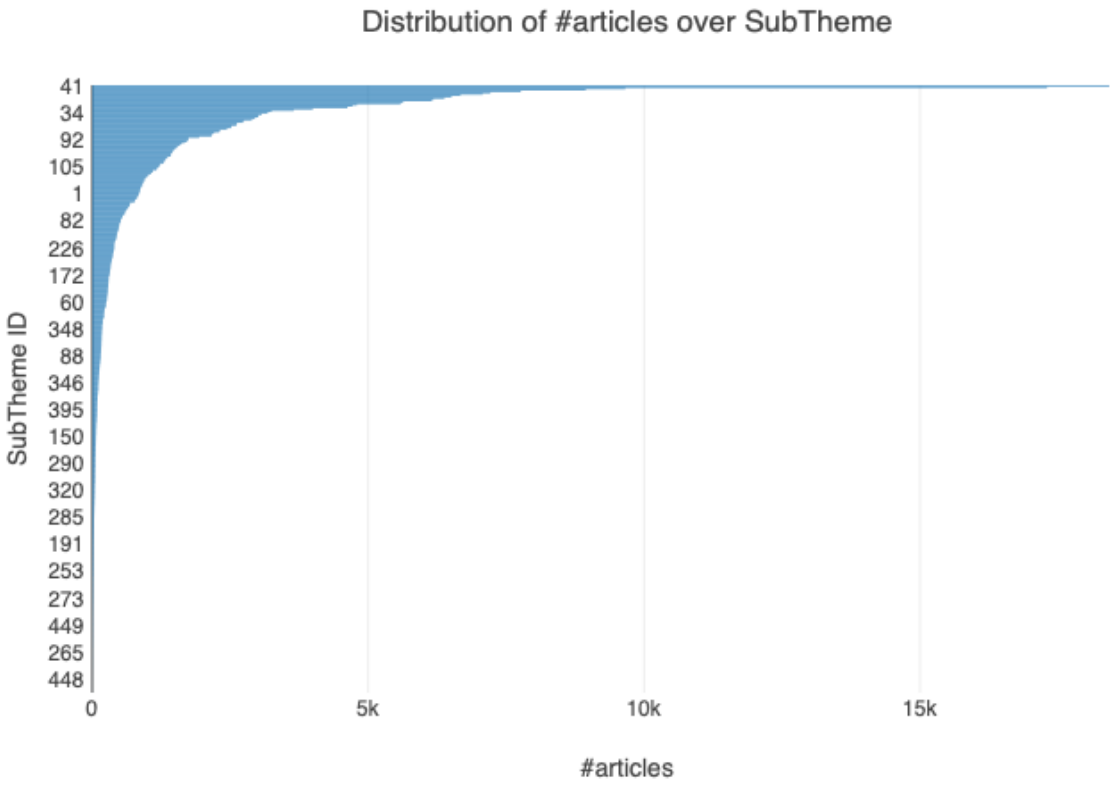
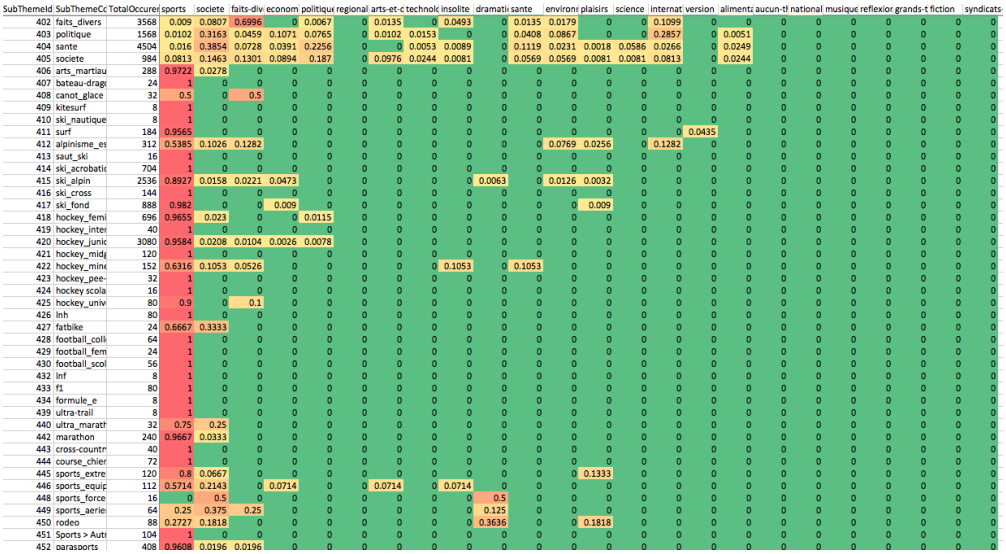


Figure 3: Sorted distribution of the sub-themes over the number of articles, showing that the distribution of sub-themes is also imbalanced.

SubThemeId	SubThemeCc	TotalOccure	sports	societe	faits-div	economi	politiqu	regional	arts-et-c	technol	insolite	dramati	sante	environ	plaisirs	science	internat	version	alimenti	aucun-tl	national	musique	reflexior	grands-t	fiction	syndicats
62	obesite	456	0	0.2807	0	0	0.0526	0	0	0	0	0	0.5263	0	0	0.0877	0	0	0.0351	0	0	0	0	0.0175	0	0
63	patinage	10456	0.9748	0.013	0	0.0008	0.0054	0.0008	0.0015	0	0	0.0031	0	0	0.0008	0	0	0	0	0	0	0	0	0	0	0
64	pauvrete	12720	0.0006	0.8327	0.0138	0.0327	0.0585	0.0013	0.0038	0	0.0013	0	0.0082	0	0.0006	0.0006	0.0201	0.005	0.0189	0	0.0019	0	0	0	0	0
65	phenomene-	31976	0.0035	0.1531	0.2274	0.0178	0.0078	0.0033	0.0008	0.0005	0.003	0.0008	0.0043	0.5462	0.0018	0.0023	0.0188	0.0065	0.001	0	0.0013	0	0	0	0.0003	0
66	plongeon	2664	0.976	0.003	0.012	0	0.009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
67	pollution	8064	0	0.0496	0.0228	0.0486	0.0397	0.006	0	0.001	0	0	0.0208	0.7688	0	0.0188	0.0208	0.001	0	0	0.002	0	0	0	0	0
68	psychologie	816	0.0098	0.2745	0	0	0.0196	0	0.0196	0.0098	0	0	0.1667	0	0.0196	0.4804	0	0	0	0	0	0	0	0	0	0
69	recensement	48	0	0.6667	0.1667	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1667	0	0	0
70	recettes	1160	0	0.0345	0	0.0069	0	0	0	0	0	0.0069	0	0	0.0621	0	0	0	0.8897	0	0	0	0	0	0	0
71	religion	11408	0.0035	0.6992	0.0884	0.0021	0.0715	0.0063	0.0077	0.0007	0.0028	0	0.0077	0.0028	0.0049	0.0028	0.0926	0.0028	0	0	0.0007	0	0.0035	0	0	0
72	reseaux-soci	6552	0.0049	0.1807	0.0415	0.0232	0.0488	0.0037	0.0208	0.5592	0.0085	0.0012	0.0098	0.0134	0	0.0049	0.072	0	0.0037	0.0012	0	0.0012	0	0.0012	0	0
73	ressources-n	11920	0	0.1154	0.0268	0.4819	0.0718	0.0329	0.0013	0	0.0027	0	0.0013	0.2456	0.0007	0.002	0.0121	0.0007	0	0	0.0034	0	0	0.0007	0	0.0007
74	musique-roc	1416	0.0056	0.0339	0.0282	0.0282	0.0056	0	0.791	0	0	0.0565	0	0	0.0056	0.0056	0	0	0	0	0	0.0395	0	0	0	0
75	securite-nati	12360	0.0136	0.4641	0.2434	0.0175	0.0919	0	0	0.0382	0.0019	0.0045	0.0304	0.0472	0.0026	0.0026	0.0317	0.0032	0	0	0.0058	0	0	0.0013	0	0
76	famille	5184	0.0154	0.6682	0.0478	0.0247	0.0864	0.0031	0.0247	0.0031	0.0046	0.0231	0.037	0.0015	0.0448	0.0015	0.0015	0.0015	0.0093	0	0	0.0015	0	0	0	0
77	sida	776	0	0.1237	0.0722	0	0	0	0	0	0	0	0.5979	0	0	0.1959	0	0	0	0	0.0103	0	0	0	0	0
78	ski	21832	0.9601	0.0106	0.0066	0.0128	0	0.0018	0.0004	0	0	0.0026	0	0.0026	0.0018	0	0	0.0007	0	0	0	0	0	0	0	0
79	soccer	51064	0.9913	0.0041	0.0017	0.0006	0.0003	0	0	0.0002	0	0.0002	0	0	0	0.0002	0.0011	0	0	0	0	0	0	0	0	0
80	systeme-de-s	38376	0.0025	0.3046	0.0567	0.0113	0.0905	0.0054	0.0002	0.0006	0.0015	0	0.468	0.0229	0.001	0.0088	0.014	0.0056	0.0058	0	0.0004	0	0.0002	0	0	0
81	tennis	49208	0.9976	0.0011	0.0011	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0002	0	0	0	0	0	0	0
82	terrorisme	4144	0	0.2587	0.1795	0.0019	0.0367	0	0.0019	0	0	0	0.0019	0.0019	0	0.0019	0.4961	0.0019	0	0	0.0135	0	0	0.0039	0	0
83	tourisme-et-	18536	0.0138	0.3751	0.0341	0.3634	0.0263	0.0052	0.0173	0.0013	0.0043	0.0302	0.0043	0.0466	0.0531	0.0017	0.0108	0.0065	0.0047	0	0.0004	0	0.0009	0	0	0
84	transport	57704	0.0019	0.4601	0.1079	0.2312	0.1171	0.0043	0.0007	0.0064	0.0031	0.0004	0.0057	0.0347	0.0019	0.0033	0.0126	0.0071	0.0001	0	0.0012	0	0	0.0001	0	0
85	urbanisme	5424	0.0029	0.5265	0.0265	0.0811	0.233	0.0074	0.0162	0.0029	0.0044	0	0.0015	0.0678	0.0074	0.0015	0.0029	0.0162	0.0015	0	0	0	0	0	0	0
86	vin	2272	0	0.1021	0.0528	0.2641	0.0387	0	0.007	0	0.0035	0.0282	0.0106	0.0141	0.0951	0.007	0.007	0	0.3697	0	0	0	0	0	0	0
87	activites-spo	10560	0.3606	0.3629	0.0523	0.0402	0.0311	0	0.0053	0.003	0.0053	0.0227	0.0402	0.028	0.0424	0.003	0.0015	0.0008	0.0008	0	0	0	0	0	0	0
88	biologie	1240	0	0.0452	0.0065	0.0065	0.0129	0	0.0194	0	0	0.0065	0.0387	0.2129	0	0.6387	0	0	0.0129	0	0	0	0	0	0	0
89	jeunesse	8168	0.0803	0.5073	0.0715	0.0235	0.0372	0	0.093	0.0078	0.002	0.0274	0.0627	0.0147	0.0186	0.0167	0.002	0.0264	0.0039	0	0	0	0.0049	0	0	0
90	industrie-du-	1720	0.0093	0.493	0.1488	0.1535	0.0791	0	0.0093	0.0186	0.0233	0.0326	0	0	0.0047	0	0.0047	0.0233	0	0	0	0	0	0	0	0
91	evenements-	11440	0.014	0.1294	0.0392	0.0231	0.0084	0	0.7007	0	0.0014	0.0608	0.0035	0.0028	0.0042	0	0.0021	0	0.0021	0	0	0.0084	0	0	0	0
92	affaires	13928	0.0086	0.0953	0.0299	0.7565	0.0402	0	0.0075	0.0126	0	0	0.0052	0.0184	0.0023	0.0006	0.0069	0.0017	0.0126	0	0.0011	0.0006	0	0	0	0
93	investisseme	2256	0.0106	0.1135	0.0177	0.695	0.0851	0	0.0213	0.0071	0	0	0.0106	0.0284	0	0.0071	0.0035	0	0	0	0	0	0	0	0	0
94	relations-de-	26000	0	0.5188	0.0529	0.2606	0.1028	0	0.0006	0.0003	0.0003	0.0003	0.0502	0.0006	0.0003	0.0012	0.0043	0.0052	0.0003	0	0.0009	0	0.0003	0	0	0
95	affaires-muni	62128	0.0067	0.3887	0.0831	0.0743	0.3831	0	0.0049	0.0024	0.0031	0.0004	0.0048	0.0467	0	0.0005	0.0001	0	0.0009	0	0.0001	0	0.0001	0	0	0
96	premieres-n	53384	0.0109	0.6012	0.0851	0.0418	0.1257	0	0.0477	0.0007	0.0001	0.0019	0.0286	0.0391	0.0034	0.0022	0.0043	0.0036	0.0015	0	0.0006	0.0004	0.0007	0	0	0
97	telecommun	4320	0.0019	0.2426	0.0315	0.3722	0.0537	0	0.013	0.237	0.0019	0	0.0037	0.0074	0	0.0056	0.0204	0.0037	0	0	0.0056	0	0	0	0	0
98	olympiques	24064	0.9711	0.0076	0.0037	0.0027	0.008	0	0.0003	0.0003	0	0	0	0	0	0.0003	0.0057	0	0	0	0.0003	0	0	0	0	0
99	rechauffeme	6560	0.0024	0.0329	0.0098	0.0378	0.1244	0	0.0049	0.0012	0	0	0.0061	0.6341	0	0.0659	0.0756	0.0024	0.0024	0	0	0	0	0	0	0
100	ecologie	1528	0	0.178	0.0052	0.0785	0.0524	0	0.0209	0	0	0	0.0052	0.6073	0.0052	0.0262	0.0209	0	0	0	0	0	0	0	0	0
101	politique-fed	56592	0.0008	0.0806	0.0158	0.0647	0.7822	0	0.0025	0.0017	0.0003	0.0003	0.0095	0.0199	0.0004	0.0013	0.017	0.0001	0.0006	0	0.0023	0	0	0	0	0
102	politique-prc	138408	0.0009	0.0887	0.0144	0.0573	0.7875	0	0.002	0.0005	0.0002	0.0001	0.0225	0.0224	0.0002	0.0004	0.0004	0.0018	0.0004	6E-05	6E-05	6E-05	6E-05	0	0	0
103	musique-cou	1320	0.0061	0.0303	0.0121	0	0.0061	0	0.8303	0	0	0.097	0	0	0.0061	0	0	0	0	0	0	0	0.0121	0	0	0
104	musique-hip	408	0	0.0588	0.0196	0	0	0	0.8431	0	0	0.0392	0	0	0	0	0	0	0	0	0	0	0.0392	0	0	0
105	industrie-for	9672	0.0008	0.1282	0.0099	0.6849	0.0314	0	0	0.0008	0.0008	0.0008	0.0008	0.1356	0	0.0033	0	0.0017	0.0008	0	0	0	0	0	0	0

Figure 4: Excerpt of the heat-map of themes (horizontal) and sub-themes (vertical) correspondence, showing sub-themes spread over numerous themes. This indicates that some sub-themes are not limited to specific themes.



SubThemeId	SubThemeCc	TotalOccurer	sports	societe	faits-div	economi	politiqu	regional	arts-et-c	technol	insolite	dramati	sante	environ	plaisirs	science	internat	version	aliments	aucun-th	national	musique	reflexior	grands-t	fiction	syndicats
402	faits_divers	3568	0.009	0.0807	0.6996	0	0.0067	0	0.0135	0	0.0493	0	0.0135	0.0179	0	0	0.1099	0	0	0	0	0	0	0	0	0
403	politique	1568	0.0102	0.3163	0.0459	0.1071	0.0765	0	0.0102	0.0153	0	0	0.0408	0.0867	0	0	0.2857	0	0.0051	0	0	0	0	0	0	0
404	sante	4504	0.016	0.3854	0.0728	0.0391	0.2256	0	0	0.0053	0.0089	0	0.1119	0.0231	0.0018	0.0586	0.0266	0	0.0249	0	0	0	0	0	0	0
405	societe	984	0.0813	0.1463	0.1301	0.0894	0.187	0	0.0976	0.0244	0.0081	0	0.0569	0.0569	0.0081	0.0081	0.0813	0	0.0244	0	0	0	0	0	0	0
406	arts_martiau	288	0.9722	0.0278	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
407	bateau-drag	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
408	canot_glace	32	0.5	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
409	kitesurf	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
410	ski_nautique	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
411	surf	184	0.9565	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0435	0	0	0	0	0	0	0	0
412	alpinisme_es	312	0.5385	0.1026	0.1282	0	0	0	0	0	0	0	0	0.0769	0.0256	0	0.1282	0	0	0	0	0	0	0	0	0
413	saut_ski	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
414	ski_acrobatic	704	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
415	ski_alpin	2536	0.8927	0.0158	0.0221	0.0473	0	0	0	0	0	0.0063	0	0.0126	0.0032	0	0	0	0	0	0	0	0	0	0	0
416	ski_cross	144	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
417	ski_fond	888	0.982	0	0	0.009	0	0	0	0	0	0	0	0	0.009	0	0	0	0	0	0	0	0	0	0	0
418	hockey_femi	696	0.9655	0.023	0	0	0.0115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
419	hockey_inter	40	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
420	hockey_junic	3080	0.9584	0.0208	0.0104	0.0026	0.0078	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
421	hockey_midg	120	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
422	hockey_mine	152	0.6316	0.1053	0.0526	0	0	0	0	0	0.1053	0	0.1053	0	0	0	0	0	0	0	0	0	0	0	0	0
423	hockey_pee-	32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
424	hockey scola	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
425	hockey_univ	80	0.9	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
426	lnh	80	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
427	fatbike	24	0.6667	0.3333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
428	football_coll	64	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
429	football_fem	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
430	football_scol	56	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
432	lnf	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
433	f1	80	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
434	formule_e	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
439	ultra-trail	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
440	ultra_marath	32	0.75	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
442	marathon	240	0.9667	0.0333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
443	cross-countr	40	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
444	course_chier	72	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
445	sports_extre	120	0.8	0.0667	0	0	0	0	0	0	0	0	0	0	0.1333	0	0	0	0	0	0	0	0	0	0	0
446	sports_equip	112	0.5714	0.2143	0	0.0714	0	0	0.0714	0	0.0714	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
448	sports_force	16	0	0.5	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
449	sports_aerie	64	0.25	0.375	0.25	0	0	0	0	0	0	0.125	0	0	0	0	0	0	0	0	0	0	0	0	0	0
450	rodeo	88	0.2727	0.1818	0	0	0	0	0	0	0	0.3636	0	0	0.1818	0	0	0	0	0	0	0	0	0	0	0
451	Sports > Aut	104	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
452	parasports	408	0.9608	0.0196	0.0196	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5: Excerpt of the heat-map of themes (horizontal) and sub-themes (vertical) correspondence, showing sub-themes very focused on one of very few themes. This indicates that some sub-themes are rather "faithful" to a specific theme.

5 Data cleaning and splitting

As a group, we agreed on the need to clean the data and use the same subsets to train our models and evaluate ourselves if we wanted our results to be comparable.

As we said on Section 3, the sub-theme is not a mandatory requisite for publishing an article on *ici.radio-canada.ca*. This means that a lot of the articles have a theme but no sub-theme. In order to work on equal ground when using themes and sub-themes classifiers, we decided to remove all articles that did not contain both a theme and (at least) one sub-theme.

There is one theme that corresponds to "No theme selected". Since this null-theme matches no specific domain, we chose to remove all articles labeled with this theme from our data set.

After cleaning, the total number of articles fell from around 900 000 to, approximately, 240 000. After this, the remaining data was split into 3 parts: the Training set, the Validation set and the Test set. After having randomized the whole data set, of the 240 000 total articles, 80% were assigned to the Train set, 10% to the Validation set and another 10% to the Test set. This was made once and distributed to the whole team so we could train, validate and test each model on the same subsets and obtain comparable results.

We also decided of a specific output format of the various classifiers we tested and an evaluation protocol for which an algorithm was implemented in order to assure evaluation uniformity for every team.

These tasks mainly fell on Fabrizio Gotti's shoulders, since they were necessary to start working on the problems as soon as possible and he already had an advance knowledge of the data, the format and the API.

Despite one week for preparing data before the workshop, the team spent more than a day to decide of a way to clean the data, and to proceed.

Table 1: Classification task scores.

Sub-task	Classifier	Precision @1	Recall @1	F1 @1
Theme identification	BERT	75.4%	75.4%	75.4%
Theme identification	Fasttext	77.6%	77.6%	77.6%
Theme identification	Logistic regression	77.8%	77.8%	77.8%
Theme identification	**SVM**	**78.7%**	**78.7%**	**78.7%**
Sub-theme identification	BERT	7.1%	5.8%	6.4%
Sub-theme identification	**Fasttext**	**71.1%**	**58.1%**	**63.9%**
Sub-theme identification	Logistic regression	na	na	na
Sub-theme identification	SVM	na	na	na

6 Theme/Sub-theme classifiers

The first problematic consisted in training in a supervised way some classifiers to predict the theme or the sub-themes on an article, based on its contents, that is, its words.

Due to time constraints, we tested two well-established feature-based approaches, namely Linear Regression and Support Vector Machines (Cortes et Al. [2]), as well as two approaches based on deep learning, namely BERT (Devlin et al. [3]) and Fasttext (Joulin et al. [4]). It is important to stress that many of those algorithms are controlled by hyper-parameters that we did not have time to investigate.

All resulting scores are shown on Table 1.

6.1 Logistic Regression

We first experimented with logistic regression, where each article is represented by a vector of dimension V, where V is the vocabulary size (the size of the list of each word in the training material). The value of each coefficient is computed with the standard TF-IDF score; very roughly frequent words in a document get rewarded by this score if not too frequent in the other articles. Training such a model amounts to learn to weight each dimension (word) of the associated vector space representation so to minimize error classification. The advantage of such a technique is that we can observe the learned weights and see what the model thinks about the importance of a specific word for a given theme (or sub-theme). We used Scikit-Learn for training the model nearly in its default setting. A modest number of variants have been tried that rendered very close performances in the end.

6.2 SVM

Support Vector Machine has been repeatedly reported as a robust classifier. Therefore we tested one such approach. The feature representation was very similar to the one given to the linear regression approach. We used the implementation in Scikit-Learn and its SGDClassifier function with almost all of the default arguments except for the random state value (42), the maximum of iterations (limited to 16), and tolerance (None).
This model rendered the best performance as a theme classifier.

Both the Logistic Regression and SVM models can run on ordinary CPUs and do not require a state-of-the-art computation power.

6.3 Bert

BERT is a pretrained model. It provides contextual embeddings of the words in a sentence and can ultimately be used to extract a vector representation of a sequence of tokens. The model can be further fine-tuned to obtain better representation for a particular task.

For the theme/sub-themes classification task, we fed to BERT the first 128 tokens of the body and we took the embedding of the first token (corresponding to the "Start" symbol) to be our sentence representation. This representation

is then fed to two distinct linear layers followed by softmax layers (for themes and sub-themes). We used a cross-entropy loss for training. For sub-themes, experiments have been made using a multi-label objective but this method showed really weak results and we finally changed it to be a cross-entropy loss using only the first sub-theme as gold label.

Training is done using Adam with a learning rate of 10^{-5}. For these experiments we used a pytorch implementation of BERT (pytorch-transformers) and the "bert-base-multilingual-cased" model.

6.4 Fasttext

Another popular and recent neural network pre-trained model we tested was Fasttext, from Facebook's AI Research lab.

We used the basic available model and chose, as hyper-parameters, a learning rate of 1.0, using a vector dimension of 50, a word ngram length of 2, a bucket value of 200 000 and we trained for 40 epoch. As a result, we got a good score for theme classification and the best score for sub-theme classification. It is noteworthy to say that these sub-theme results rendered the best scores for the sub-theme classification sub-task.

Contrary to the Logistic Regression and SVM models, the BERT and Fasttext models require to run on highly potent GPUs in order to get a relatively fast result. BERT was trained using four Nvidia RTX 2080 Ti and convergence is achieved after approximately 2 hours. BERT was trained using one Titan XP and convergence is achieved after approximately 4 hours.

6.5 Future work

One of the most evident conclusions the classifier task teams arrived to, is that in order to better understand the data, we need to make a better analysis of the labeling done by the journalists. Understanding where are their priorities when they label an article with a specific theme or sub-theme is key to understand how clean or uniform the data is.

For the theme classification task, the best performing model was SVM. We expected it to outperform the Linear Regression model, but we were more surprised by the lower performance of the deep learning approaches. Several reasons can explain this, among which the relatively small training set, and the multilingual embeddings we used to seed the model.

As we can observe, there is room for improvement yet. Even with our best efforts, there is only so much we can do in just one week of work with the data. Our highest F1 score for theme classification is 78.7% and 63.9% for sub-theme classification. We believe these scores can be improved by testing other models, better adjustment on the hyper-parameters or even an automatic data cleaning for ambiguously labeled themes.

7 Keyword extraction

In order to allow an improved article access to the potential reader, the theme is not always enough. This is because themes and sub-themes only represent a more or less general thematic domain while, very often, the reader has a greater interest in a specific subject. This is where keywords become useful.

Since the database does not have any keyword annotations and we lack a large number of reader-annotators to manually extract, pinpoint or label each one of the 900 000 articles with multiple keywords, we moved to an automatic keyword extraction method. By which we mean to analyze the content of the article in order to deduce what words (or group of words) would be representative keywords.

The greater problem we encountered with this task, was the evaluation side of it. Without a keyword annotated corpus, we were unable to produce any evaluative test other than the superficial human analysis of some very limited examples. We still spent some time deploying mainly 2 core technologies that we detail hereafter.

7.1 DBpedia Spotlight

DBpedia Spotlight (Mendes et al. [5]) is an open-source free tool that allows to connect text to existing entries of Wikipedia and DBpedia. Basically, by giving the body of the article to the tool, we are able to extract the words and group of words that appear both in the article and as encyclopedic/ontological entries in those resources.

After running this tool we were able to extract 14 867 719 keywords from the 901 156 articles. This represents an average of 16.5 keywords per article. An excerpt of an article with the extracted keywords can be seen in Figure 6.

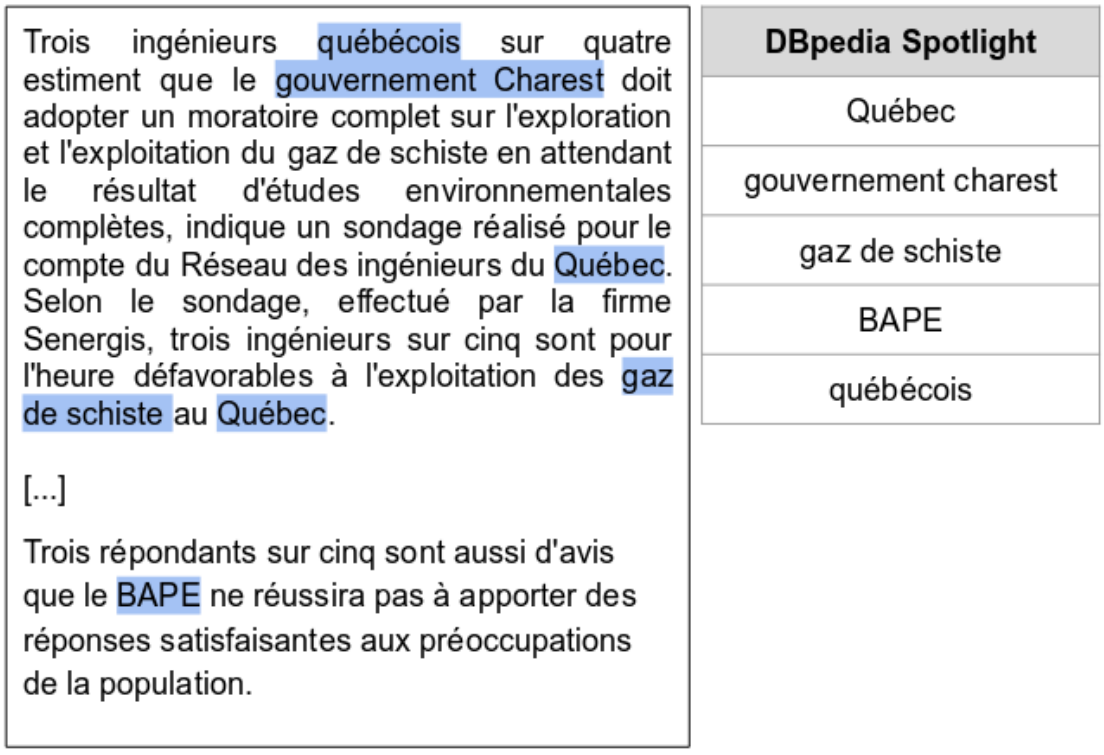


Figure 6: Excerpt of an article and extracted keywords from DBpedia Spotlight.

7.2 YAKE

Yet Another Keyword Extractor (YAKE) (Campos et al. , [1]) is also a free open-source tool to extract keywords. The main difference is that YAKE is completely language independent and instead of mapping words to encyclopedic entries, it is based on the word frequencies, their location in sentence, named entities indicators, etc.
This means that it extracts much more potential keywords but also much more noise as we can see in Figure 7.

7.3 Future work

Concerning the keyword extraction task, we already have results but we lack the evaluation tools to test them. We thought of multiple ways we could have done so, if we had more time.

One way would have been to run both these tools on keyword labeled data and compare the output result with the human labels. Our guess is that this method would have rendered a very low precision score, since the algorithms are not able to detect which keywords are representative of the whole article the same way a human annotator can.

Another evaluation strategy we could have used is the human evaluation. By detecting all the keywords from the keywords extraction tools and having them analyzed by a jury of annotators, we would be able to get an extremely precise way to know if the task is up to human standards. Nevertheless, this is extremely time and cost intensive

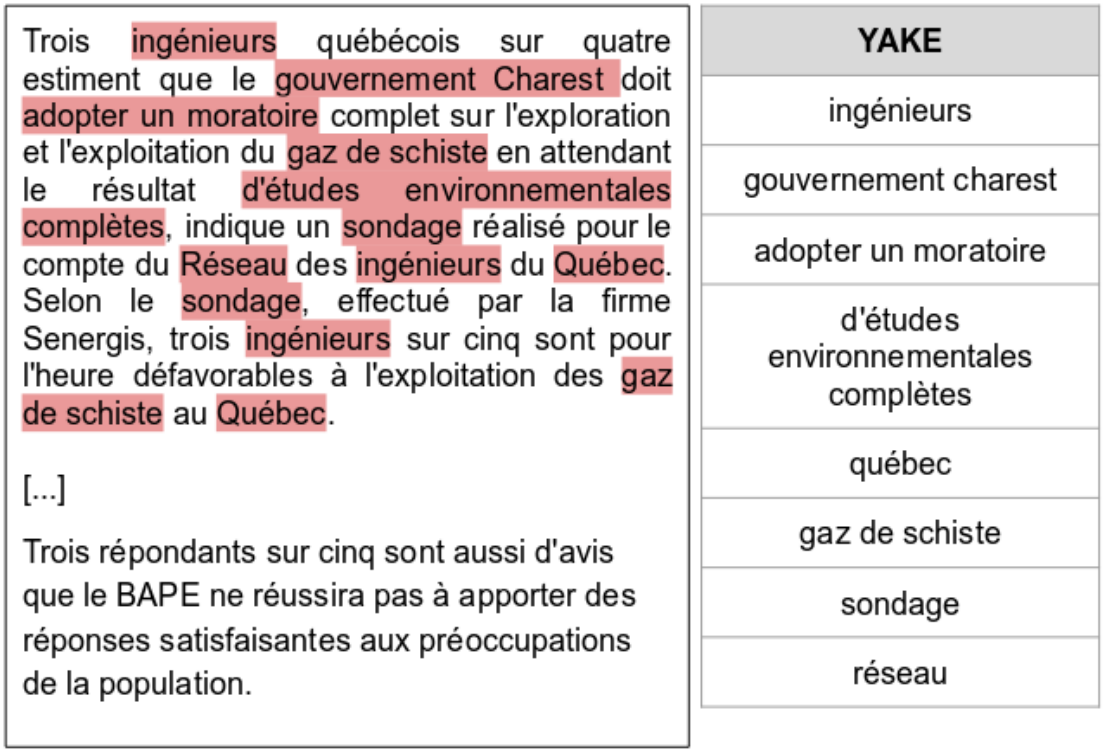


Figure 7: Excerpt of an article and extracted keywords from DBpedia Spotlight.

since we would need at the very least 3 human annotators reading each article and analyzing the keywords that were extracted, and on an sample set big enough to be representative.

Another choice would be to annotate our corpus with the keywords from each system and seeing if they improve the results on our theme/sub-theme classification task.

8 Conclusion

We had 2 tasks for this workshop: to classify articles by themes and sub-themes based on their content and to automatically extract the keywords. For the first task we proposed four different systems rendering very close results: one based on Google's BERT model, one based on Facebook's Fasttext model, one based on the logistic regression approach and one based on the SVM approach. The classification sub-tasks were evaluated by comparing the top 1 prediction of the model to the human annotated label. For the theme classification sub-task, the SVM returns the best result of all but just by a 3% difference. For the sub-theme classification problem, we experimented with only two out of four methods, of which, Fasttext rendered the best result.

The theme/sub-theme classifiers allow to assist the journalists and content editors and save them the time to search potential themes and sub-themes to classify each new article. The keyword extractor allows to add content metadata to the already existing article database. Both tasks could be further improved in many ways but this would require an even more profound analysis of the data annotation and an implementation of keyword extraction evaluation.

References

[1] Campos, Ricardo and Mangaravite, Vítor and Pasquali, Arian and Jorge, Alípio Mário and Nunes, Célia and Jatowt, Adam. A text feature based automatic keyword extraction method for single documents. In *European Conference on Information Retrieval*, pages 684–691. Springer, 2018.

[2] Cortes, Corinna and Vapnik, Vladimir. Support-vector networks. In *Machine learning*, volume 20, number 3, pages 273–297. Springer, 1995.

[3] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint*, arXiv:1810.04805, 2018.

[4] Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Mikolov, Tomas. Bag of tricks for efficient text classification. In *arXiv preprint*, arXiv:1607.01759, 2016.

[5] Mendes, Pablo N and Jakob, Max and García-Silva, Andrés and Bizer, Christian. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM, 2011.