CRM / IVADO Industrial Problem Solving Workshop

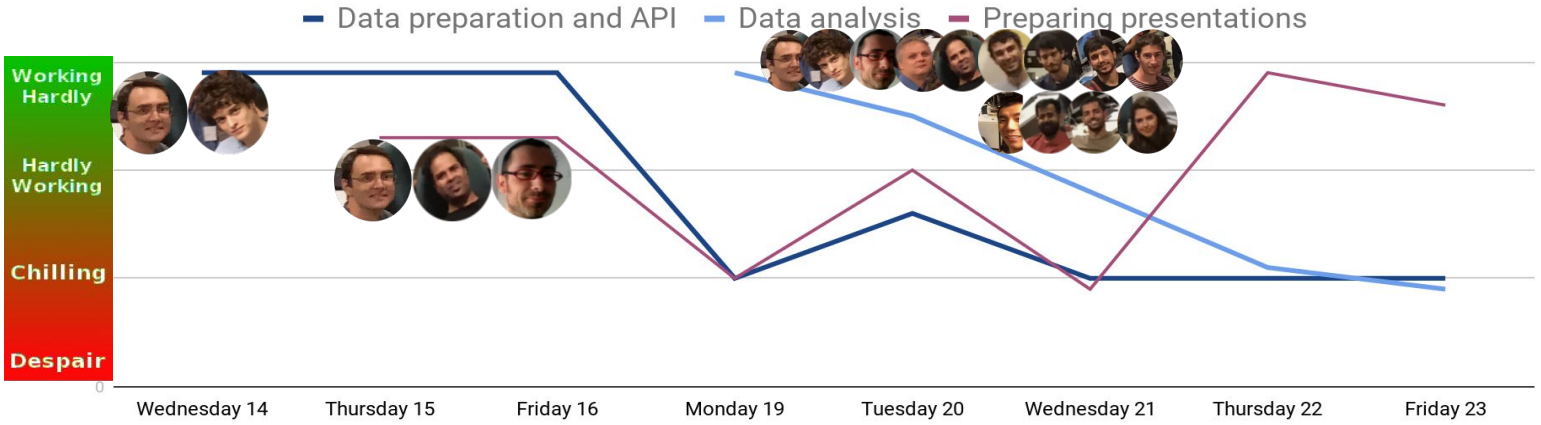
Automatic Content Categorization and Description

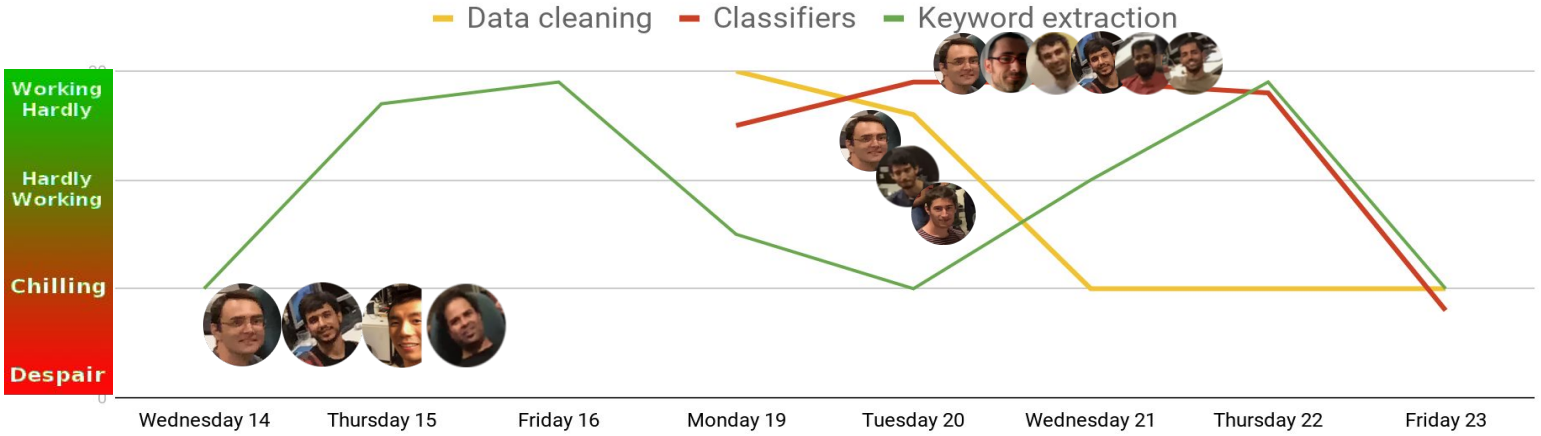
Médias numériques

Université de Montréal

Our Team

Our Schedule





The problems & objectives

- **Problem 1 :** Journalists are required to choose a theme for their article but subthemes are optional.
    - Train a classifier that will suggest themes and subthemes based on the content of the article.
- **Problem 2 :** Journalists are not required to annotate the articles with keywords.
    - Extract and suggest keywords based on the content of the article.

Raw resource:

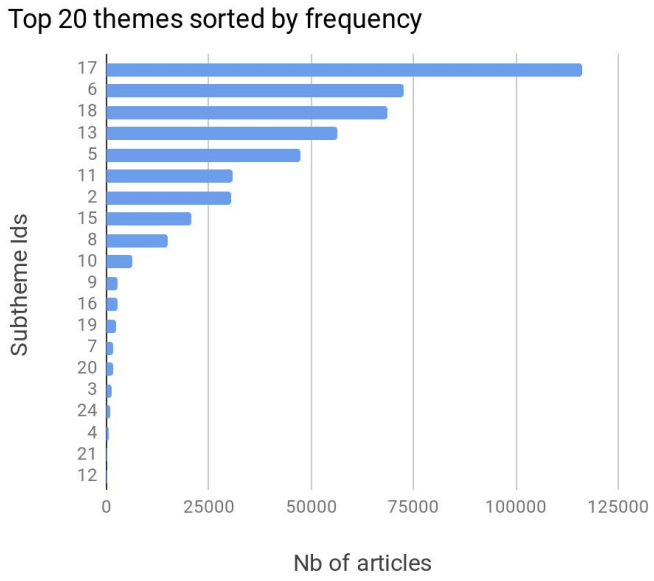
- 900k radio-canada articles

Daily growth:

- 450-600 articles/day

Theme/subtheme distribution

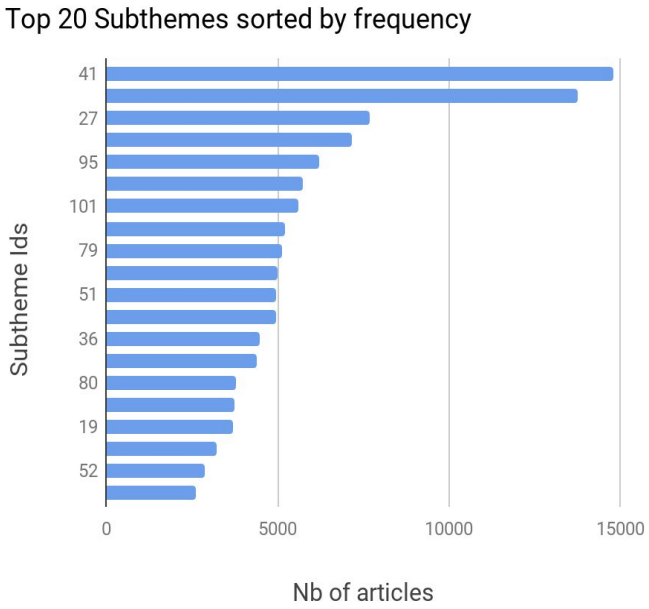
Themes :



Top 3 Themes:

- 17 - Société
- 6 - Sports
- 18 - Justice et faits divers

Subthemes :



Top 3 Subthemes:

- 41 - Hockey (theme Sport)
- 102 - Politique provinciale (theme Politique)
- 27 - Éducation (theme Société)

Clean resource:

- 240k radio-canada articles

Nb of Themes:

- 26

Nb of Subthemes:

- 466

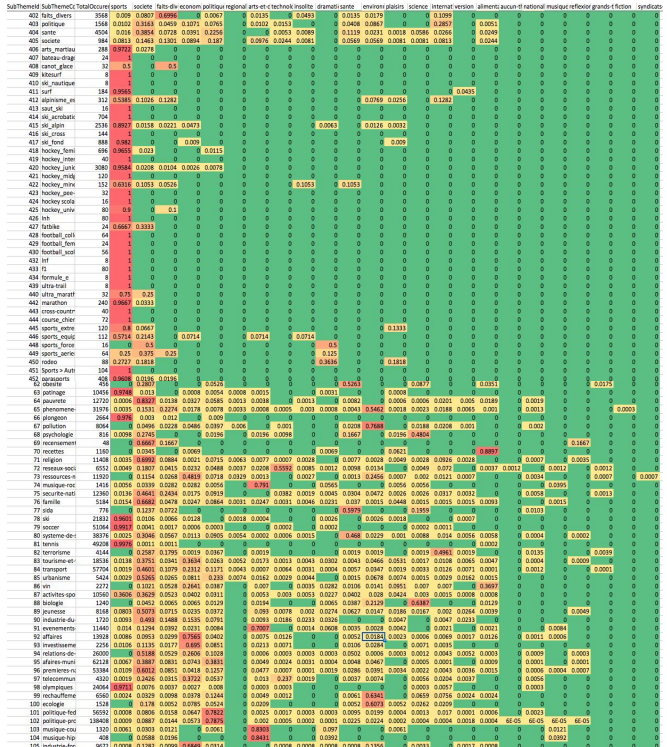
More Subtheme Stats

SubThemeId	SubThemeCc	TotalOccurer	sports	societe	faits-div	econom	politique	regional	arts-et-c	technolc	insolite	dramati	sante	environ	plaisirs	science	internat
177	finance	1544	0.0052	0.1088	0.0466	0.6062	0.1554	0	0.0104	0.0363	0	0	0.0104	0	0	0	0.0207
178	forces-policie	11656	0.0027	0.256	0.6452	0.0055	0.0638	0	0.0007	0.0041	0.0041	0	0.0055	0.0014	0.0007	0.0021	0.0055
179	futur	32	0	0	0	0	0	0	0	0.75	0	0	0	0	0.25	0	0

In short :

Even after cleaning, Subtheme labeling is messy.

Some appear in only one Theme, others in MANY.



Clean resource:

- 240k radio-canada articles

Nb of Themes:

- 26

Nb of Subthemes:

- 466

Theme/subtheme classifiers

Task	Classifier	Precision @1	Recall @1	F1 @1
Theme Identification	Bert	73.8%	73.8%	73.8%
	Fasttext	77.6%	77.6%	77.6%
	Logistic regression	77.8%	77.8%	77.8%
	SVM	78.7%	78.7%	78.7%
Subtheme Identification	Bert	7.1%	5.8%	6.4%
	Fasttext	71.1%	58.1%	63.9%
	Logistic regression	NA	NA	NA
	SVM	NA	NA	NA

Classifier Facts

In feature-based classifiers, features = presence/absence of specific words.
Most correlated features are typically good representatives of each class

- **Theme 5 (Économie) :** entreprise, économique, entreprises, finances, commerce, économie, emploi, usine, entrepreneurs, brunswick, constructeur, commerces.
- **Theme 18 (Justice et faits divers)** : homme, police, sherbrooke, incendie, olympique, incident, accident, cour, drummondville, juge, rappelons.
- **Theme 17 (Société) :** ottawa, reportage, automobilistes, alward, autochtones, brunswick, csf, couronne, mâmawi, histoire, ailleurs, recteur, patients.
- **Theme 15 (Environnement) :** environnement, écologique, climatiques, durable, inondations, climatique, sinistrés, conservation, environnementale, déchets.

Theme (truth)	SVM (predicted)	Lin. Reg. (pred)	Fasttext (pred.)
Économie	Économie	Économie	Économie
Économie	Justice et faits divers	Justice et faits divers	Justice et faits divers
Économie	Société	Environnement	Société

Keyword Extraction

Trois ingénieurs québécois sur quatre estiment que le gouvernement Charest doit adopter un moratoire complet sur l'exploration et l'exploitation du gaz de schiste en attendant le résultat d'études environnementales complètes, indique un sondage réalisé pour le compte du Réseau des ingénieurs du Québec. Selon le sondage, effectué par la firme Senergis, trois ingénieurs sur cinq sont pour l'heure défavorables à l'exploitation des gaz de schiste au Québec.

[...]

Trois répondants sur cinq sont aussi d'avis que le BAPE ne réussira pas à apporter des réponses satisfaisantes aux préoccupations de la population.

DBpedia Spotlight	YAKE
Québec	ingénieurs
gouvernement charest	gouvernement charest
gaz de schiste	adopter un moratoire
BAPE	d'études environnementales complètes
québécois	québec
	gaz de schiste
	sondage
	réseau

In Conclusion

- Automatic and assisted labeling into Themes is possible (accuracy ~ 80%).
- Automatic and assisted labeling into Subthemes is more complex.
- Keyword extraction seems promising (and potentially helpful for Content Indexing).

Future Work

- Better analysis of the labeling done by journalists.
- Classifiers results can be improved.
- Keyword extraction should go through a human evaluation process.
- Sleeping.

Questions?