Soutenance de DataStream

Classification et clustering de tweets dans des topics

SAMO KAMGA Marius Bartel, TORY Zakaria, ZAFY Karine

M2 Data Sciences - Institut Polytechnique de Paris

15 janvier 2023

Table of Contents

Contexte et objectifs

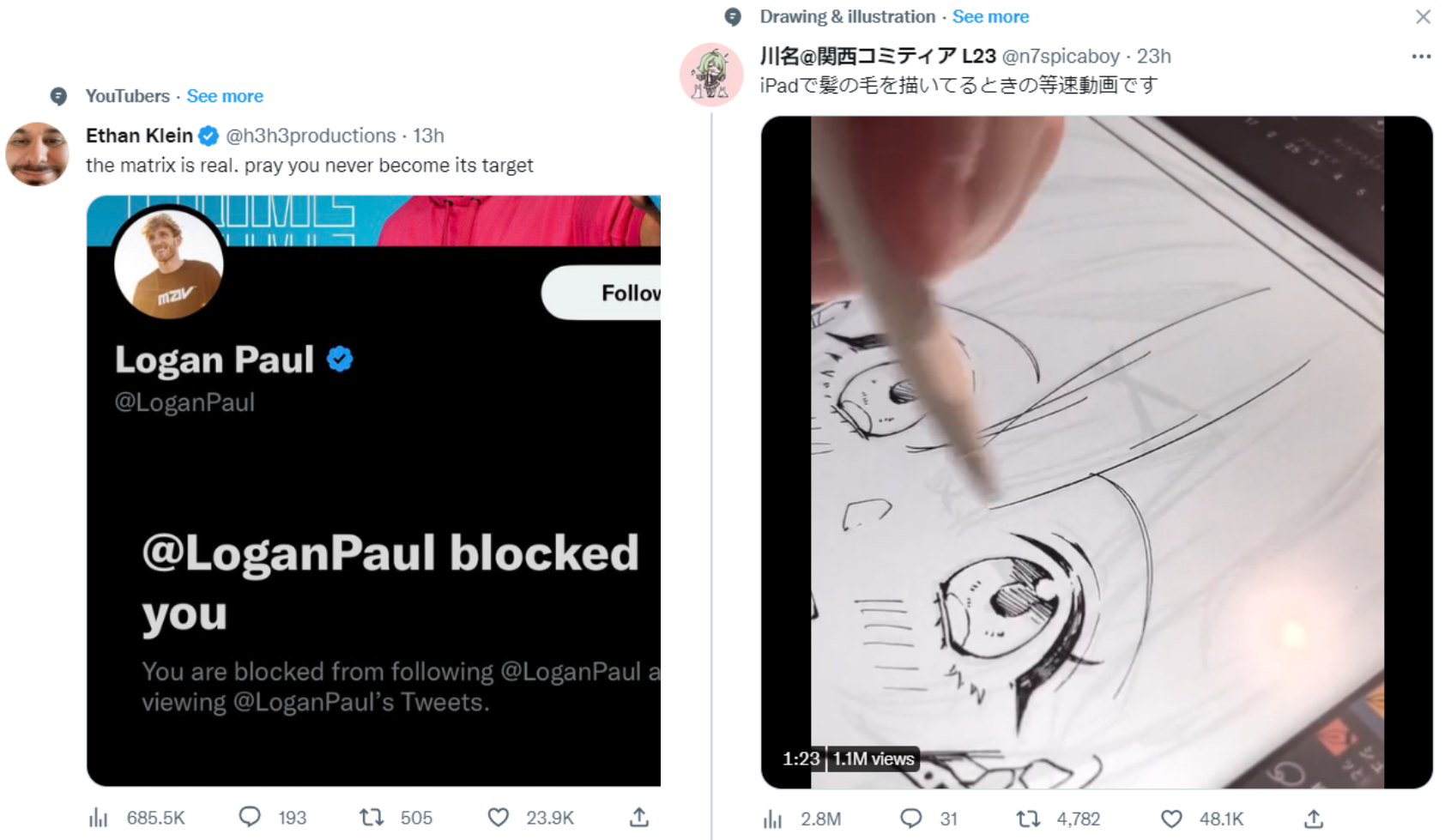


Figure – Tweets automatiquement classés dans des topics par Twitter

Contexte et objectifs

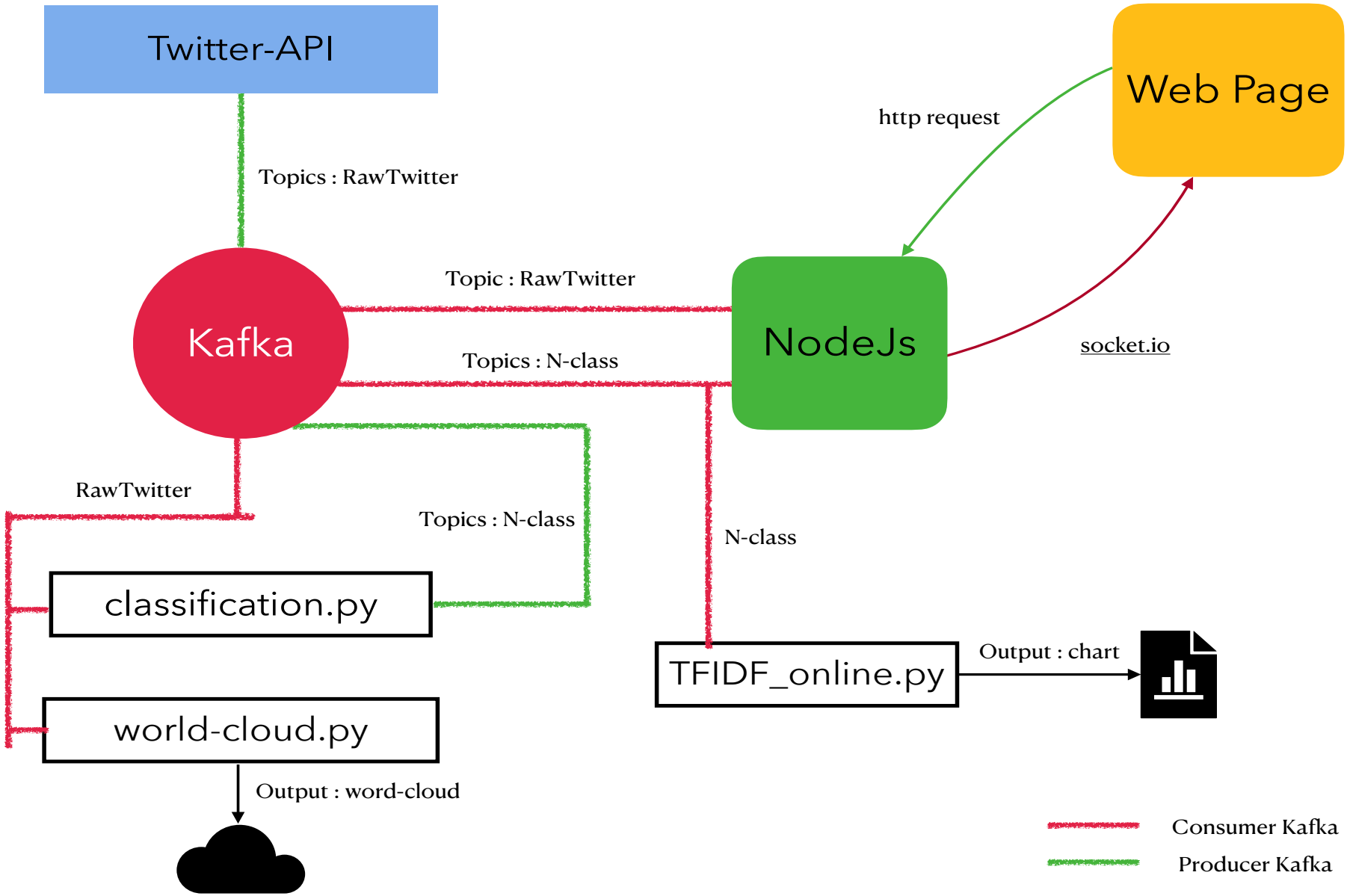
Objectifs :

- Identifier à quel topic un tweet (texte) appartient quand ce n'est pas toujours explicité dans le tweet

Intérêts de la modélisation online :

- Apparition de nouveaux trends, topics
- Adaptation du modèle à ces nouveaux trends

Architecture de l'application web



Classification

- Stream de tweets avec les queries : "politics, "manga", "health", "music", "school" (5 classes)
- Labelisation d'un tweet par identification de la query dans le tweet (méthode imparfaite)
- Embeddings pré-entrainés : TFIDF, BOW, BERT
- KNN

N(neighbors)	nb tweets	Embedding	Accuracy (%)
10	50K(∞)	BOW	64-65
100	50K (∞)	BOW	58-60
10	2K (∞)	BERT	78-84

Classification - Top words, TFIDF de chaque classe

Class 1- POLITICS				
politics	people	like	amp	know
0.054	0.014	0.013	0.011	0.009
Class 2 - MANGA				
man	manga	music	read	like
0.045	0.044	0.0019	0.018	0.016
Group 3 - HEALTH				
health	mental	good	care	amp
0.050	0.025	0.019	0.018	0.014
Group 4 -MUSIC				
music	love	like	new	video
0.059	0.017	0.014	0.012	0.011
Group 5 - SCHOOL				
school	high	like	old	kids
0.055	0.022	0.013	0.010	0.009

Clustering - Topic modelling

- Comparaison online VS batch
- Stream de tweets avec les queries : "politics, "manga", "health", "music", "school" (5 classes))
- Embeddings pré-entrainés : 'glove-twitter-200', word2vec_twitter_model
- Afin de determiner les topics le modèle consite a effectuer un clustering puis à resumer les classes à l'aide tu TF-IDF, LDA est une autre approche probabiliste pour la détection de topic.

Modèle	nb de tweets	répartition des clusters
Online Kmeans	∞	mauvaise et non discriminante
BatchKmeans	16000	parfois discriminante
LDA	16000	pas très discriminante

Clustering - MiniBatchKmeans

Group 1				
music	school	amp	health	biblebuild
0.032	0.019	0.011	0.010	0.010
Group 2				
bulan	unk	dir	testi	pinned
0.014	0.008	0.005	0.004	0.003
Group 3				
education	health	gpa	mental	care
0.149	0.100	0.084	0.030	0.027
Group 4				
flower	beautiful	unfold	bath	autumn
0.062	0.047	0.047	0.035	0.024

Table – TFIDF of the top words of each clusters, mini-batch clustering