



# Projet P9 - Parcours IA



Réalisez une application de recommandation de contenu



# Contexte

---



Recommandation de contenus de lectures pertinents

Construction de MVP : recommandation d'articles et livres aux particuliers

# Objectifs

---

**Réaliser un système de recommandation de contenu de lecture pertinent**

**Fonctionnalité la plus critique**

Recommandation de 5 articles pour l'utilisateur

**Contrainte critique pour l'architecture**

Prise en compte de l'ajout de nouveaux utilisateurs / articles

# Livrables

---

## **Prototype de recommandation de contenu**

Approches collaborative filtering &  
content-based

## **Déploiement Cloud**

Architecture Serverless

**Réflexion sur  
l'architecture pour  
intégration de nouveaux  
utilisateurs et articles**

# Jeu de données

# Jeu de données

## News Portal User Interactions by Globo.com

kaggle



clicks



articles\_metadata.csv



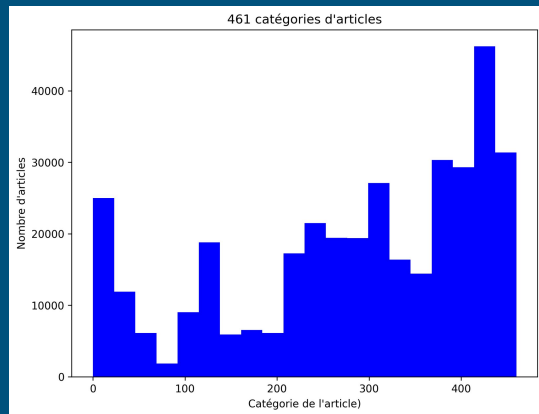
articles\_embeddings.pickle

- Informations sur les articles
- Comportements utilisateurs
- Interactions utilisateurs - articles

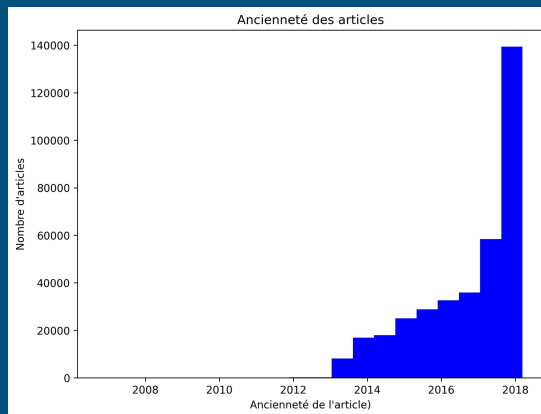
[Icons by Icons8](#)

# Jeu de données

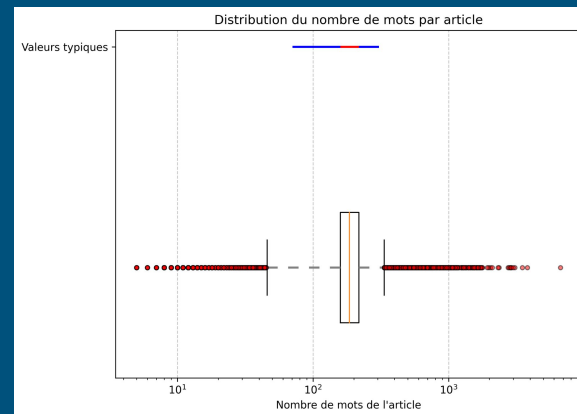
Dataset articles\_metadata : 364 047 articles



461 catégories d'articles



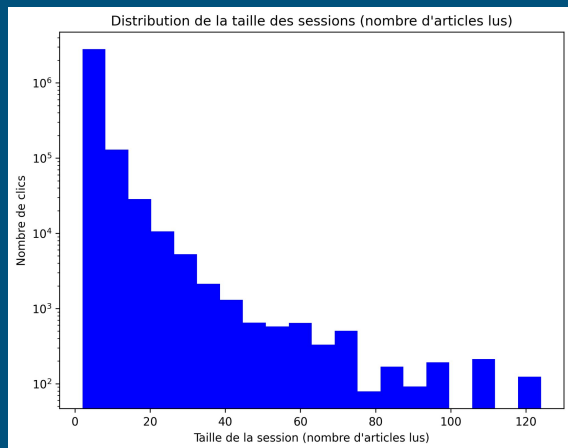
Articles publiés entre  
2012 et 2018



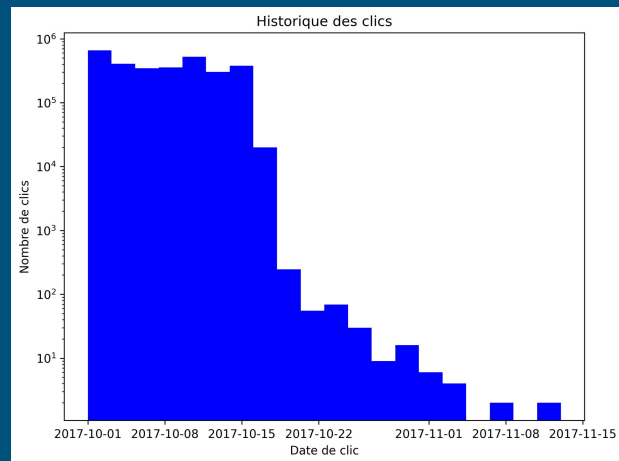
200 mots en moyenne

# Jeu de données

Dataset merged\_clicks : 2 988 181 clics, 322 897 utilisateurs, 46 033 articles consultés



Chaque utilisateur a consulté entre 2 et 124 articles (4 en moyenne)

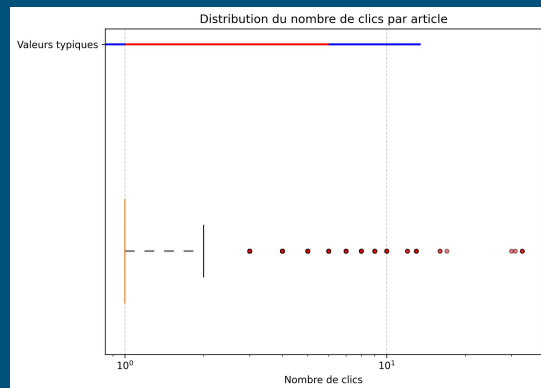
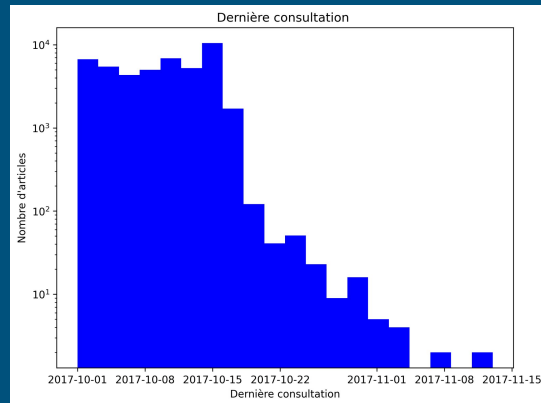


Date du premier article consulté : 2017-10-01  
Date du dernier article consulté : 2017-11-13



# Jeu de données

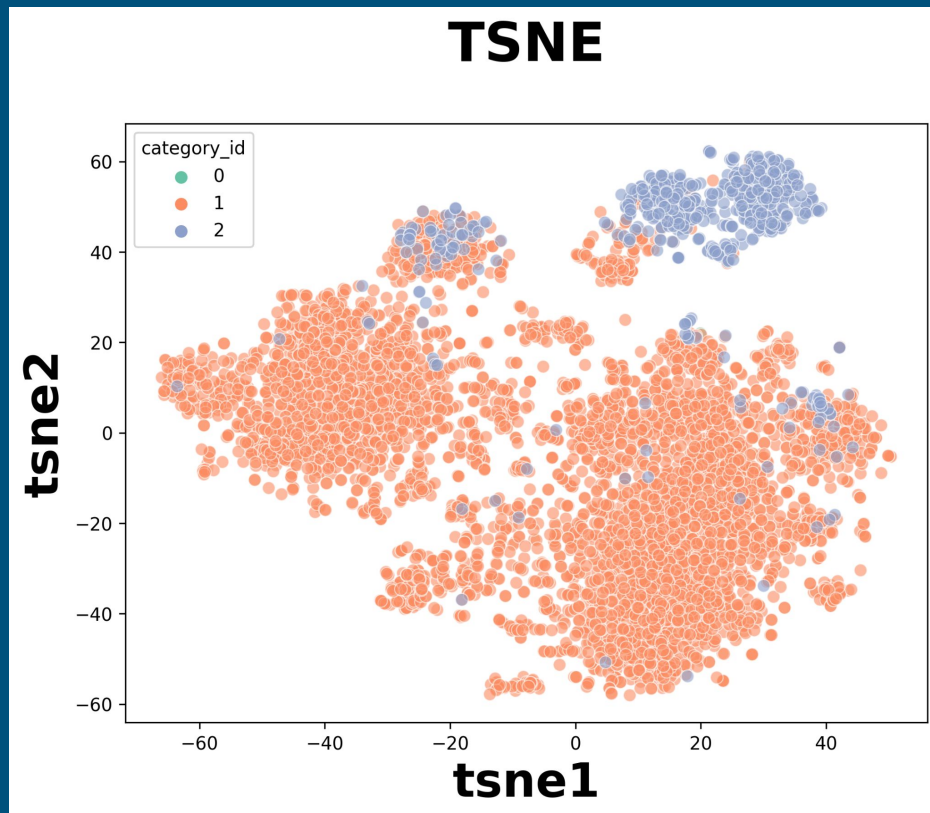
- La plupart des clics datent des 15 premiers jours sur les 43 enregistrés
- Chaque utilisateur a consulté 4 articles / cliqué 4 fois en médiane
  - 75% des utilisateurs ont réalisé 10 clics au maximum
  - La taille de session médiane par utilisateur est de 3
- Seuls 13% des articles ont été consultés
  - Plus de 50% des articles lus n'ont été lus que par un seul lecteur
- Pour plus de 75% des articles consultés les utilisateurs ne cliquent qu'une fois
  - Les utilisateurs cliquent au maximum 33 fois par article



# Jeu de données

## Matrice d'embedding

- 364 047 embeddings d'articles de dimension 250
- Représentation tsne des 5000 premiers articles
- Les articles semblent classés par catégorie



# Approches de recommandations

# Approches de recommandations

---

## Content-based

Recommandation d'**articles similaires aux articles consommés** par le lecteur

Préférences du lecteur calculées selon ses **lectures précédentes**

## User-based collaborative filtering

Recommandation d'articles basées sur les **lectures de lecteurs similaires**

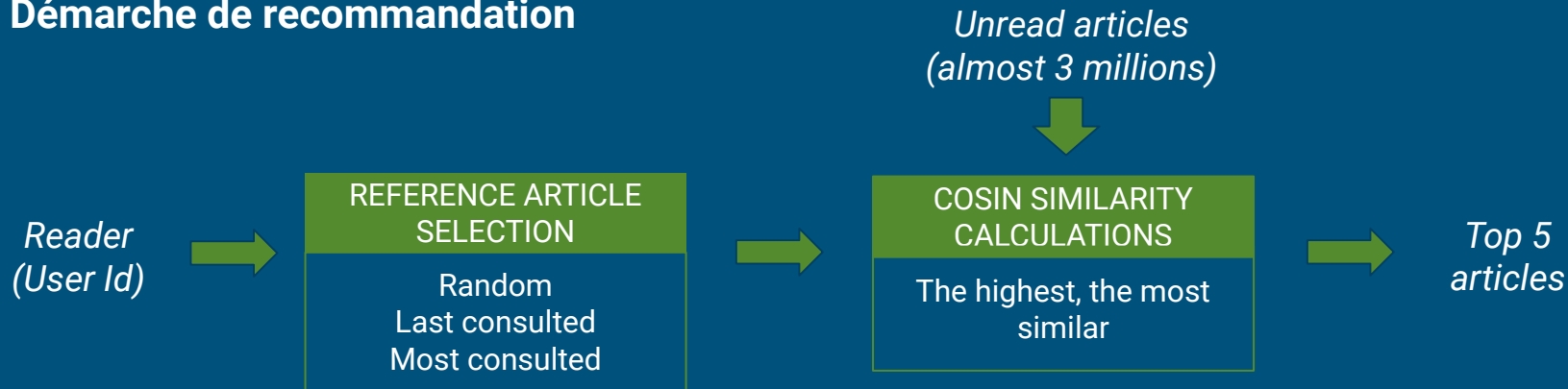
Préférences du lecteur calculées selon celles **des lecteurs similaires**

# Approche content-based

## Mesure de cosin-similarity entre deux articles

- Mesure de similarité entre deux articles représentés par des vecteurs d'embedding
- Matrice d'embeddings : près de 3 millions d'articles dans espace vectoriel de dimension 250

## Démarche de recommandation



# Approche content-based

## Recommandations pour le lecteur 140 711 basées sur un article lu au hasard

SELECTED USER  
User id : 140711

--  
REFERENCE ARTICLE USED FOR RECOMMENDATION  
Article id: 284547

	article_id	category_id	words_count	creation date
	284547	412	182	2017-10-03

--  
OTHER ARTICLES READ BY THE USER

	article_id	category_id	words_count	creation date
	71076	136	240	2017-10-04
	174236	299	251	2017-10-04
	271551	399	224	2017-10-03
	284547	412	182	2017-10-03

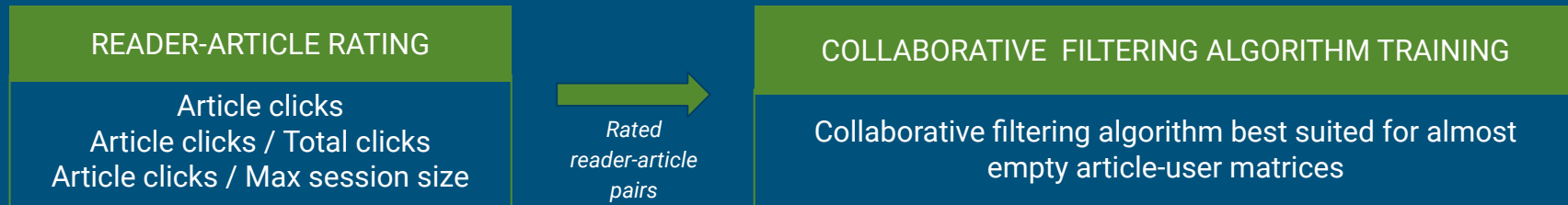
--  
RECOMMENDED ARTICLES : [279547, 285342, 285435, 286161, 286351]

	article_id	category_id	words_count	creation date
	279547	412	200	2015-02-10
	285342	412	202	2017-10-03
	285435	412	237	2017-10-03
	286161	412	239	2017-10-03
	286351	412	156	2017-05-02

Articles majoritairement, consécutifs de même catégorie et taille

# Approche collaborative filtering

## Note lecteur-article et entraînement de modèle

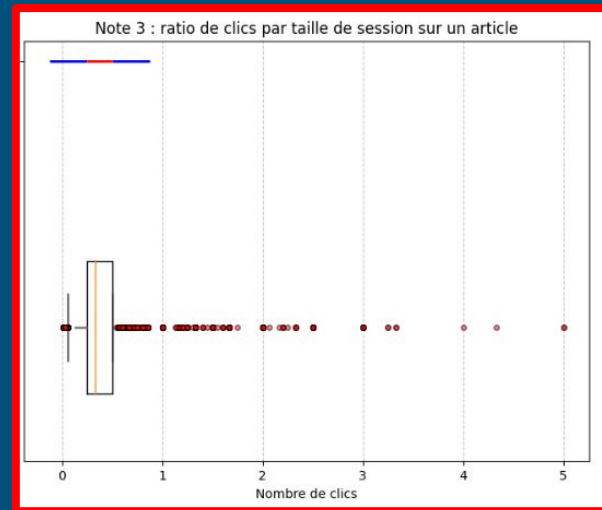
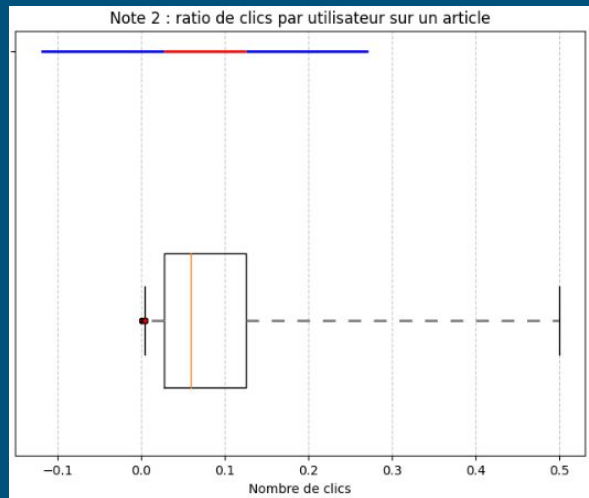
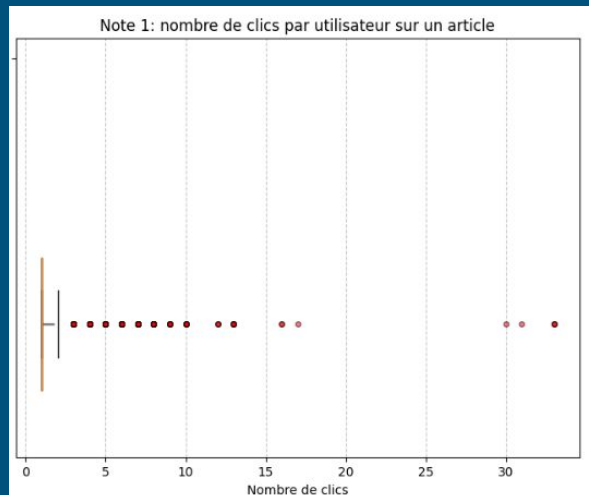


## Démarche de recommandation



# Approche collaborative filtering

Le ratio de clics par taille maximale de session possède la meilleure variabilité



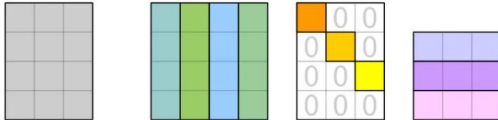


# Approche collaborative filtering

## Choix de l'algorithme SVD (Singular Value Decomposition)

- Matrice lecteur-article quasiment vide (0.02% de remplissage)
- L'utilisation de la factorisation matricielle permet d'améliorer les performances de l'algorithme

```
Nombre d'utilisateurs : 322897  
Nombre total d'articles consultés : 46033  
Nombre de paires utilisateur-article évaluées : 2.95 millions  
Taille de la matrice user-article : 14.86 milliards  
Taux de remplissage de la matrice user-article : 0.02 %
```


$$\begin{matrix} \mathbf{M} \\ m \times n \end{matrix} = \begin{matrix} \mathbf{U} \\ m \times m \end{matrix} \begin{matrix} \mathbf{\Sigma} \\ m \times n \end{matrix} \begin{matrix} \mathbf{V}^* \\ n \times n \end{matrix}$$

# Approche collaborative filtering

---

## Réglage du modèle et des hyperparamètres

- Rating scale : (0, 5) soit l'échelle de valeurs de la note **clics\_by\_session**
- GridSearchCV
  - n\_factors : [50, 100, 140]
  - n\_epochs : [10, 20, 50]
  - lr\_all : [0.002, 0.005, 0.007]
  - reg\_all : [0.01, 0.02, 0.03]
  - CV = 3 folds
  - Evaluation : métriques RMSE et MAE

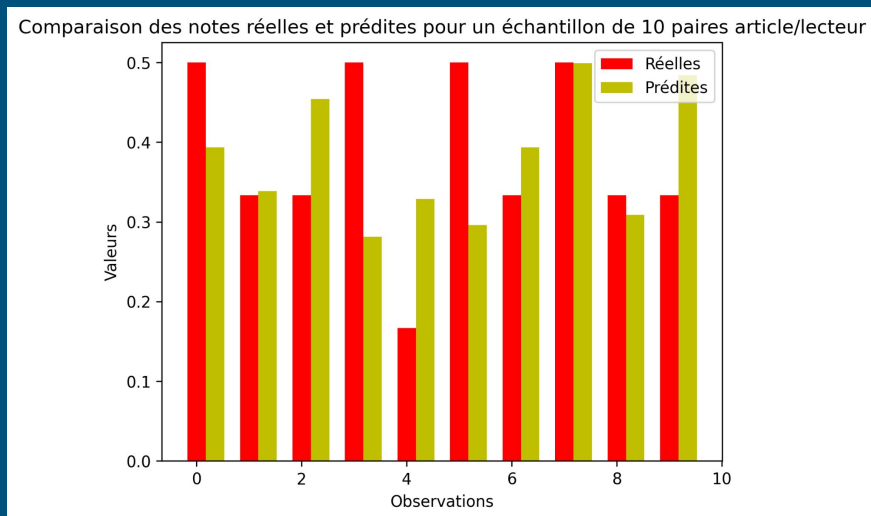
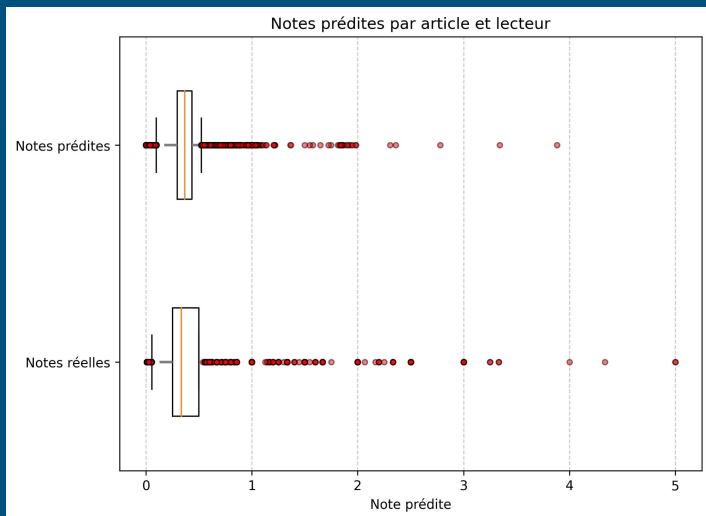
Les valeurs des notes prédites vs. réelles sont visualisées pour quelques paires lecteur-article choisies.

# Approche collaborative filtering

## Modèle final

n\_factors : 20, n\_epochs : 70, lr\_all : 0.009, reg\_all : 0.01

RMSE : 0.12, MAE : 0.10, Accuracy : 0.12



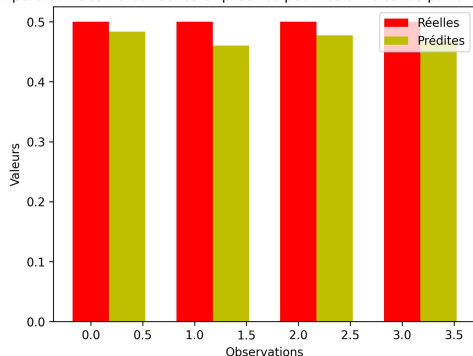
# Approche collaborative filtering

## Recommandations pour le lecteur 140 711 basées sur la note *clicks\_by\_session*

SELECTED USER : READ ARTICLES  
User id : 140711

	article_id	category_id	words_count	creation date
1	71076	136	240	2017-10-04
2	174236	299	251	2017-10-04
3	271551	399	224	2017-10-03
4	284547	412	182	2017-10-03

Comparaison des notes réelles et prédites pour les articles lus par un utilisateur



--  
RECOMMENDED ARTICLES : [69463 73431 70095 68851 25058]

	article_id	category_id	words_count	creation date
1	69463	136	172	2017-08-21
2	73431	138	183	2017-01-13
3	70095	136	293	2017-09-26
4	68851	136	278	2017-07-30
5	25058	25	133	2014-04-03

Articles non consécutifs, parfois peu ressemblants à ceux consultés

# Synthèse

## Content-based

	article_id	category_id	words_count	creation date
	279547	412	200	2015-02-10
	285342	412	202	2017-10-03
	285435	412	237	2017-10-03
	286161	412	239	2017-10-03
	286351	412	156	2017-05-02

Plus grand choix d'articles (articles non lus par d'autres lecteurs)

## Articles lus par l'utilisateur 140 711

	article_id	category_id	words_count	creation date
	71076	136	240	2017-10-04
	174236	299	251	2017-10-04
	271551	399	224	2017-10-03
	284547	412	182	2017-10-03

## Collaborative filtering

	article_id	category_id	words_count	creation date
	69463	136	172	2017-08-21
	73431	138	183	2017-01-13
	70095	136	293	2017-09-26
	68851	136	278	2017-07-30
	25058	25	133	2014-04-03

Choix d'articles limité aux lectures d'autres utilisateurs

**Une hybridation des deux systèmes permettrait une recommandation plus large et plus originale aux lecteurs**

Les articles proposés restent similaires à l'historique du lecteur (quantité, variété)

Plus grande originalité dans les articles proposés (goûts des autres lecteurs)

# Application Serverless

# Application Serverless

## Architecture choisie

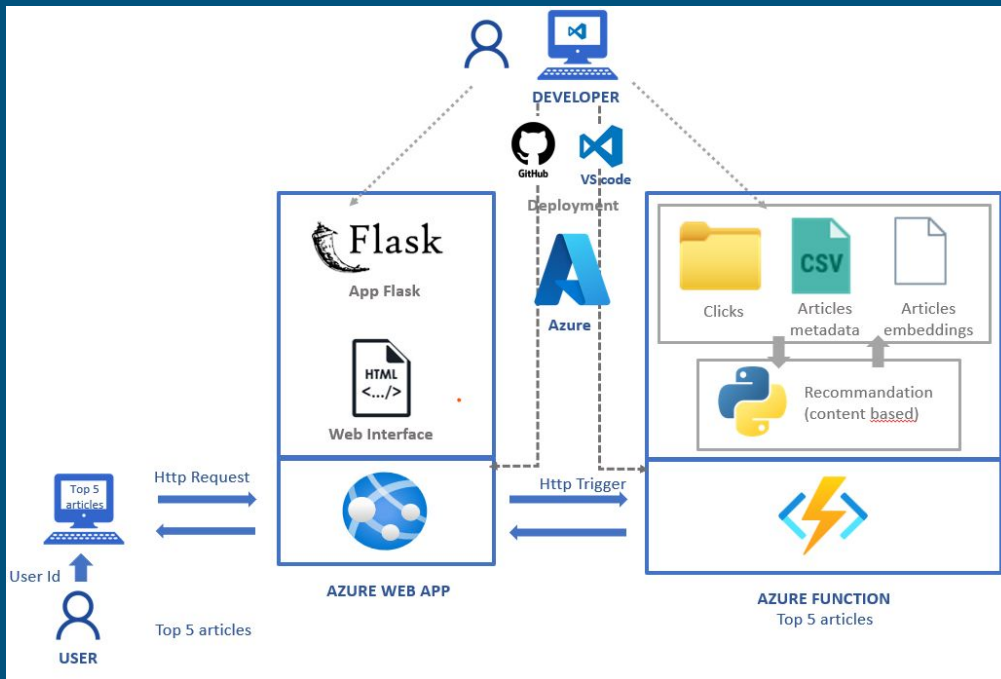
Application Azure Web App  
+  
Azure Function

### URL Azure Function

<https://p9testapp.azurewebsites.net/api/httpexample>

### URL Application Web

<http://p9recodashboard.azurewebsites.net/>



# Application Serverless

## Architecture cible

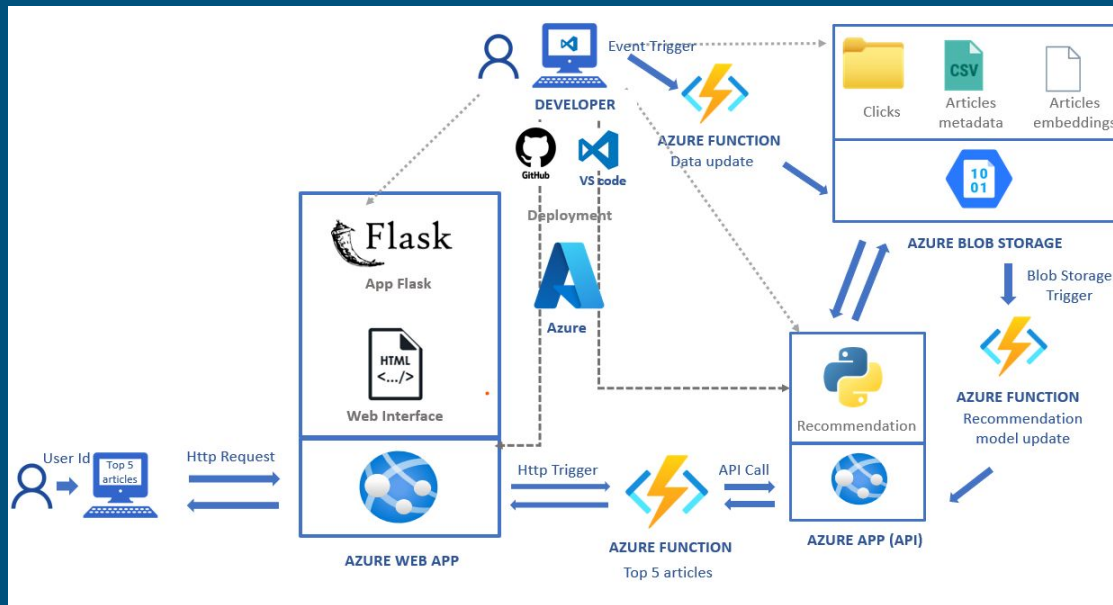
Azure Web App (Web interface & API)

+

Azure Functions

+

Azure Blob Storage





# Conclusions et Perspectives

---

## Systèmes de recommandation

- Le modèle content-based a l'avantage de ne pas être limité aux articles lus par l'ensemble des lecteurs, mais il ne recommande que des articles similaires à ceux déjà consultés
- Le modèle basé sur une approche de filtrage collaboratif permet de recommander des contenus plus originaux pour le lecteur, puisqu'ils sont basés sur les goûts de lecteurs similaires. Il ne permet toutefois pas d'accéder à des articles déjà non consultés
- Une solution plus intéressante serait de combiner ces deux approches, permettant ainsi au lecteur d'accéder à la fois à une offre de contenu plus large mais aussi plus originale

## Architecture Serverless

- La mise en place d'une architecture Serverless avec des Azure functions permet d'optimiser les coûts liés à la fréquence relative des différentes requêtes dans une application
- La mise à jour du jeu de données et du modèle de recommandation peut être automatisée par l'ajout d'Azure functions déclenchées par des événements prédéfinis

# ANNEXES

---

# Jeu de données - Questions A MAJ

---

- Quel est le nombre d'articles lus sur la base ?
- Quel est le nombre d'articles lus par utilisateur ?
- Par combien d'utilisateurs chaque article est-il lu ?
- Comment classer les articles pour un utilisateur ?
  - Récence
  - Nombre de clics par article

# Jeu de données - Merge par user\_id

**Chaque utilisateur a consulté 4 articles /cliqué 4 fois en médiane**

- 75% des utilisateurs ont réalisé 10 clics au maximum
- La taille de session médiane par utilisateur est de 3

	total clics	largest session size	last activity	total articles
user_id				
0	8	2	2017-10-09 01:54:46.617	8
1	12	2	2017-10-17 01:01:09.398	12
2	4	2	2017-10-17 01:15:09.725	4
3	17	4	2017-10-12 03:19:55.593	17
4	7	3	2017-10-10 03:15:20.003	7
5	87	10	2017-10-16 22:19:20.852	84

	total clics	largest session size	total articles
count	322897.000000	322897.000000	322897.000000
mean	9.254285	3.266156	9.138239
std	14.946358	2.147438	14.344677
min	2.000000	2.000000	2.000000
25%	2.000000	2.000000	2.000000
50%	4.000000	3.000000	4.000000
75%	10.000000	4.000000	10.000000
max	1232.000000	124.000000	1048.000000

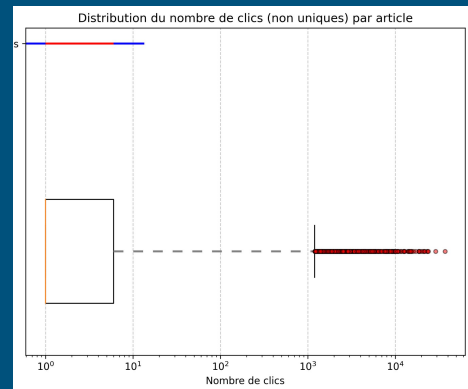
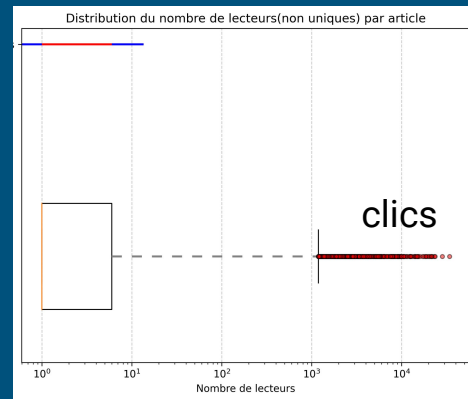
# Jeu de données - Merge par article\_id

Seuls **13% des articles ont été consultés**

- Plus de 50% des articles consultés n'ont été consultés qu'une seule fois (médiane 1)
- Plus de 50% des articles lus n'ont été lus que par un seul lecteur (médiane 1)

	total clics	last consultation	total readers
click_article_id			
3	1	2017-10-09 18:06:18.399	1
27	1	2017-10-04 19:21:06.293	1
69	1	2017-10-12 18:14:29.718	1
81	2	2017-10-02 19:56:55.111	2
84	1	2017-10-09 14:48:05.152	1

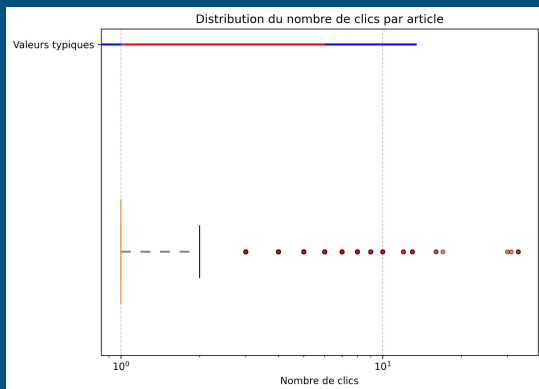
	total clics	total readers
count	46033.000000	46033.000000
mean	64.913888	64.099885
std	629.322888	617.366810
min	1.000000	1.000000
25%	1.000000	1.000000
50%	1.000000	1.000000
75%	6.000000	6.000000
max	37213.000000	34145.000000



# Jeu de données - Merge user-article

Pour plus de 75% des articles consultés les utilisateurs ne cliquent qu'une fois

- Les utilisateurs cliquent au maximum 33 fois par article



		clics per article	last click date
user_id	click_article_id		
0	68866	1	2017-10-01 03:00:58.020
	87205	1	2017-10-09 01:54:46.617
	87224	1	2017-10-09 01:54:16.617
	96755	1	2017-10-07 01:33:53.155
	157541	1	2017-10-01 03:00:28.020
	160158	1	2017-10-07 14:54:50.915
	233470	1	2017-10-07 14:55:20.915
	313996	1	2017-10-07 01:34:23.155
	36162	1	2017-10-17 01:00:39.398
	59758	1	2017-10-01 17:36:36.845
1	96663	1	2017-10-01 03:04:07.951

		clics per article
count	2.950710e+06	
mean	1.012699e+00	
std	1.344699e-01	
min	1.000000e+00	
25%	1.000000e+00	
50%	1.000000e+00	
75%	1.000000e+00	
max	3.300000e+01	