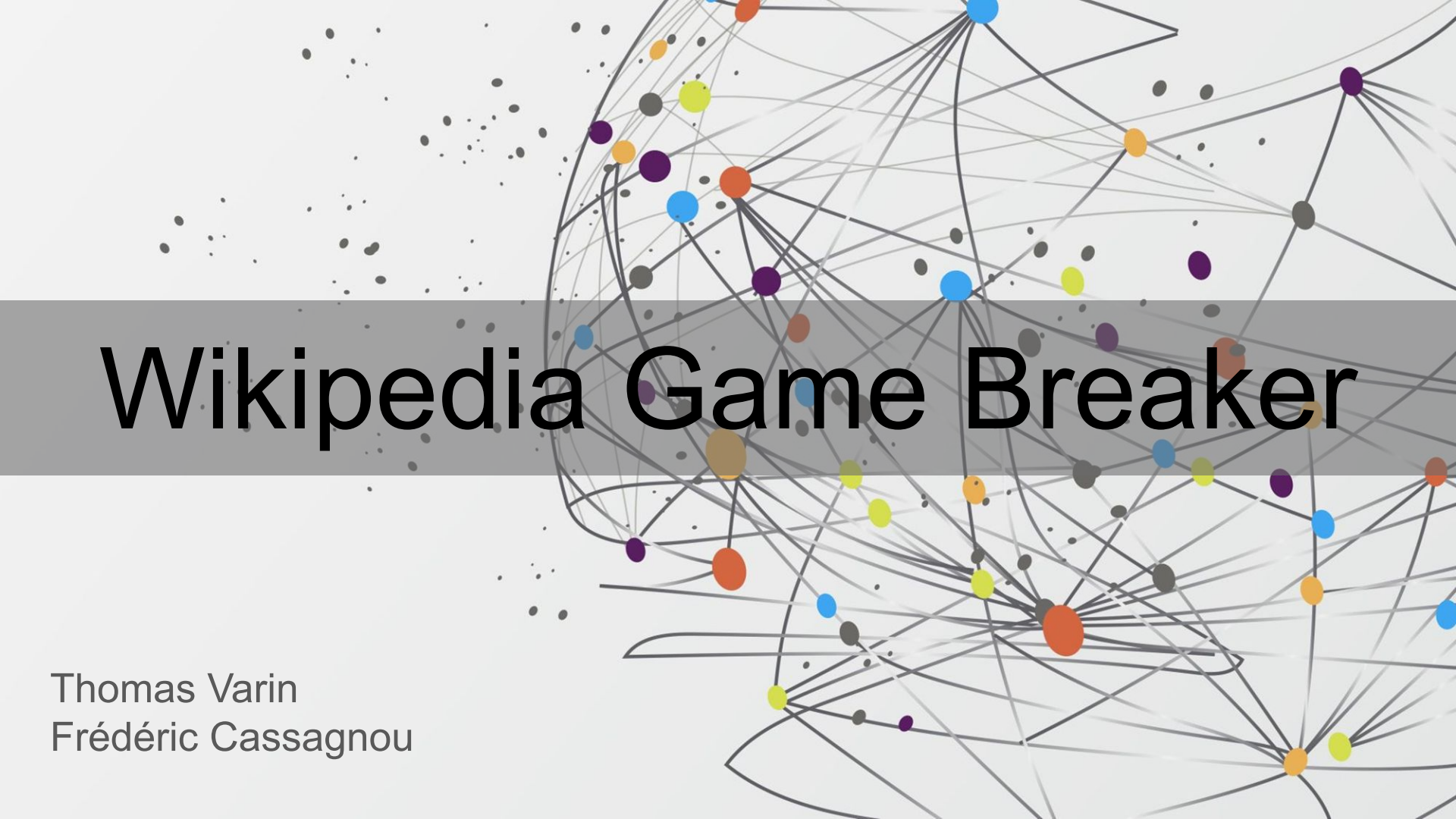


# Wikipedia Game Breaker

Thomas Varin
Frédéric Cassagnou

# Présentation du projet

**Wikipédia :**

- Encyclopédie gratuite en ligne
- Une encyclopédie par langue, avec ses propres règles
- Chaque article peut en citer d’autres

**Jeu de Wikipédia :**

- Passer d’une page à une autre en utilisant uniquement les liens bleus

# Exemples d’articles


# Estatistiques

| Page d'estatistiques | |
|---|---|
| Noumbe d'articlles | 5 031 |
| Pages<br>(Toutes les pages du wiki, y compris les pages de discussion, les redirections, etc.) | 10 638 |
| Fichiers téléversés | 0 |
| **Estatistiques d's éditiouns** | |
| Modifications de pages depuis l'installation de Wikipedia | 220 731 |
| Nombre moyen de modifications par page | 20,75 |
| **Estatistique des féchouneus** | |
| Utilisateurs inscrits (liste des bénoums) | 14 229 |
| Utilisateurs actifs (liste des bénoums)<br>(Utilisateurs ayant fait au moins une action durant les 30 derniers jours) | 17 |
| Robots (liste des bénoums) | 23 |
| Counétablles d'ouovrage (liste des bénoums) | 2 |
| Administrateurs d'interface (liste des bénoums) | 0 |
| Bureaucrates (liste des bénoums) | 0 |
| Masqueurs de modifications (liste des bénoums) | 0 |
| Stewards (liste des bénoums) | 0 |
| Créateurs de comptes (liste des bénoums) | 0 |
| Importateurs (liste des bénoums) | 0 |
| Importateurs transwiki (liste des bénoums) | 0 |
| Exemptés de blocage IP (liste des bénoums) | 0 |
| Vérificateurs d'utilisateurs (liste des bénoums) | 0 |
| Utilisateurs bloqués de l'outil d'informations sur une IP (liste des bénoums) | 0 |
| Utilisateurs confirmés (liste des bénoums) | 0 |
| **Autres statistiques** | |
| Mots dans toutes les pages de contenu | 429 394 |

# Objectifs

Crawler un Wikipédia complet

Trouver un chemin optimal entre toutes les pages Wikipédia


Trouver les cas impossibles

Interpréter les résultats pour définir des stratégies


Jouer avec les poids pour se rapprocher des vraies conditions

# Méthodologie et outils

- **Python 3.10**
- **Crawler** “homemade” en Python
- **Pyspark(SQL)**
- **Pickle** pour stocker les données (pourra être remplacé)
- **Matrice d’adjacence** pour le graphe

# Résultats

Distances depuis chaque article

| index | max_distance |
|---|---|
| 0 | 11 |
| 1 | 10 |
| 2 | 11 |
| 3 | 11 |
| 4 | 11 |
| 5 | 10 |
| 6 | 11 |
| 7 | 11 |
| 8 | 11 |
| 9 | 11 |

Distances from Oûtriche

| dist | count |
|---|---|
| -1 | 855 |
| 0 | 1 |
| 1 | 58 |
| 2 | 517 |
| 3 | 2558 |
| 4 | 1425 |
| 5 | 233 |
| 6 | 62 |
| 7 | 21 |
| 8 | 4 |
| 9 | 1 |
| 10 | 5 |

Distances from 1322

| dist | count |
|---|---|
| -1 | 855 |
| 0 | 1 |
| 1 | 25 |
| 2 | 620 |
| 3 | 2175 |
| 4 | 1612 |
| 5 | 354 |
| 6 | 71 |
| 7 | 16 |
| 8 | 4 |
| 9 | 1 |
| 10 | 1 |
| 11 | 5 |

| max(max_distance) |
|---|
| 11 |

# Résultats

Nombre de pages atteignables

| Name | Reachable |
|---|---|
| Câtel_de_Vinchennes | 4892 |
| Edmund_Blampied | 4885 |
| 1416 | 4885 |
| Rimeus | 4885 |
| 1382 | 4885 |
| Maûsolée | 4885 |
| 1656 | 4885 |
| Bailliage_dé_Gùer... | 4885 |
| 24_Janvyi | 4885 |
| 1952 | 4885 |
| Jèrri | 4885 |
| 27_Novembre | 4885 |
| Ésope | 4887 |
| 26_Août | 4885 |
| Ph'lippe_Duron | 4885 |
| Berlanga_de_Duero | 4886 |
| 25_Juilet | 4885 |
| 777 | 4885 |
| Nouormandie | 4885 |
| Quien | 4885 |

# Résultats

## Chemins les plus courts entre deux articles

```
5 ['Crouésade', 'Jèrriais', 'Langue', 'Laungue_construite', 'Angllais']
No path from Clloque to Abejar
3 ['757', 'IXe_s.', '854']
5 ['Ouothelle_(frouque)', 'Page_dé_garde', 'Giographie', 'Liste_des_pays_du_monde', 'Kosovo']
4 ['Afrique-du-Su', 'Monnaie', 'Liste_des_dus_de_Normaundie', '927']
4 ['Tokyo', 'Japon', 'Liste_des_pays_du_monde', 'Irlaunde']
4 ['1890', 'XIXe_s.', 'Janvyi', '6_Janvyi']
4 ['Annaées_1360', 'Page_dé_garde', 'Fraunce', 'Crna_Gora']
4 ['Cannot', 'Page_dé_garde', 'Mouogeâle', 'Viande']
4 ['1645', 'Normaundie', 'Semanne', 'Djödi']
5 ['Trochet', 'Page_dé_garde', 'Aungllèterre', '1603', 'Annaées_1620']
6 ['Cité', 'Ville', 'Liste_des_capitales', 'Muscade', 'Épice', 'Moutarde']
4 ['Histouère', 'Page_dé_garde', 'Baêtes', 'Mammifère']
3 ['Langue_romanne', 'Laungue', 'Français']
7 ['Mongolie', 'Page_dé_garde', 'Mouogeâle', 'Pain', 'Sexualité', 'Contraception', 'Capote']
4 ['Scooby-Doo', 'Améthique', 'Ûrope', 'Mathe']
4 ['1623', 'Page_dé_garde', 'Islam', 'Jésû']
6 ['Capote', 'Page_dé_garde', 'Islam', 'Jésû', 'Fei', 'Dieu']
4 ['1905', 'Page_dé_garde', 'Afrique', 'Vèrt_Cap']
3 ['Sŷmbole', 'Ûro', 'Espangne']
```

# Défis rencontrés

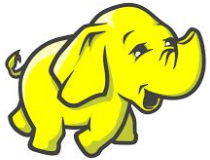
**Pas de dataset** public de taille raisonnable

Le crawling est très **long** et **incertain**

Messages d’erreurs de **Java**

# Défis rencontrés

# Pistes d’amélioration



Utiliser HDFS pour le stockage distribué

Remplacer pickle par Parquet

Utiliser un Wikipédia plus grand