Vers une base de connaissances africaine des données journalistiques (articles et commentaires)

Expérimentation sur le projet Check4Decision

Dr Edouard Ngor SARR
Université Assane SECK de Ziguinchor-UASZ
Check4Decision Project
CERIDES UCAO

Du 25 au 27 Octobre 2022 à COTONOU -BENIN

Extraction & Agrégation & la réutilisation de données journalistiques issues de sources web ?

Check4Decision Research Project

Présentation

Projet de recherche scientifique basé à l'Université de THIES au Sénégal : http://check4decision.univ-thies.sn/

- Un consortium de 04 universités
- Un financement initial du CEA MITIC

Equipe de Recherche et Partenaires

Prof. Ousmane SALL
Enseignant-Chercheur

Dr. Babiga BIRREGAH
Enseignant-Chercheur

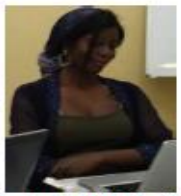
Dr. Edouard Ngor SARR
Enseignant-Chercheur

Prof. Mamadou Bousso
Enseignant-Chercheur

Dr. Marie NDIAYE DIOP
Enseignante-Chercheure

Sény N. Mbaye (UT)

Mme. Senghor (UCAO)



Aminata MAIGA (UT)

Lamine FATY (UASZ)

Mory M.Koulibaly(AIMS)

Tony Tona Landau (AIMS)

Moustapha Sissoko (UT)

Abdou L. Péne (UT)

Check4Decision Research Project

A l'origine (2017): Fact-checking

Google images

TinEye

Amnesty International

Youtube DataViewer

Google Fact Check Tools

France 24 The Observers

Africa Check

- Journal *Le Monde* en France (https://www.lemonde.fr/verification/)

ClaimBuster

- **Du NFS aux Etats Unis**
- Un modèle d'IA pour détecter automatiquement les affirmation à vérifier.
- https://idir.uta.edu/claimbuster/

InVID

In video veritas

InVid et CrossCheck lancés par le réseau de fact-checking First Draft News (https://fr.firstdraftnews.com/) et financés par Google News Lab. http://invid.condat.de/

	Mode d'utilisation	Type de données en entrée				
	Extension navigateur	Plateforme en ligne	Image	Audio/Vidéo	Email	Texte
VERA		✓				✓
FactMinder		✓				✓
Claimbuster		✓				✓
Truth Teller		✓		✓		✓
MapChecking		✓	✓			
CrossCheck		✓				✓
Hoaxbuster		✓			✓	
Decodex	✓					✓
Check News		✓				✓
Google Images		✓		✓		
TinEye		✓		✓		
YouTube DataViewer		✓		✓		
FullFact	✓		✓	✓	✓	✓
LazyTrusth		✓			✓	
Holy Grail		✓		✓		✓
Trooclick	✓					✓

Full Fact fights bad information

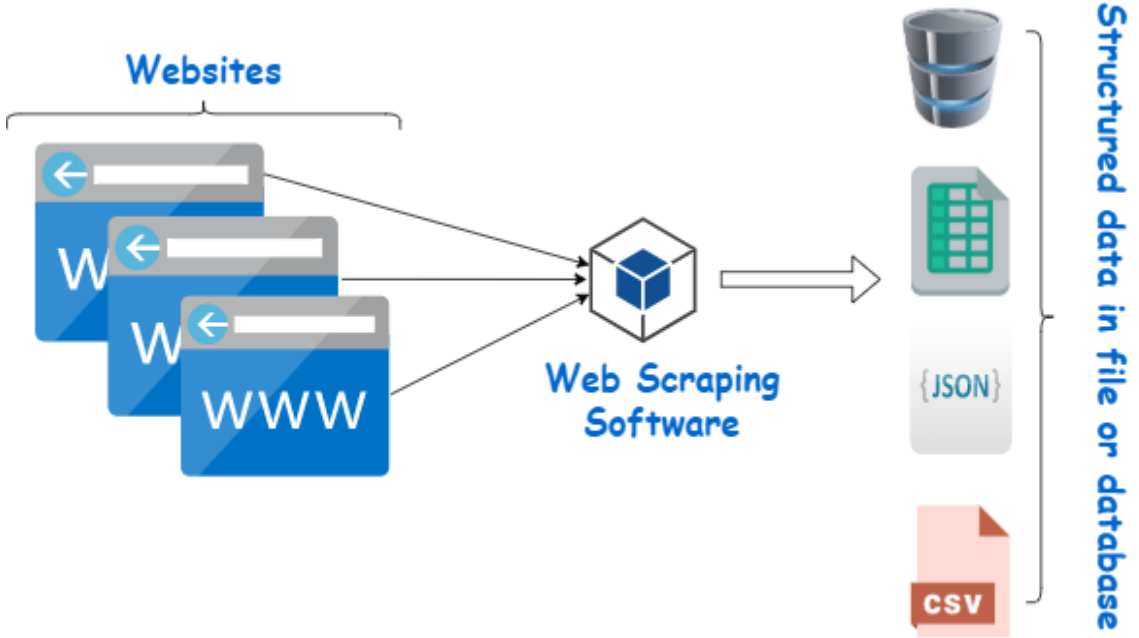
fullfact.org

https://fullfact.org

Check4Decision Research Project

Acquisition des données par ML

- Web Crawling:
  - Action d'indexer toute l'information d'une page web pour y avoir accès plus facilement.
- Web Scraping :
  - Action d'extraire / récupérer des données à partir de sites web, grâce à un programme ou un logiciel.
  - Automatisation intelligente du processus dans lequel une personne réaliserait des copier/coller manuels

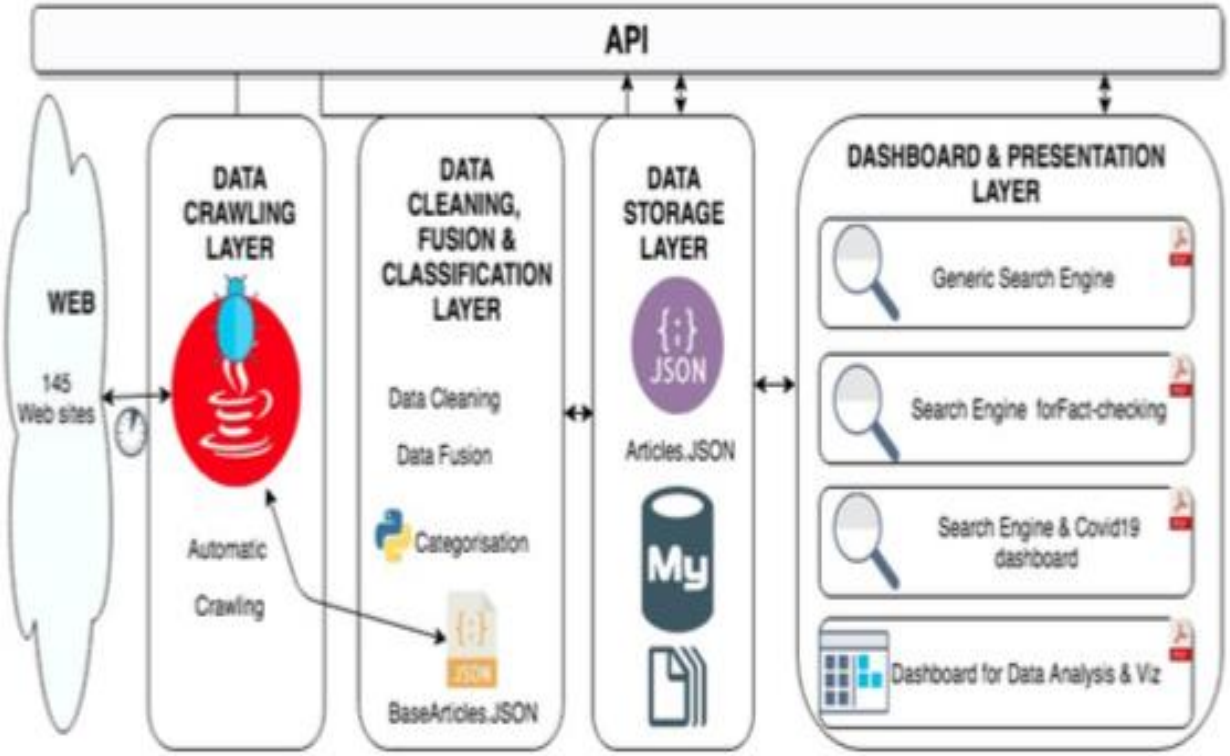


Utilisation:

- Moteurs de recherche
- Comparateurs de prix
- Veille concurrentielle

Check4Decision Research Project

Acquisition des données par ML



Activités

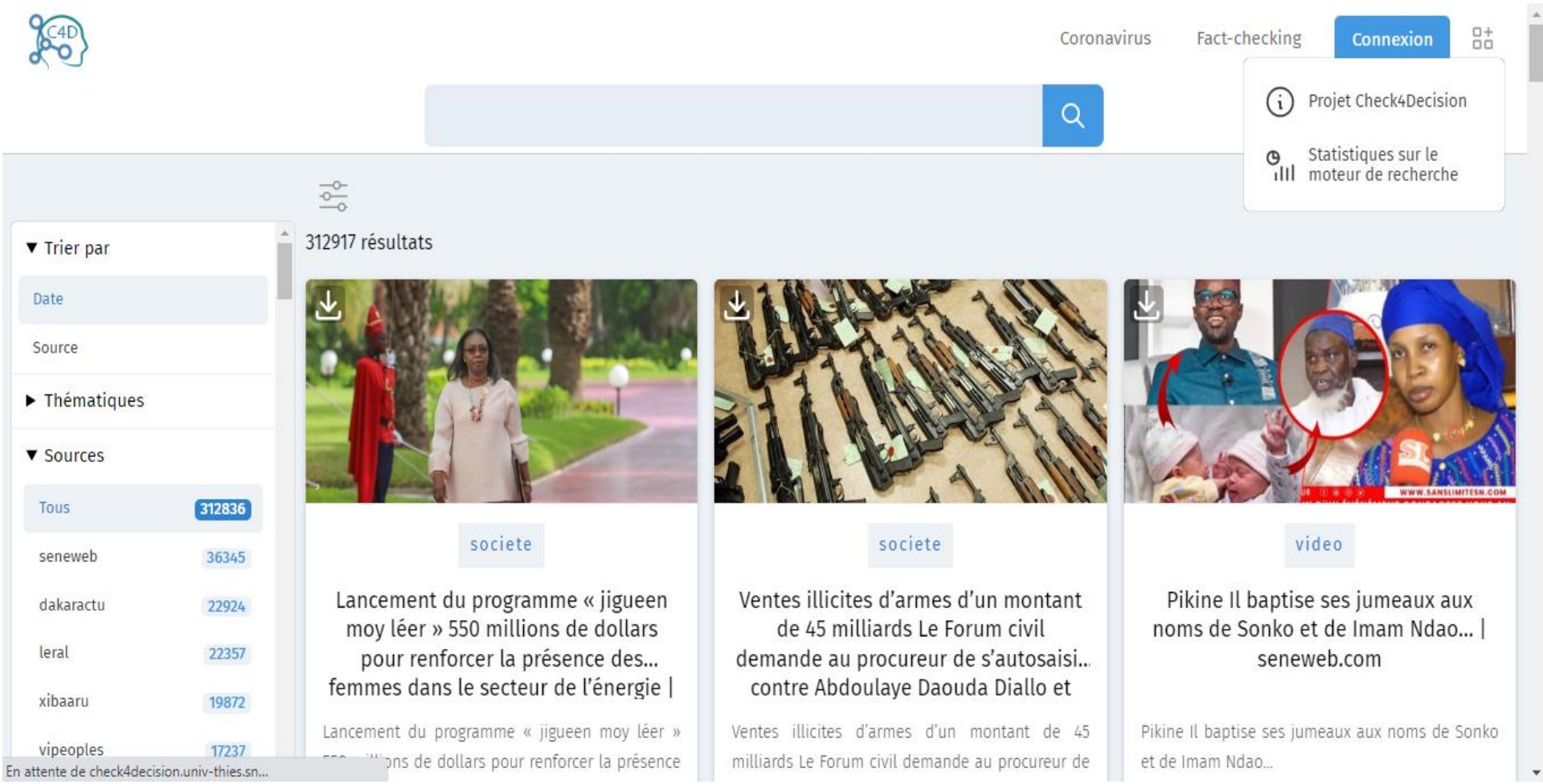
- Indexation des sources (Web Crawling)
- Extraction des articles & commentaires (web Scraping)
- Formatage et uniformisation (Hétérogénéité)
- Structuration et stockage (Format Leger : CSV, JSON, BDR)
- Catégorisation des données (Machine Learning)
- Analyse et Présentation (Moteur de recherche, Tableau de bord ..)

Check4Decision Research Project

Notre Corpus

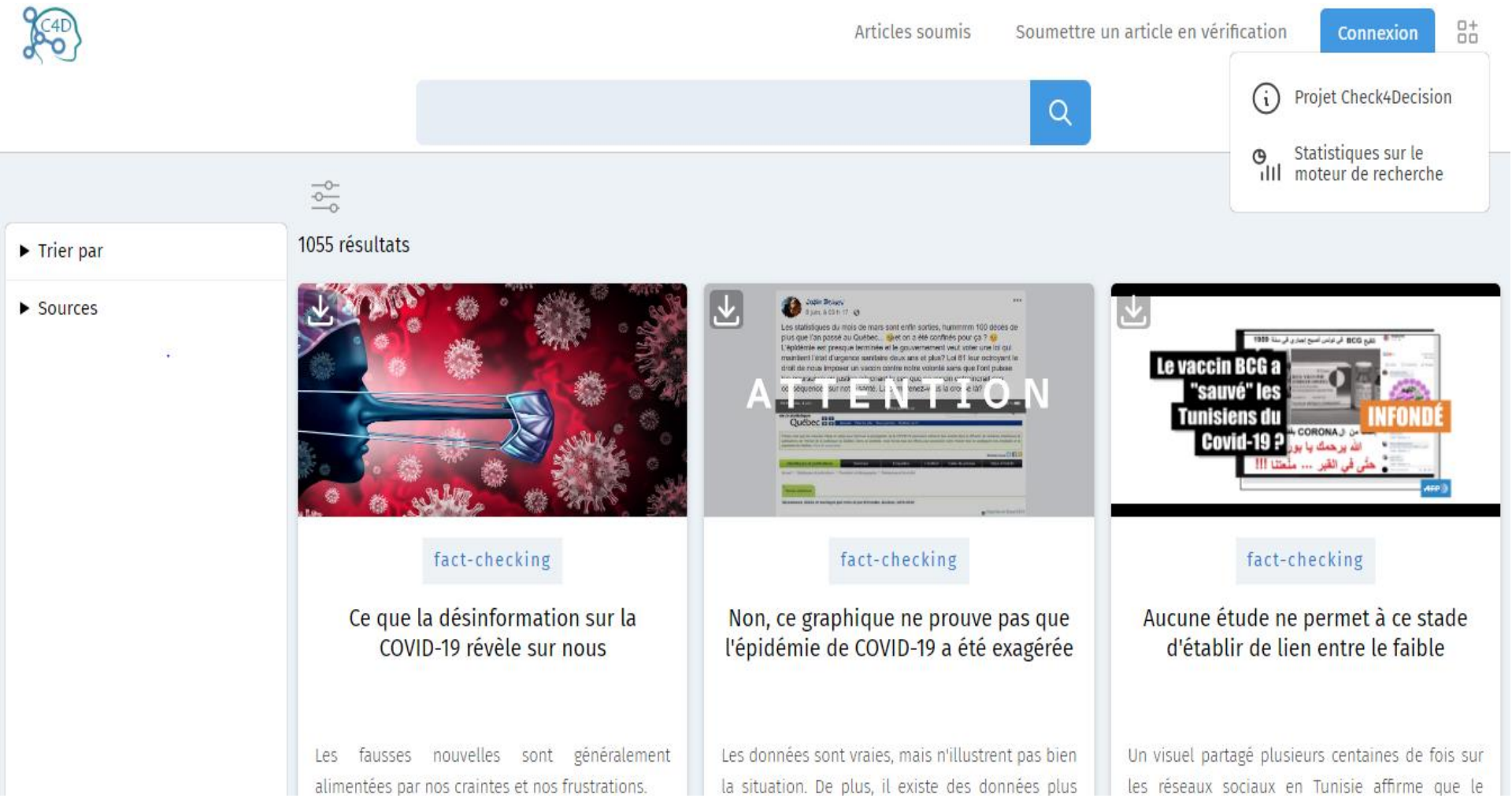
- Lieu de Stockage : Serveur de UIDT (SSH, FTP & BD)
- Type de sources: Sites Web / Sénégalais
- Type de données : Articles et commentaires
- Format de données : Texte (CSV, JSON ou BDR) et Image (JPG & PNG)
- Nombre de sources : 111
- Fréquence de la collecte: 3 par jours
- Taille du corpus:
  - Articles: +500 000
  - Commentaires: + 10 000 000
- Mode d'accès: OPEN
- Exploitation des données : Plateforme & Tableau de bord
  - Moteur de recherche:
    - https://check4decision.univ-thies.sn/search/
  - Tableau de bord:
    - https://check4decision.univ-thies.sn/search/dashboard/articles

Projets Check4Decison
Résultats: Moteur de recherche



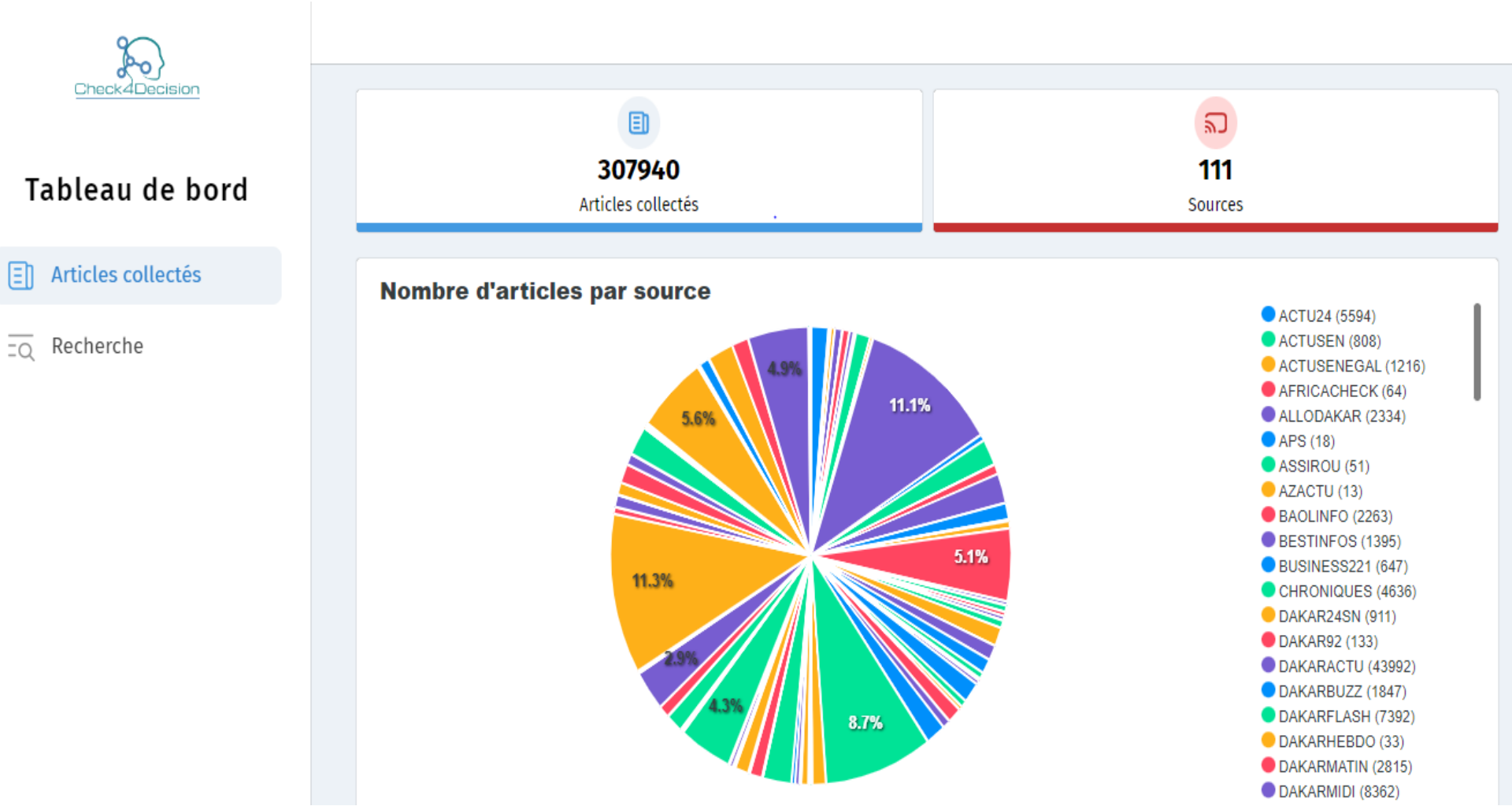
https://check4decision.univ-thies.sn/search

Projets Check4Decison
Résultats: Moteur de Fact-checking



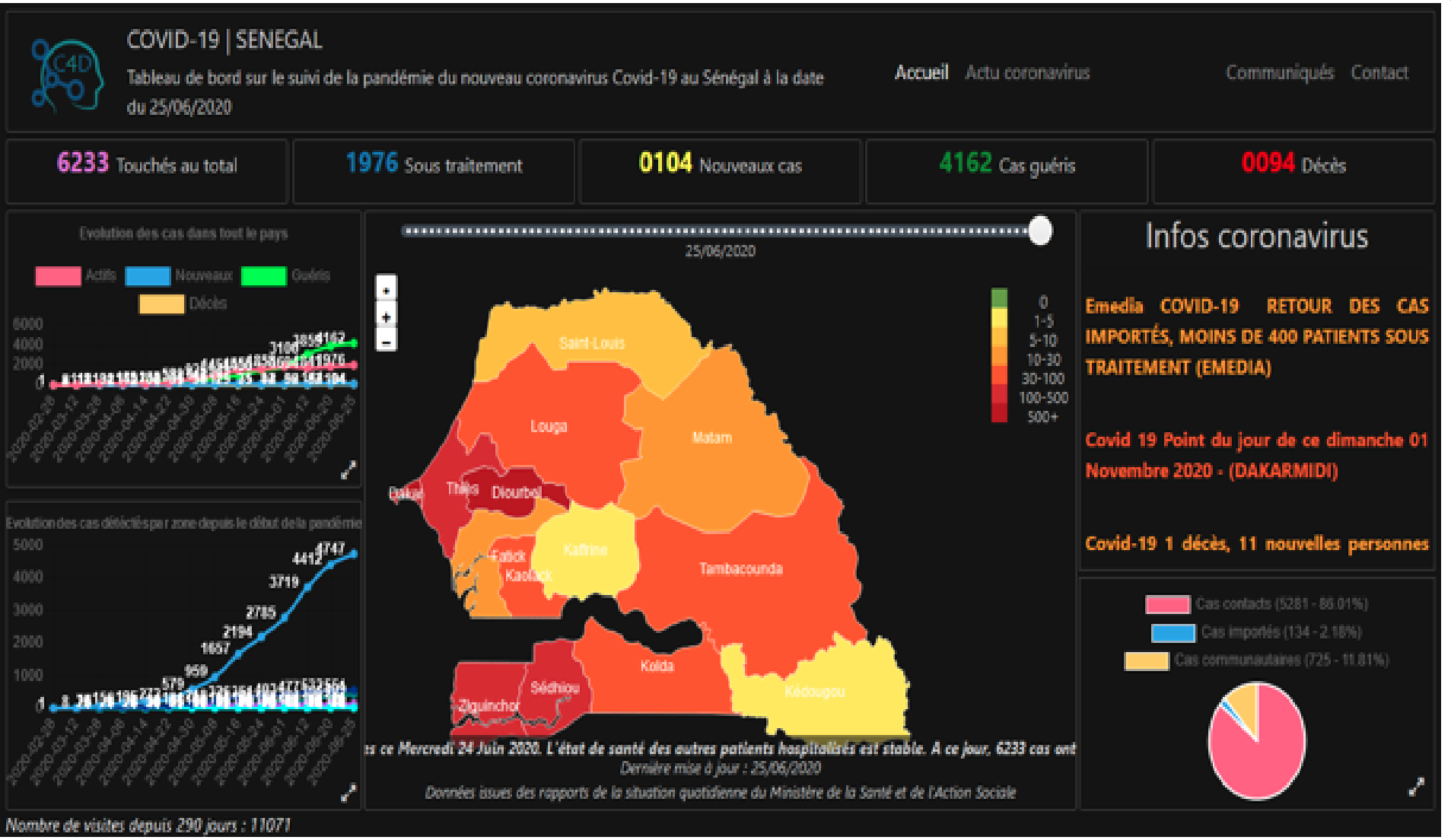
https://check4decision.univ-thies.sn/search/checked-articles

Projets Check4Decison
Résultats: Dashboard



https://check4decision.univ-thies.sn/search/dashboard/articles

Projets Check4Decison
Résultats: Moteur Fact-Checking



https://check4decision.univ-thies.sn/search/corona/

Réutilisation de la Data
Travaux en cours

- Crawling (Indexation) Intelligent par apprentissage profond des données (UADB)
- Extraction large spectre de données Web (Utt)
- La classification automatique des articles de presse par machine Learning (UCAD)
- Modération intelligente des contenus journalistique (UASZ
- Modélisation de la propagation des fakenews dans la presse en ligne (Dr Mansal-UCAO)
- Identification automatique des faits dans les affirmations journalistique (UIDT)
- Impacts des fakenews dans le choix des électeurs : le cas du Sénégal (Jean d'arc Post BAC-UCAO)
- Opinion Mining sur les commentaires journalistique (Dr Faty)
- Aspects légaux du web Scraping et du Crawling (Dr Bassene-UASZ)

Perspectives
Une base de connaissance

- **Une base de connaissance:**
  - Regroupe des connaissances spécifiques à un domaine spécialisé donné, sous une forme exploitable.
  - Va centraliser toutes les informations disponibles par rapport à un sujet ou un domaine bien donné.
- A court terme : CEDEAO
- A moyen terme: Presse Francophone Africaine
- A long terme : Afrique

Pour plus d'informations

From Intelligent Crawling to Inclusive Fact-Checking: An End-to-End System

Edouard Ngor SARR
UFR SES
Université Assane SECK de Ziguinchor
Ziguinchor-SENEGAL
edouard-ngor.sarr@univ-zig.sn

Lamine FATY
Check4Decision Research Project
Université Assane SECK de Ziguinchor
Ziguinchor-SENEGAL
lamine.faty@univ-zig.sn

Moussa Déthié SARR
UFR-Sciences et Technologies
Université de Thiès
Thiés-SENEGAL
mdsarr@univ-thies.sn

Mouhamadou Moustapha SISSOKHO
Check4Decision Research Project
Université de Thiés
Thiés-SENEGAL
siskomouhamed@gmail.com

Fatima TOURE
Check4Decision Research Project
Université de Thiés
Thiés-SENEGAL
toure.fatima947@gmail.com

Ousmane SALL
UFR-Sciences et Technologies
Université de Thiés
Thiés-SENEGAL
osall@univ-thies.sn

Abstract: *In this article, we present an aggregation platform of journalistic contents based on an intelligent crawler of articles from Online Press, of an inclusive fact-checking approach and an Opinion Mining method. More than 300 000 press articles from more than 145 Senegalese online information websites have been collected, processed, analyzed, aggregated, stored, and classified by category and theme, thanks to Machine Learning. The primary objective in these researches is to provide Web surfers, mainly journalists and fact-checkers, with a unique platform aggregating all the data related to Senegalese current affairs with analysis options, display and fact-checking.*

Keywords: *Fact-checking, Crawling, Scraping, Machine Learning, Classification, Search engine.*

I. INTRODUCTION & MOTIVATIONS

- The first search[5] engine is generic and has more than 300 000 articles from the Senegalese online press.
- The second search engine is the result of a filter related to Covid-19.
- The third and last one (fact-checking search engine) only shows press articles which have been already checked by certified fact-checkers journalists. That last search engine is directly interfaced with many famous sites such as AfricaCheck[6], ICI Radio Canada[7], Factuel AFP[8], etc. Its role is to minimize the negative impacts of fake news on public opinion.

An inclusive fact-checking opinion is also inserted into each of these search engines as well as two dashboards; one about Covid-19 in Senegal and another one for the analysis

Check4Decision

Merci de votre attention

CEA MITIC