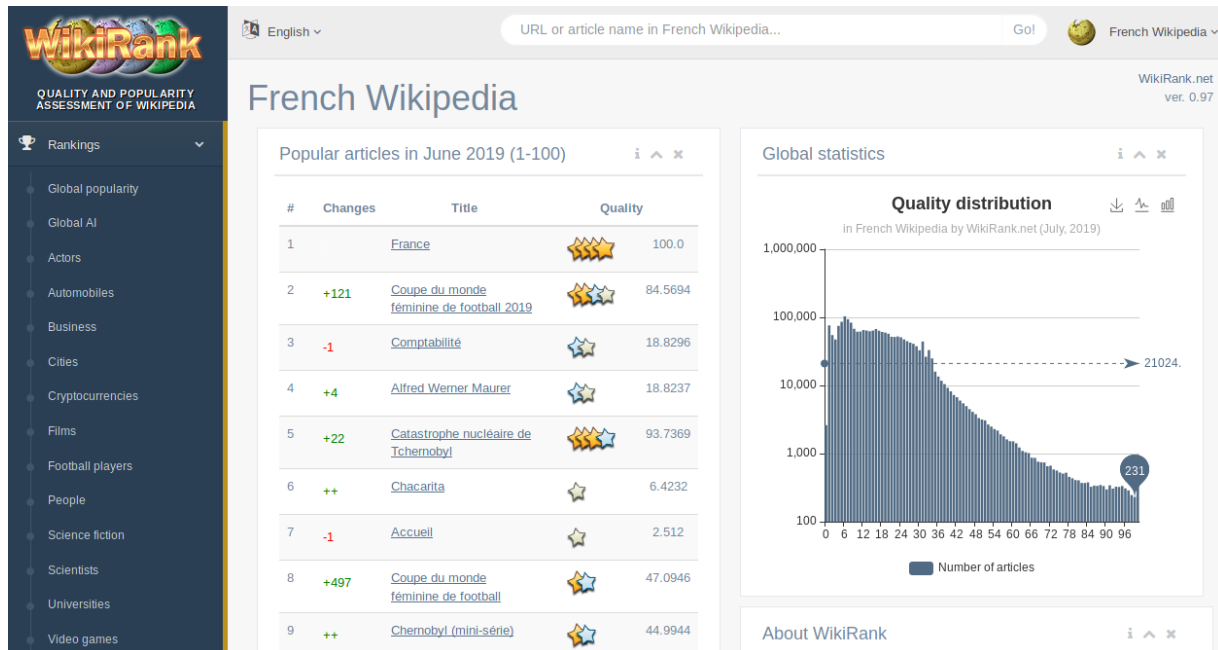


Big Data. TD 1

Sergey Kirgizov

ESIREM

La popularité des articles de Wikipédia



Dataset : *WikiRank 05.2019 – quality, popularity and Authors’ Interest for Wikipedia articles.*

https://figshare.com/articles/WikiRank_05_2019_-_quality_scores_popularity_and_AI_for_Wikipedia_articles/8231273/2

Licence : domaine public, CC0¹

La description originale :

This dataset includes a list of over 39 million Wikipedia articles in 55 languages with quality scores by WikiRank (<https://wikirank.net>). Quality scores of articles are based on Wikipedia dumps from May, 2019. Popularity and Authors’ Interest based on activity in April 2019.

Format

- page_id – The identifier of the Wikipedia article (int), e.g. 4519301
- page_name – The title of the Wikipedia article (utf-8), e.g. General relativity
- wikirank_quality – quality score for Wikipedia article in a scale 0-100 (as of May 1, 2019)
- popularity – median of daily number of page views of the Wikipedia article during April 2019
- authors_interest – number of authors of the Wikipedia article during April 2019

Méthodes et technologies : à votre choix.











Définitions :

Une **donnée aberrante** est une valeur ou une observation qui est “distante” des autres observations effectuées sur le même phénomène, c’est-à-dire qu’elle contraste grandement avec les valeurs “normalement” mesurées. — Wikipedia

Exemple : un article peu populaire ayant un grand nombre d’auteurs.

1. <https://creativecommons.org/publicdomain/zero/1.0/>

Veuillez utiliser vos technologies et langages préférés.

-  EXERCICE 1 : Télécharger le fichier <https://kirgizov.link/teaching/esirem/bigdata-2021/dataset/wikirank-fr.tsv.gz> contenant les scores de qualité des articles basés sur les données de Wikipédia (mai 2019), Les indices de popularité et d'intérêt des auteurs basé sur l'activité d'avril 2019.
-  EXERCICE 2 : Décompresser le fichier (en utilisant gunzip par exemple).
-  EXERCICE 3 : Regarder les données, leur structure.
-  EXERCICE 4 : Compter le nombre d'articles (nombre des lignes).
-  EXERCICE 5 : Regarder attentivement les données. Essayer de trouver des erreurs structurelles et les corriger.
-  EXERCICE 6 : Compter le nombre d'articles après avoir corrigé l'erreur.
-  EXERCICE 7 : Afficher le TOP 20 des articles les plus populaires.
-  EXERCICE 8 : Afficher le TOP 20 des articles les moins populaires.
-  EXERCICE 9 : Visualiser les corrélations entre le nombre d'auteurs et la popularité des articles.
-  EXERCICE 10 : Trouver les articles qui ont des caractéristiques aberrantes.