Measuring Quality of Collaboratively Edited Documents :

the case of Wikipedia

The presentation is given by :

Amin El Iraki

Natixis

Data Scientist

@ amine.el-iraki@dauphine.eu

Olivier Randavel

M13H
(Start-up) Data Marketing

Data Scientist

@ olivier.randavel@gmail.com

Hadrien Mariaccia

Dassault Systèmes

Data Scientist

@ hadrien.mariaccia@dauphine.eu

Hippolyte Mayard

Société Générale

Data Scientist

@ hippolyte.mayard@dauphine.eu

Valentin Vu Van

Thales

Data Scientist

@ valentin.vu-van@dauphine.eu

Measuring Quality of Collaboratively Edited Documents :

the case of Wikipedia

Use case

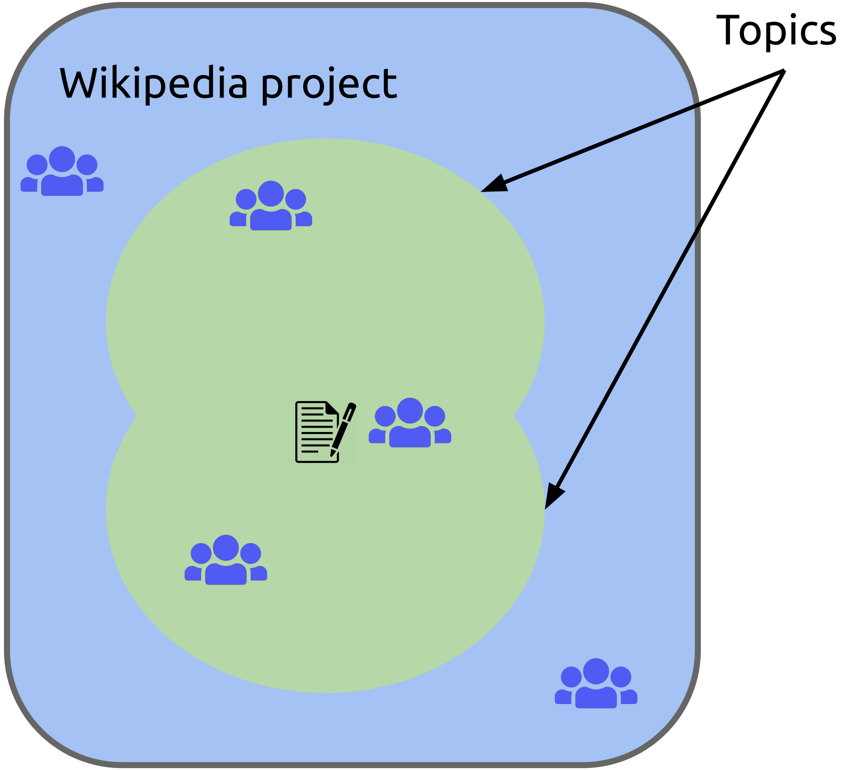
Table of contents

1. Wikipedia ranking model

1. Research paper model

1. Pros and cons

Wikipedia presentation :



Wikipedia project

Topics

49 millions articles, 6 millions in English

🟢 +300 Topics

🟦 13 Projects

+100 000 contributors per month (english language)

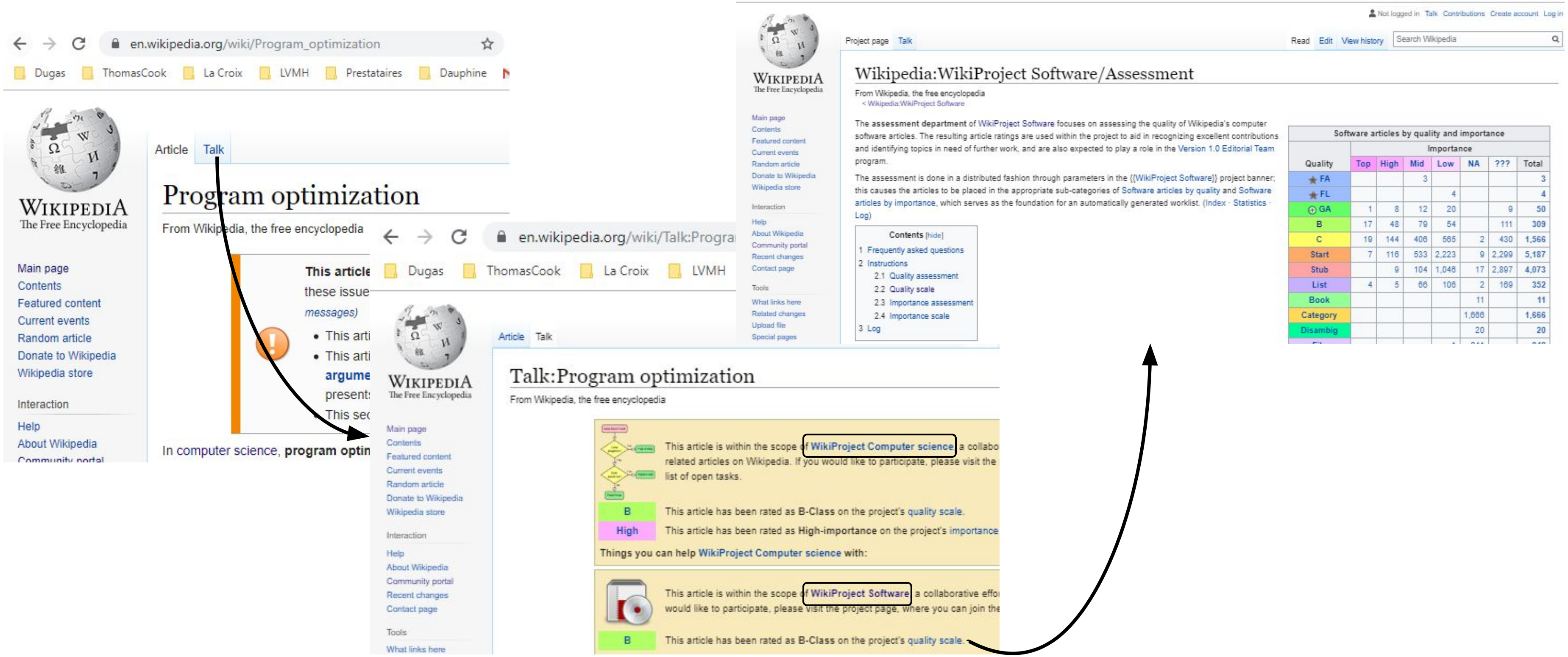
How many page views are seen by month ?

14 billion page views per month

When doing research on google, what is the chance to find a wikipedia' article in the top 5' results ?

In 96% case, top 5 results

Wikipedia ranking :



Program optimization

From Wikipedia, the free encyclopedia

This article … these issue … messages)

- This arti
- This arti … argume … present
- This sec

In computer science, **program optin**

Talk:Program optimization

From Wikipedia, the free encyclopedia

This article is within the scope of WikiProject Computer science, a collabo … related articles on Wikipedia. If you would like to participate, please visit the … list of open tasks.

B — This article has been rated as **B-Class** on the project's quality scale.

High — This article has been rated as **High-importance** on the project's importance

Things you can help WikiProject Computer science with:

This article is within the scope of WikiProject Software, a collaborative effo … would like to participate, please visit the project page, where you can join the

B — This article has been rated as **B-Class** on the project's quality scale.

Wikipedia:WikiProject Software/Assessment

From Wikipedia, the free encyclopedia
< Wikipedia:WikiProject Software

The **assessment department** of WikiProject Software focuses on assessing the quality of Wikipedia's computer software articles. The resulting article ratings are used within the project to aid in recognizing excellent contributions and identifying topics in need of further work, and are also expected to play a role in the Version 1.0 Editorial Team program.

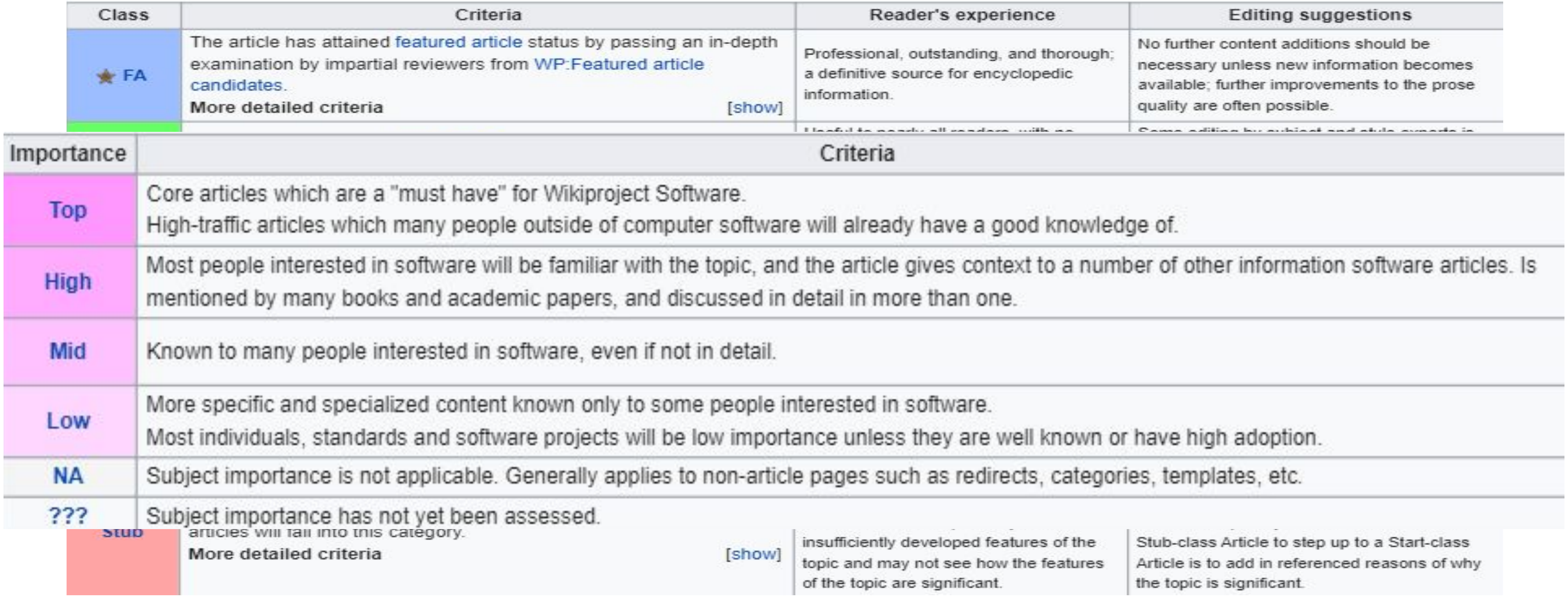
The assessment is done in a distributed fashion through parameters in the {{WikiProject Software}} project banner; this causes the articles to be placed in the appropriate sub-categories of Software articles by quality and Software articles by importance, which serves as the foundation for an automatically generated worklist. (Index · Statistics · Log)

Contents [hide]

Software articles by quality and importance

Quality	Importance						
	Top	High	Mid	Low	NA	???	Total
★ FA			3				3
★ FL				4			4
⊕ GA	1	8	12	20		9	50
B	17	48	79	54		111	309
C	19	144	406	565	2	430	1,566
Start	7	116	533	2,223	9	2,299	5,187
Stub		9	104	1,046	17	2,897	4,073
List	4	5	66	108	2	169	352
Book					11		11
Category					1,666		1,666
Disambig					20		20

Wikipedia’s quality labels:



Class	Criteria	Reader's experience	Editing suggestions
★ FA	The article has attained featured article status by passing an in-depth examination by impartial reviewers from WP:Featured article candidates. More detailed criteria [show]	Professional, outstanding, and thorough; a definitive source for encyclopedic information.	No further content additions should be necessary unless new information becomes available; further improvements to the prose quality are often possible.
		Useful to nearly all readers, with no	Some editing by subject and style experts is
Stub	articles will fall into this category. More detailed criteria [show]	insufficiently developed features of the topic and may not see how the features of the topic are significant.	Stub-class Article to step up to a Start-class Article is to add in referenced reasons of why the topic is significant.

Importance	Criteria
Top	Core articles which are a "must have" for Wikiproject Software. High-traffic articles which many people outside of computer software will already have a good knowledge of.
High	Most people interested in software will be familiar with the topic, and the article gives context to a number of other information software articles. Is mentioned by many books and academic papers, and discussed in detail in more than one.
Mid	Known to many people interested in software, even if not in detail.
Low	More specific and specialized content known only to some people interested in software. Most individuals, standards and software projects will be low importance unless they are well known or have high adoption.
NA	Subject importance is not applicable. Generally applies to non-article pages such as redirects, categories, templates, etc.
???	Subject importance has not yet been assessed.

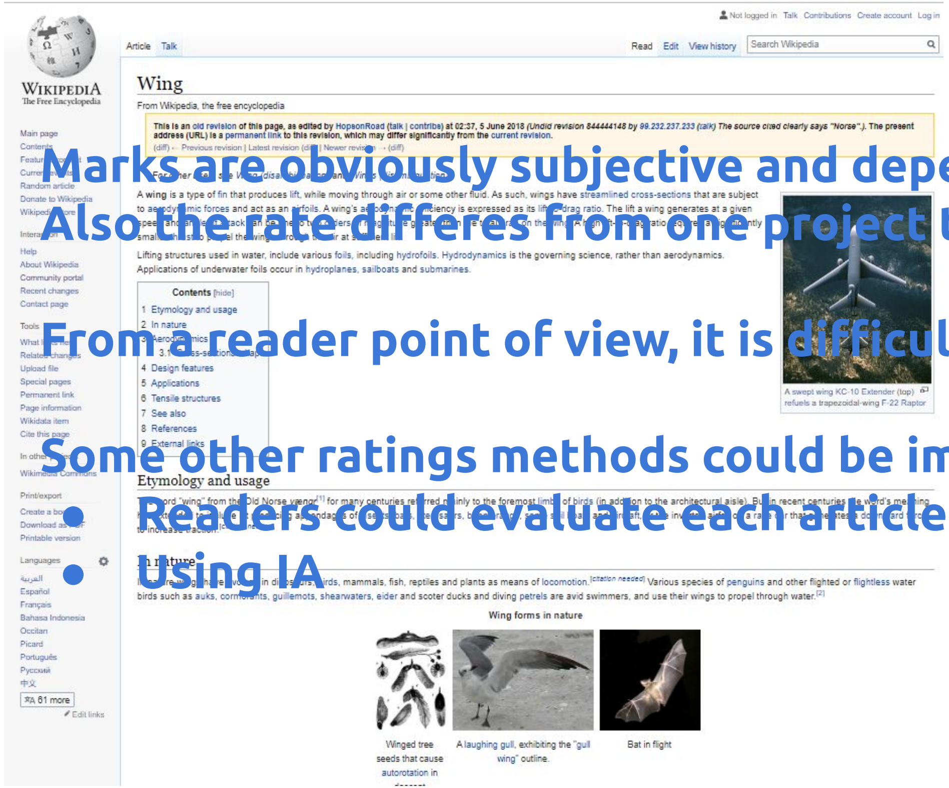
Example :

Marks are obviously subjective and dependent of the contributor’s culture.
Also the rate differes from one project to the other.

From a reader point of view, it is difficult to understand the grade.

Some other ratings methods could be implemented such as :

- Readers could evaluate each article (star rating)
- Using IA



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Wing

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by HopsonRoad (talk | contribs) at 02:37, 5 June 2018 (Undid revision 844444148 by 99.232.237.233 (talk) The source cited clearly says "Norse".). The present address (URL) is a permanent link to this revision, which may differ significantly from the current revision.
(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

A wing is a type of fin that produces lift, while moving through air or some other fluid. As such, wings have streamlined cross-sections that are subject to aerodynamic forces and act as an airfoils. A wing's aerodynamic efficiency is expressed as its lift-to-drag ratio. The lift a wing generates at a given speed and angle of attack can be one to two orders of magnitude greater than the total drag on the wing. A high lift-to-drag ratio requires a significantly smaller thrust to propel the wings through the air at sufficient lift.

Lifting structures used in water, include various foils, including hydrofoils. Hydrodynamics is the governing science, rather than aerodynamics. Applications of underwater foils occur in hydroplanes, sailboats and submarines.

A swept wing KC-10 Extender (top) refuels a trapezoidal-wing F-22 Raptor

Etymology and usage

The word "wing" from the Old Norse vængr[1] for many centuries referred mainly to the foremost limbs of birds (in addition to the architectural aisle). But in recent centuries the word's meaning has extended to include lift producing appendages of insects, bats, pterosaurs, boomerangs, some sail boats and aircraft, or the inverted airfoil on a race car that generates a downward force to increase traction.

In nature

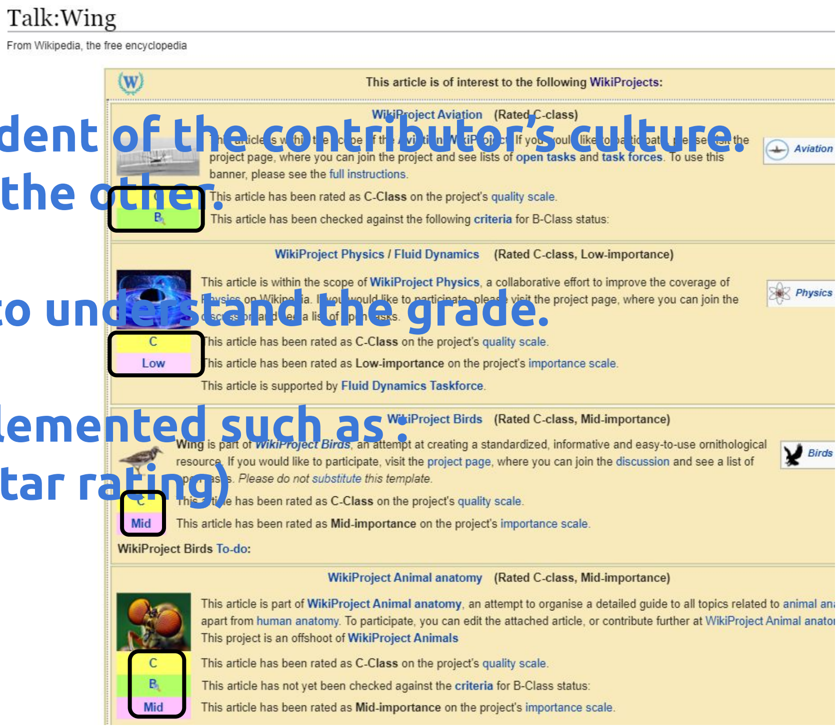
In nature, wings have evolved in dinosaurs, birds, mammals, fish, reptiles and plants as means of locomotion.[citation needed] Various species of penguins and other flighted or flightless water birds such as auks, cormorants, guillemots, shearwaters, eider and scoter ducks and diving petrels are avid swimmers, and use their wings to propel through water.[2]

Wing forms in nature

Winged tree seeds that cause autorotation in descent

A laughing gull, exhibiting the "gull wing" outline.

Bat in flight



Talk:Wing

From Wikipedia, the free encyclopedia

This article is of interest to the following WikiProjects:

WikiProject Aviation (Rated C-class)

This article is within the scope of the Aviation WikiProject. If you would like to participate, please visit the project page, where you can join the project and see lists of open tasks and task forces. To use this banner, please see the full instructions.

C B

This article has been rated as C-Class on the project's quality scale.

This article has been checked against the following criteria for B-Class status:

WikiProject Physics / Fluid Dynamics (Rated C-class, Low-importance)

This article is within the scope of WikiProject Physics, a collaborative effort to improve the coverage of Physics on Wikipedia. If you would like to participate, please visit the project page, where you can join the discussions and see a list of open tasks.

C Low

This article has been rated as C-Class on the project's quality scale.

This article has been rated as Low-importance on the project's importance scale.

This article is supported by Fluid Dynamics Taskforce.

WikiProject Birds (Rated C-class, Mid-importance)

Wing is part of *WikiProject Birds*, an attempt at creating a standardized, informative and easy-to-use ornithological resource. If you would like to participate, visit the project page, where you can join the discussion and see a list of open tasks. *Please do not substitute this template.*

C Mid

This article has been rated as C-Class on the project's quality scale.

This article has been rated as Mid-importance on the project's importance scale.

WikiProject Birds To-do:

WikiProject Animal anatomy (Rated C-class, Mid-importance)

This article is part of WikiProject Animal anatomy, an attempt to organise a detailed guide to all topics related to animal ana
apart from human anatomy. To participate, you can edit the attached article, or contribute further at WikiProject Animal anato
This project is an offshoot of WikiProject Animals

C B Mid

This article has been rated as C-Class on the project's quality scale.

This article has not yet been checked against the criteria for B-Class status:

This article has been rated as Mid-importance on the project's importance scale.

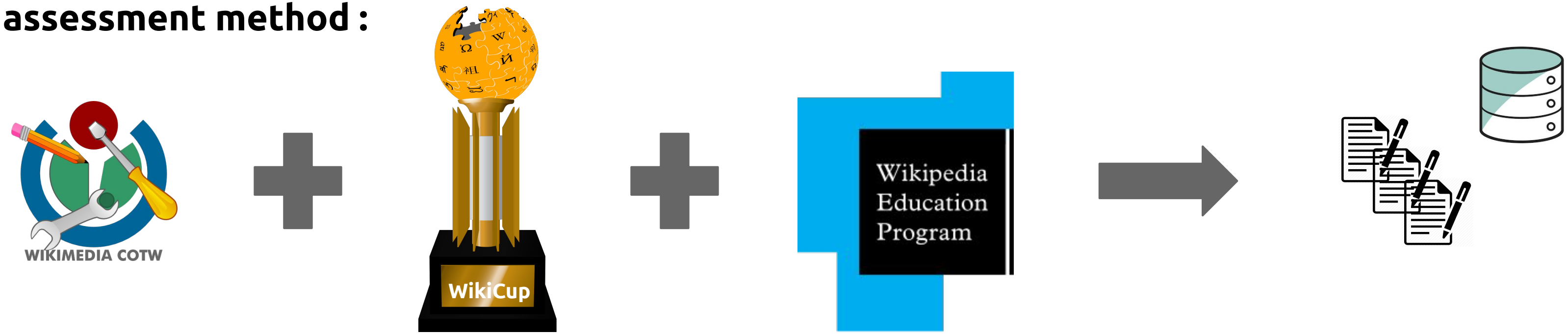
Why this article ?

- Machine learning methods to improve the precision of the quality measures of Wikipedia articles

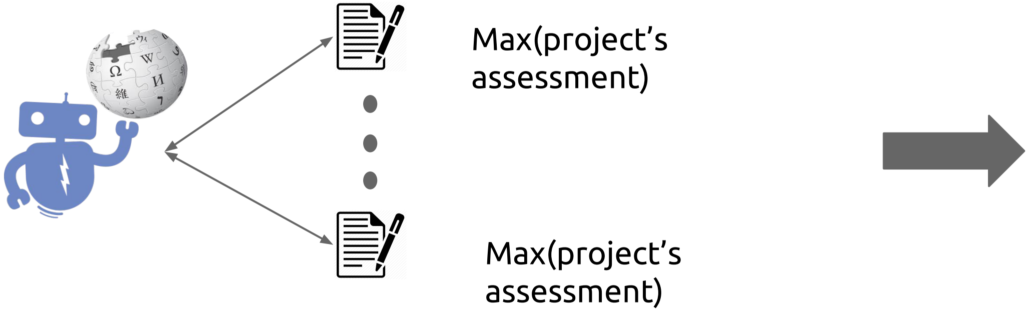
- Innovative point of view : Using structure based information AND article based information

Data set presentation :

Article assessment method :

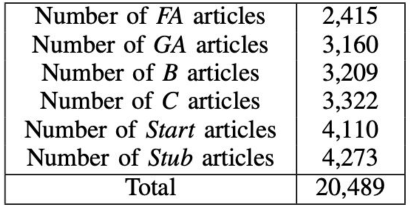


Data collection :



Max(project's assessment)

Max(project's assessment)



Number of *FA* articles	2,415
Number of *GA* articles	3,160
Number of *B* articles	3,209
Number of *C* articles	3,322
Number of *Start* articles	4,110
Number of *Stub* articles	4,273
Total	20,489

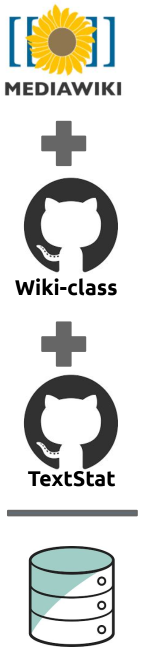
TABLE III: Distribution of the data set within different quality classes

Data set presentation :

Data cleaning :

- Remove two classes that were too small : categories A and B+
- Remove articles that have been deleted

Data preprocessing :



- Get Content
- Compute structure-based features
- Compute content-based
- Clean Data

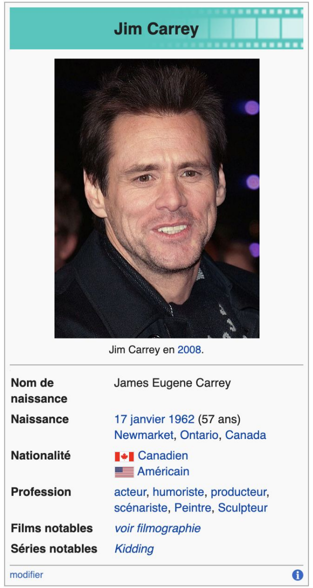
Feature selection :

Hypothesis: the writing style matters for measuring the articles quality.

Structure-based features **Wiki-class**	Content-based features **TextStat**
● Article length	● Flesch reading score (En)
● Number of references	● Flesch-Kincaid grade level (US)
● Number of outlinks to other Wikipedia pages	● Smog index (En)
● Number of citation templates	● Coleman-Liau index (US)
● Number of non-citation templates	● Automated readability index (US)
● Number of categories linked in the text	● Difficult words
● Number of images / length of article	● Dale-Chall score
● Information noise score	● Linsear write formula (US military)
● Article has an infobox or not	● Gunning-Fog index
● Number of level 2 headings/ Number of level 3+ headings	

Content-based features (Examples)

- *A wikipedia infobox*



Jim Carrey

Jim Carrey en 2008.

Nom de naissance	James Eugene Carrey
Naissance	17 janvier 1962 (57 ans) Newmarket, Ontario, Canada
Nationalité	🇨🇦 Canadien 🇺🇸 Américain
Profession	acteur, humoriste, producteur, scénariste, Peintre, Sculpteur
Films notables	*voir filmographie*
Séries notables	*Kidding*

modifier

- $$\begin{aligned} \mathit{flesch\_reading\_ease} = 206.835 \\ - (1.015 \times \mathit{avg\_sentence\_len}) \\ - (84.6 \times \mathit{avg\_syllables\_per\_word}) \end{aligned} \tag{1}$$

- $$\begin{aligned} \mathit{flesch\_kincaid\_grade} = 11.8 \times \mathit{avg\_syllables\_per\_word} \\ + 0.39 \times \mathit{avg\_sentence\_len} - 15.59 \end{aligned} \tag{2}$$

Feature selection :

Hypothesis: the writing style matters for measuring the articles quality.

Structure-based features Wiki-class	Content-based features TextStat
● Article length ● Number of references ● Number of outlinks to other Wikipedia pages ● Number of citation templates ● Number of non-citation templates ● Number of categories linked in the text ● Number of images / length of article ● Information noise score ● Article has an infobox or not ● Number of level 2 headings/ Number of level 3+ headings	● Flesch reading score ● Flesch-Kincaid grade level ● Smog index ● Coleman-Liau index ● Automated readability index ● Difficult words ● Dale-Chall score ● Linsear write formula ● Gunning-Fog index

Question: Several readability scores related ?

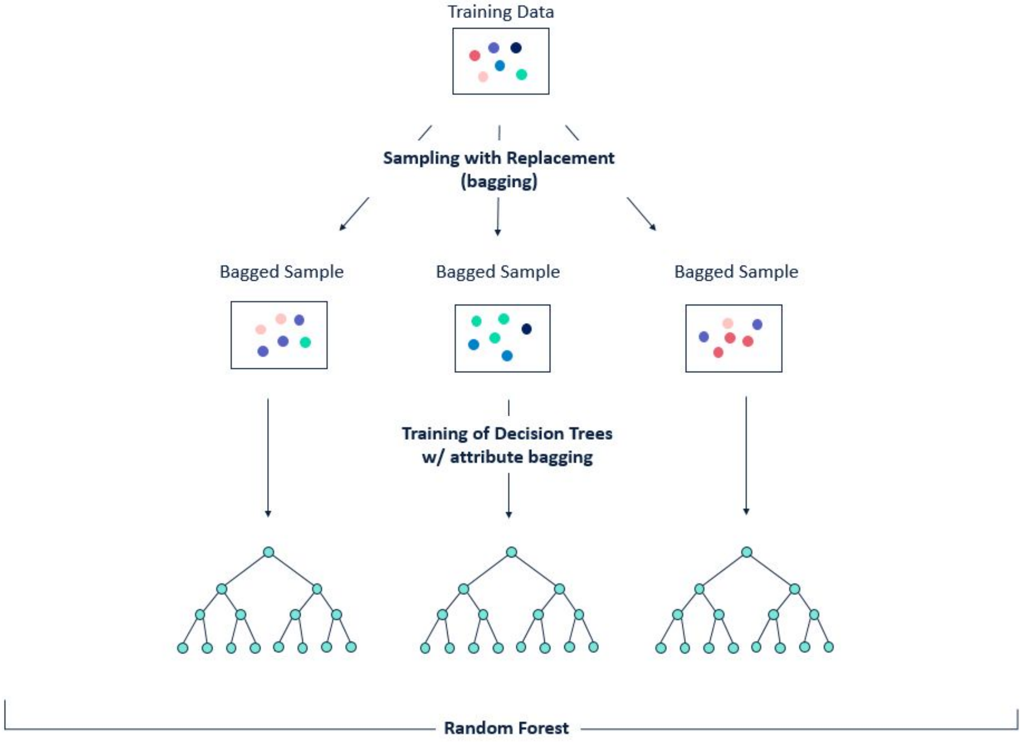
Solutions proposed in the article :

Algorithms	Hyper-parameters	Specificities	Accuracy
Linear regression	None	- dependent variable: quality class - independent variables: the features - converted the quality class to an integer: Stub to 0, Start to 1, C to 2, B to 3, GA to 4 and FA to 6	25%
Multinomial logistic regression	None	Standard	60% (5-fold Cross-Validation)
KNN	K = 3	Using the Euclidean distance	55% (5-fold Cross-Validation)
CART	None	Standard	48%
SVM	None	Standard	61% (5-fold Cross-Validation)
Random Forest	None	Applied uniquely on the structure-based features	58% (5-fold Cross-Validation)
Random Forest	None	Applied on the complete set	64% (5-fold Cross-Validation)

The most accurate model: Random Forest

Performances:

- 3 metrics:
  - Accuracy: 64%
  - AUC (Area Under Curve): 0,91
  - NDCG score: 0,987


Training Data
Sampling with Replacement (bagging)
Bagged Sample
Bagged Sample
Bagged Sample
Training of Decision Trees w/ attribute bagging
Random Forest

Pros & cons of these solutions :

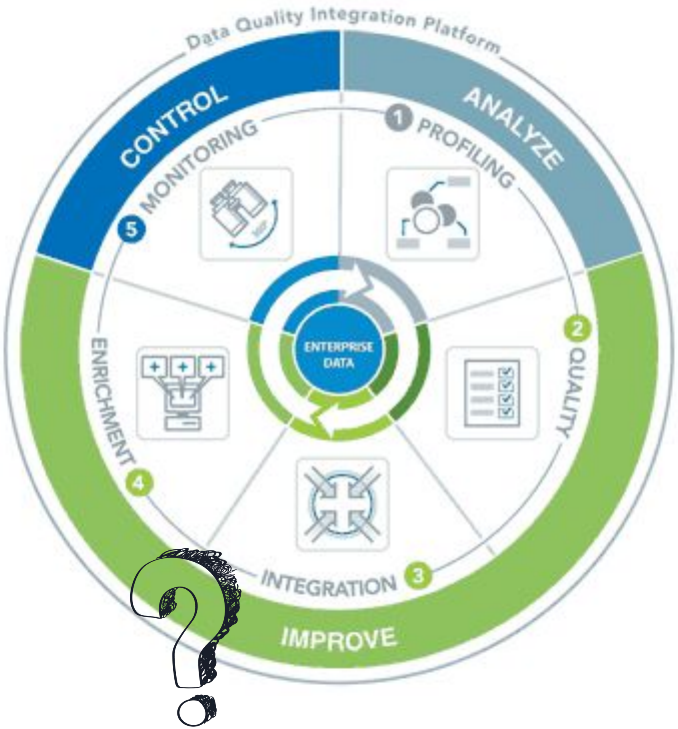
Pros

- Over-fitting problem taken avoided by the 5-fold cross validation
- The model improved the accuracy of Wikipedia quality prediction
- This paper provides advices for authors to improve the quality of Wikipedia articles

Cons

- Only for english Wikipedia articles
- The data used to evaluate the model could be reconsidered. Taking the maximum of each articles' marks is not the best practise.
- The data are manually labelled, so the marks are subjectives

Data set :



What can we say about learning on subjective evaluations :

A paper is discussing this problem : *The Success and Failure of Quality Improvement Projects in Peer Production Communities*. It aims to evaluate several rating groups that contributed to wikipedia assessments.

Mains conclusions are :

- Some articles were not correctly assessed, contributor failed to apply the assessment criteria.
- Assessments made by WEP have been given by students and it results in lower rating/writing experienced. Some groups are more efficient to produce high rated articles.

YO DAWG, IT SEEMS YOUR DATA QUALITY REPORT
NEEDS A DATA QUALITY CHECK OF IT'S OWN
memegenerator.net

References :

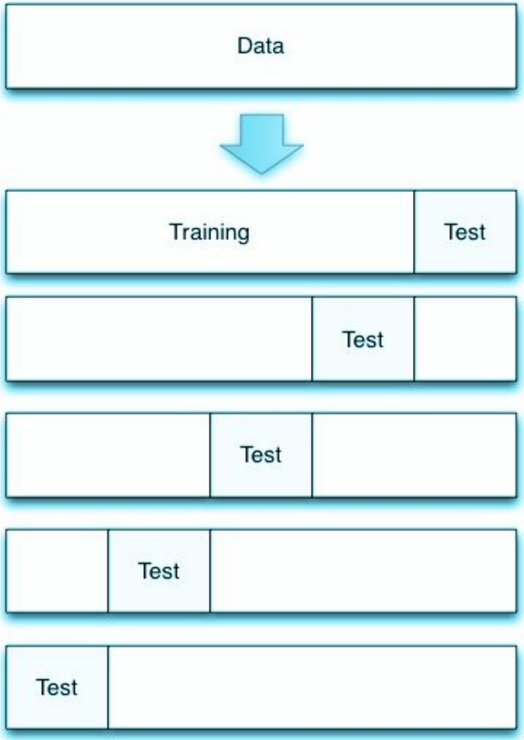
- A paper named : Measuring Quality of Collaboratively Edited Documents: the case of Wikipedia done
- A paper named : The Success and Failure of Quality Improvement Projects in Peer Production Communities
- A tool with rank suggestions : Wikirank.live
- Link to English Wikipedia Quality Asssessment Dataset. Data used to perfom classification
- Link to wikipedia assessment method

Thank you for your attention

Annexe

Model Evaluation :

- Application of different classification methods with 5-fold cross-validation techniques
    - Dataset divided into five equal parts (5- fold).
    - Four parts used as a training set / remaining part as the testing set.
    - Process repeated five times, each part being used as a testing set alternately.
- A good practical technique for bias-variance trade-off in evaluating machine learning algorithms.


Data
Training
Test
Test
Test
Test
Test