오픈소스/ 데이터 라이선스의 이해

Understanding Open Source and Data Licenses

이철남

충남대 법학전문대학원

CONTENTS

01 Open Source Licenses

02 Data Licenses

03 Open Source AI ?

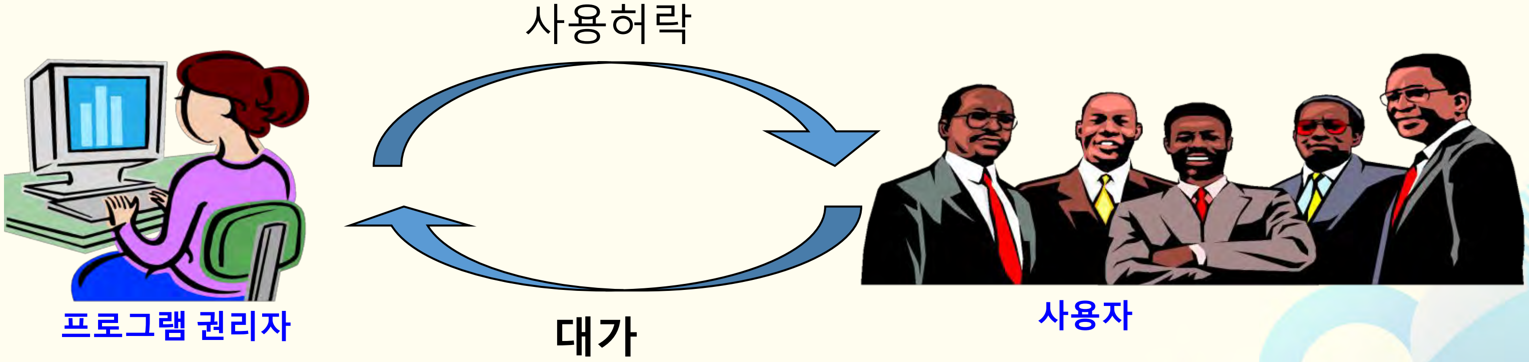
01

오픈소스 라이선스

SW 라이선스

- A software license
  is a legal instrument (usually by way of contract law, with or without printed material) governing the use or redistribution of software. (wikipedia)

  - Ownership vs. licensing



사용허락

대가

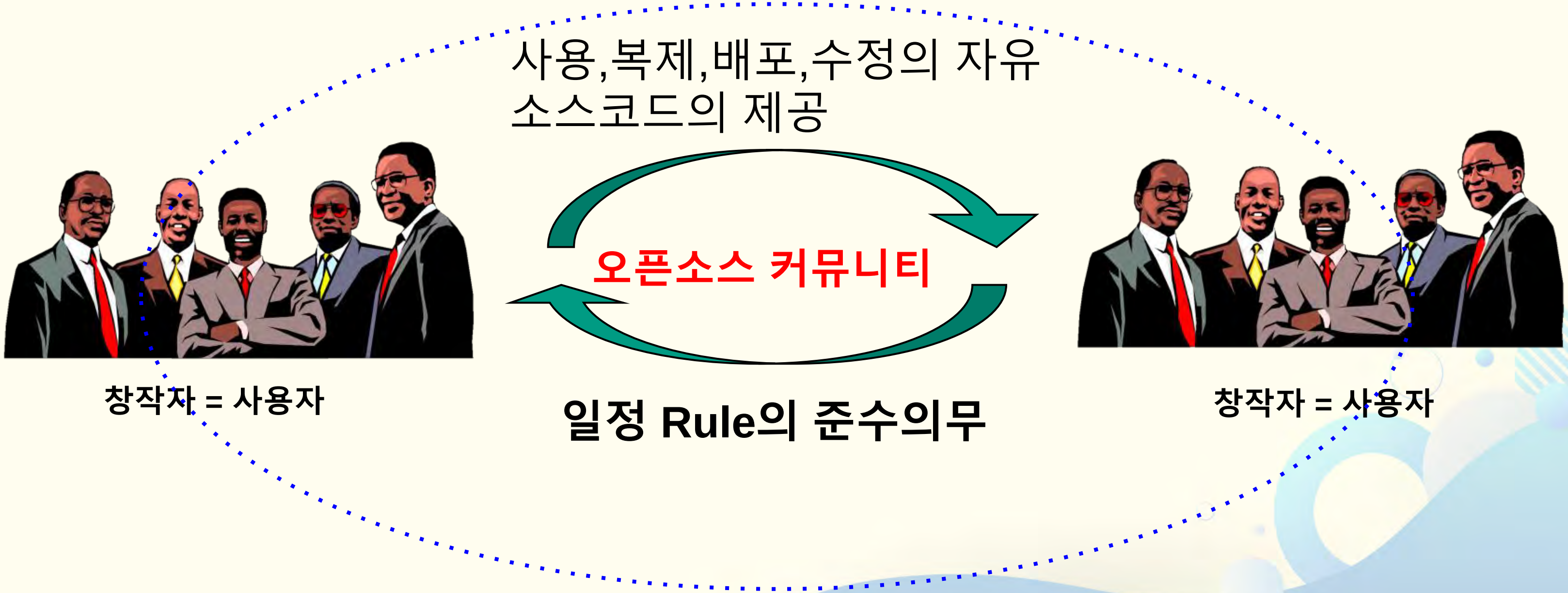
프로그램 권리자

사용자

OSS 라이선스의 의미와 기본 원리

- An open-source license

  is a type of license for computer software and other products that allows the source code, blueprint or design to be used, modified and/or shared under defined terms and conditions. (wikipedia)



사용,복제,배포,수정의 자유
소스코드의 제공

오픈소스 커뮤니티

창작자 = 사용자

일정 Rule의 준수의무

창작자 = 사용자

Permissive License

You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work (Apache 2.0)

You must cause any modified files to carry prominent notices stating that You changed the files (Apache 2.0)

You must give any other recipients of the Work or Derivative Works a copy of this License (Apache 2.0)

Copyleft License

You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License. (GPL 2.0)

You may copy and distribute the Program (or a work based on it, under Section 2) in object code … provided that you also do one of the following:

a) Accompany it with the complete corresponding machine-readable source code,

(GPL 2.0)

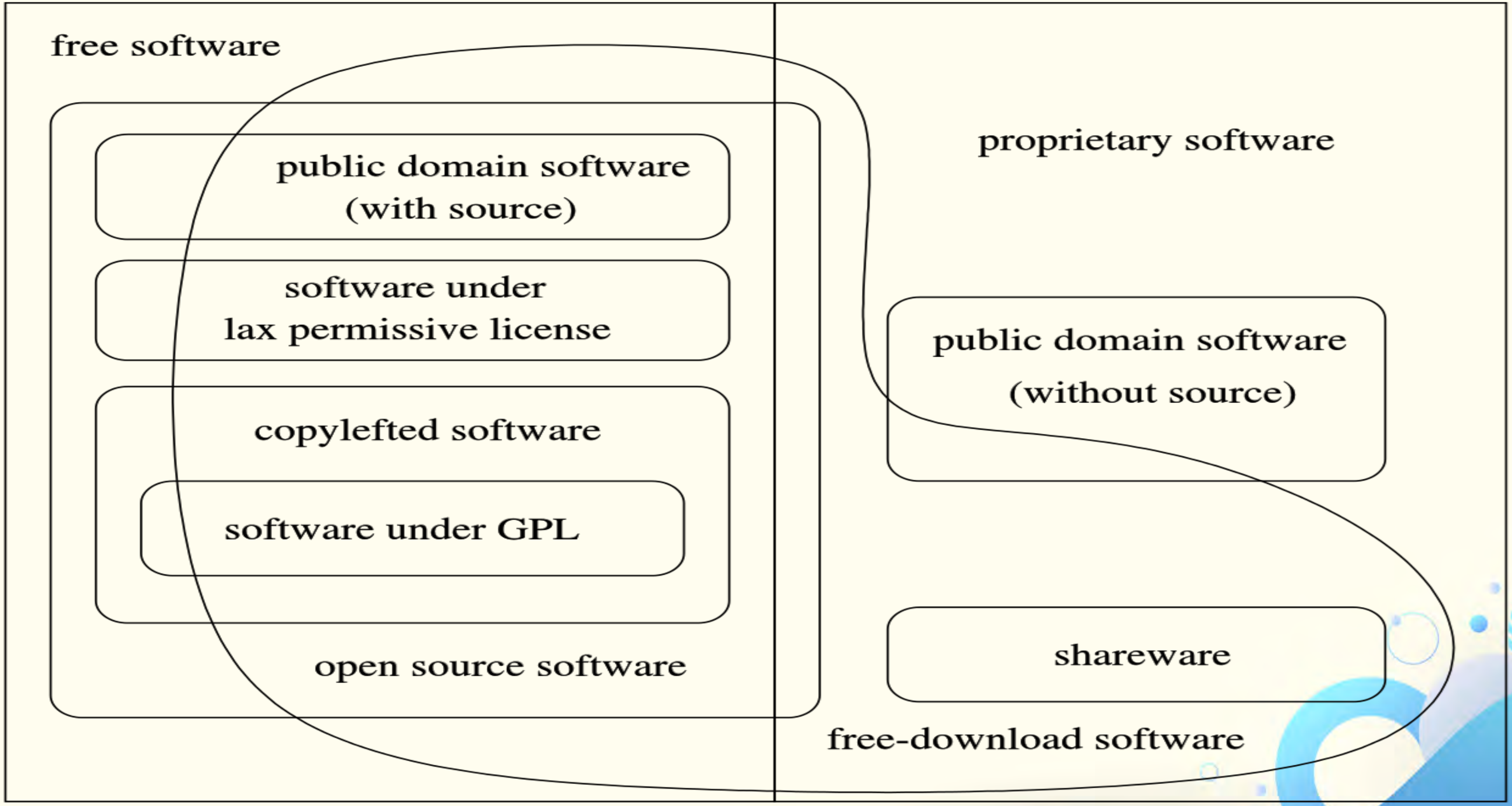
OSS의 정의

- **Open-source software (OSS)**

  is computer software that is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose. (Wikipedia)

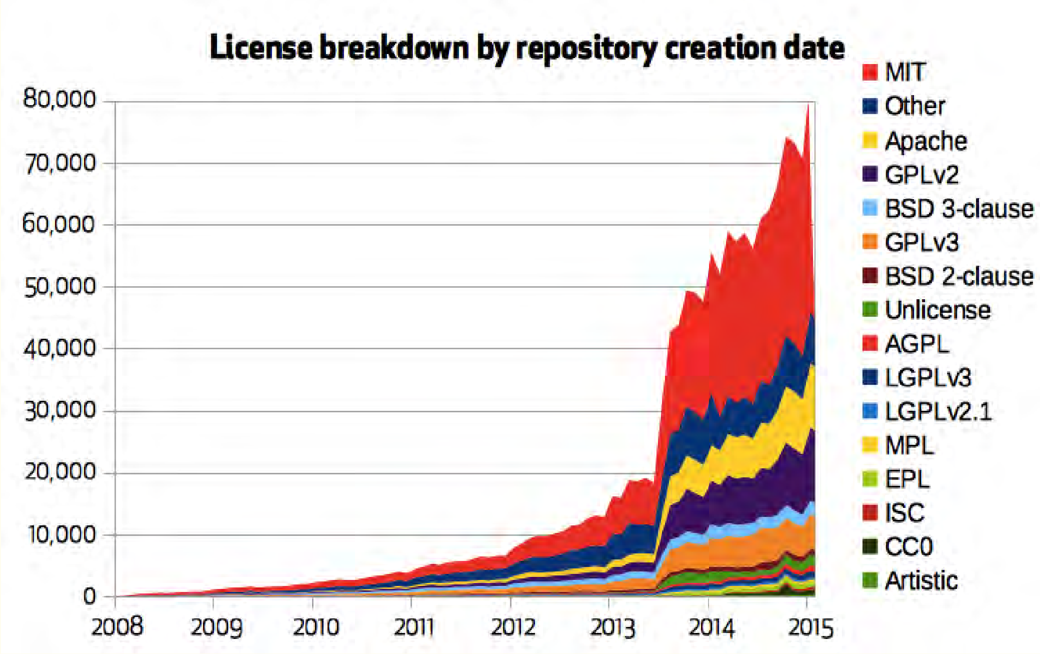
- **The Open Source Definition**
  1. Free Redistribution
  2. Source Code
  3. Derived Works
  4. Integrity of The Author's Source Code
  5. No Discrimination Against Persons or Groups
  6. No Discrimination Against Fields of Endeavor
  7. Distribution of License
  8. License Must Not Be Specific to a Product
  9. License Must Not Restrict Other Software
  10. License Must Be Technology-Neutral

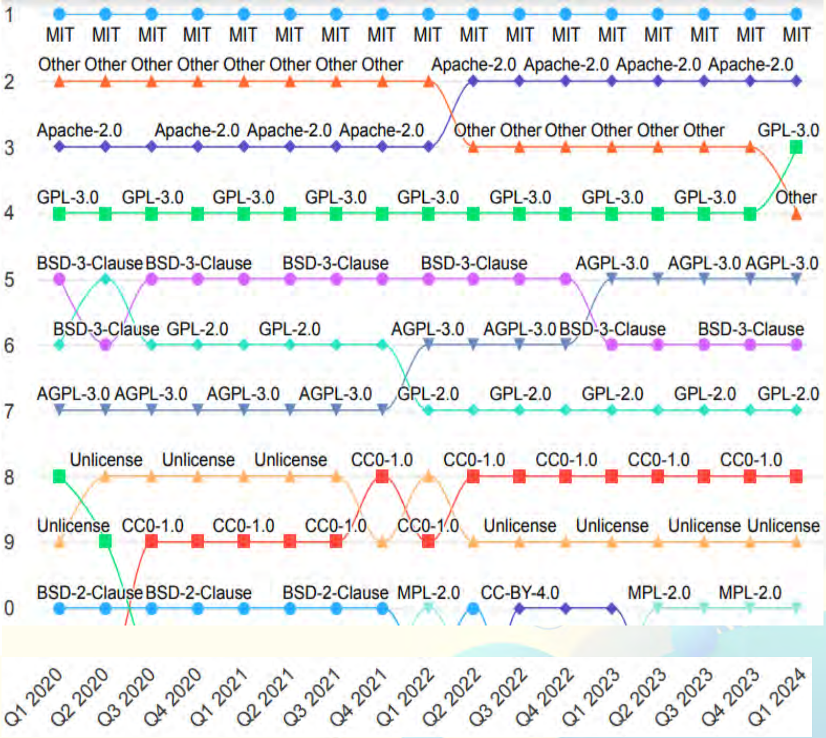
(Open Source Initiative)



(Source : Wikipedia)

오픈소스 라이선스 이용 현황



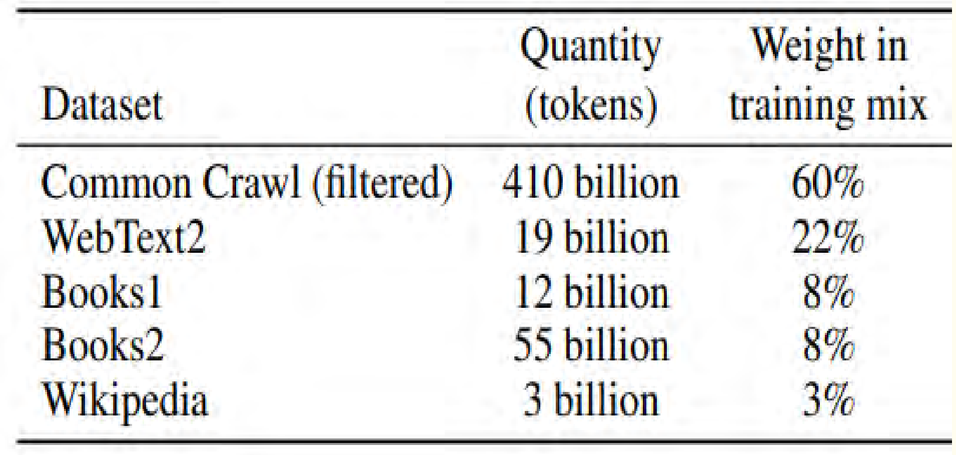


(source: github.com)

02

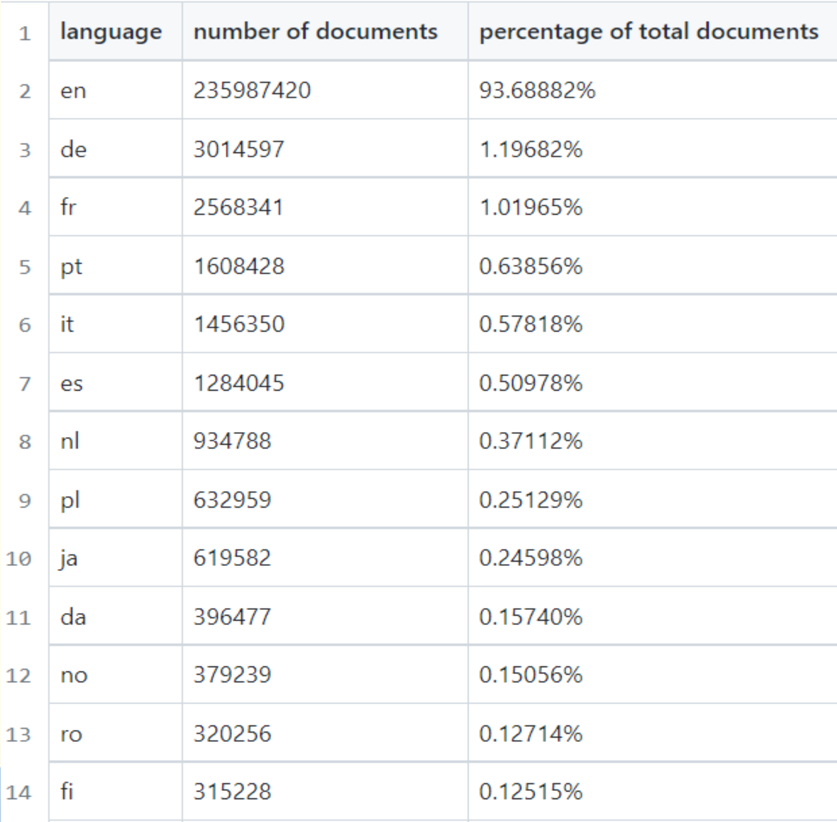
데이터 라이선스

Data : GPT-3의 학습데이터



Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

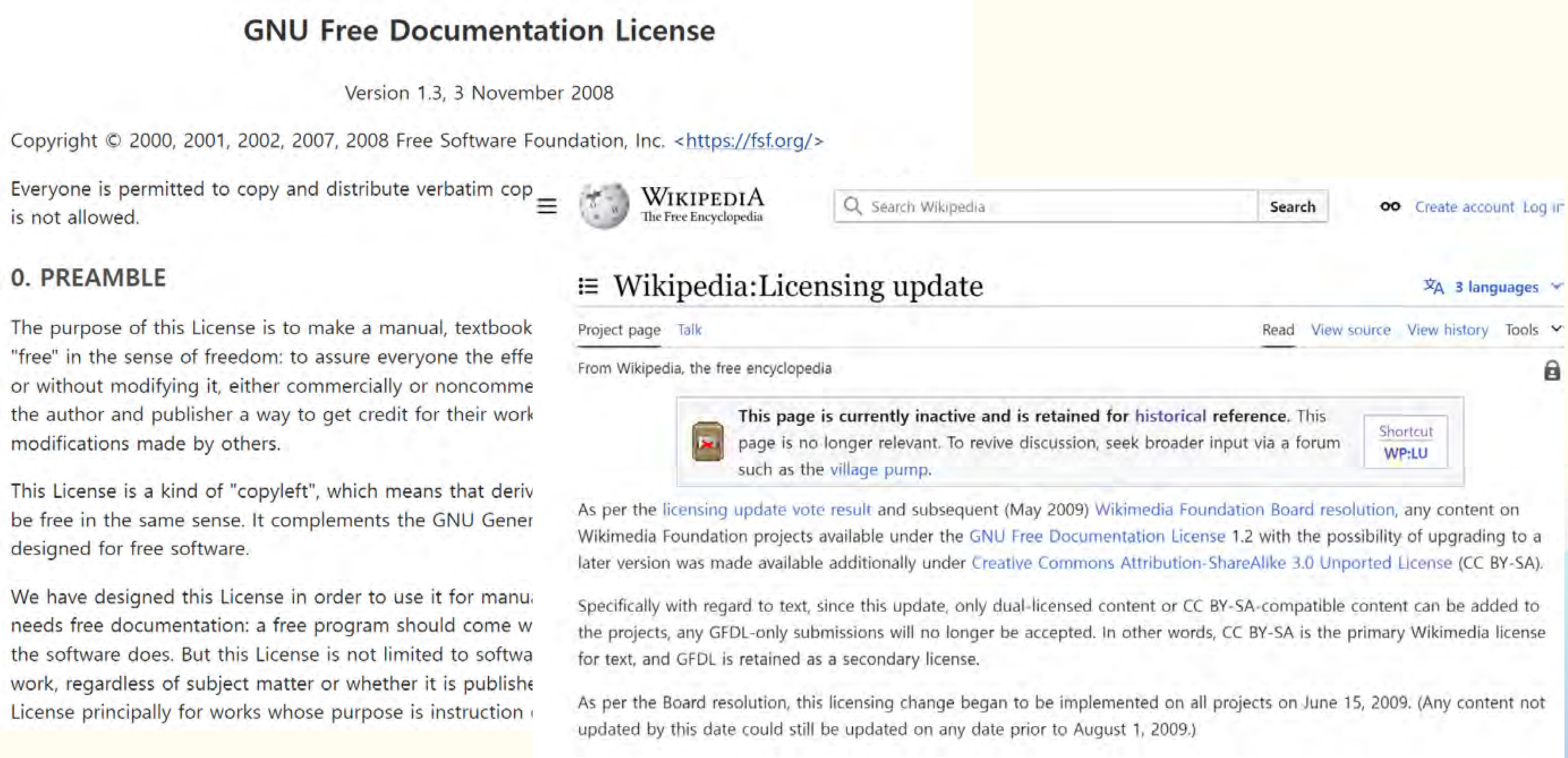
(source : arXiv:2005.14165)



language	number of documents	percentage of total documents
en	235987420	93.68882%
de	3014597	1.19682%
fr	2568341	1.01965%
pt	1608428	0.63856%
it	1456350	0.57818%
es	1284045	0.50978%
nl	934788	0.37112%
pl	632959	0.25129%
ja	619582	0.24598%
da	396477	0.15740%
no	379239	0.15056%
ro	320256	0.12714%
fi	315228	0.12515%

(source: github.com)

GFDL



GNU Free Documentation License

Version 1.3, 3 November 2008

Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc. <https://fsf.org/>

Everyone is permitted to copy and distribute verbatim cop
is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook
"free" in the sense of freedom: to assure everyone the effe
or without modifying it, either commercially or noncomme
the author and publisher a way to get credit for their work
modifications made by others.

This License is a kind of "copyleft", which means that deriv
be free in the same sense. It complements the GNU Gener
designed for free software.

We have designed this License in order to use it for manu
needs free documentation: a free program should come w
the software does. But this License is not limited to softwa
work, regardless of subject matter or whether it is publishe
License principally for works whose purpose is instruction

Wikipedia:Licensing update

From Wikipedia, the free encyclopedia

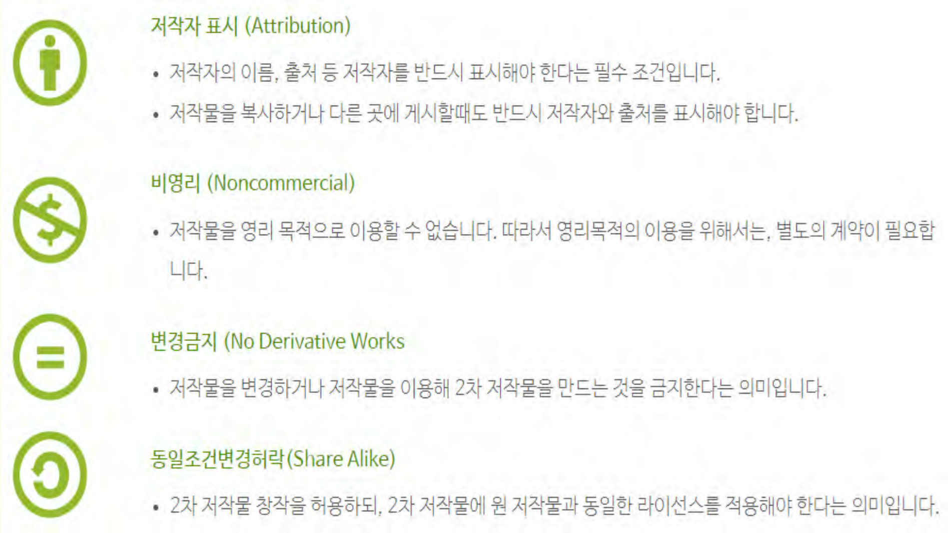
This page is currently inactive and is retained for historical reference. This page is no longer relevant. To revive discussion, seek broader input via a forum such as the village pump.

As per the licensing update vote result and subsequent (May 2009) Wikimedia Foundation Board resolution, any content on Wikimedia Foundation projects available under the GNU Free Documentation License 1.2 with the possibility of upgrading to a later version was made available additionally under Creative Commons Attribution-ShareAlike 3.0 Unported License (CC BY-SA).

Specifically with regard to text, since this update, only dual-licensed content or CC BY-SA-compatible content can be added to the projects, any GFDL-only submissions will no longer be accepted. In other words, CC BY-SA is the primary Wikimedia license for text, and GFDL is retained as a secondary license.

As per the Board resolution, this licensing change began to be implemented on all projects on June 15, 2009. (Any content not updated by this date could still be updated on any date prior to August 1, 2009.)

Creative Commons License



저작자 표시 (Attribution)

- 저작자의 이름, 출처 등 저작자를 반드시 표시해야 한다는 필수 조건입니다.
- 저작물을 복사하거나 다른 곳에 게시할때도 반드시 저작자와 출처를 표시해야 합니다.

비영리 (Noncommercial)

- 저작물을 영리 목적으로 이용할 수 없습니다. 따라서 영리목적의 이용을 위해서는, 별도의 계약이 필요합니다.

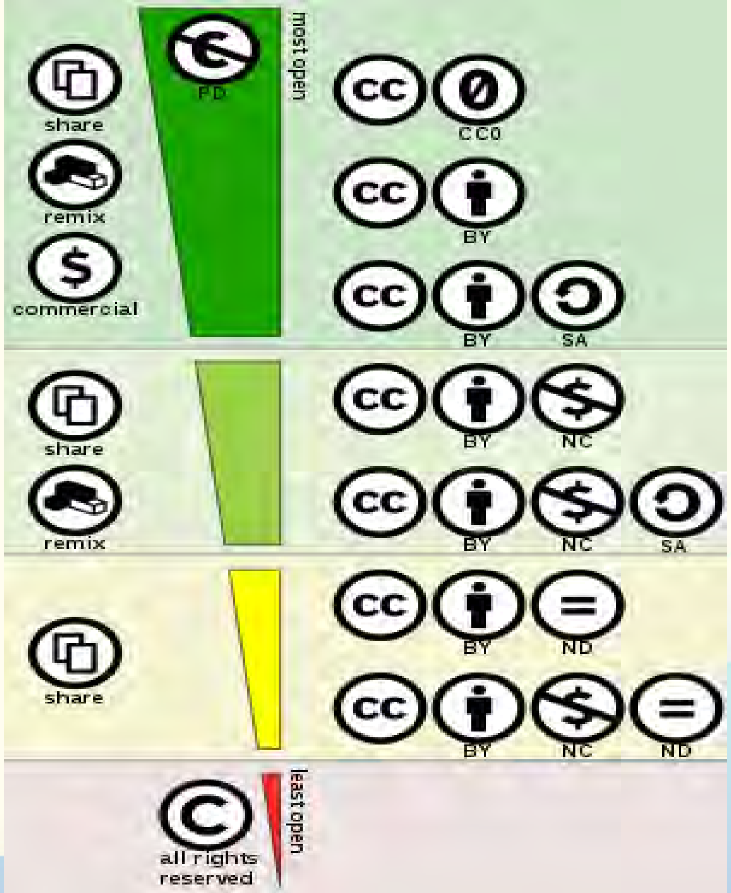
변경금지 (No Derivative Works

- 저작물을 변경하거나 저작물을 이용해 2차 저작물을 만드는 것을 금지한다는 의미입니다.

동일조건변경허락(Share Alike)

- 2차 저작물 창작을 허용하되, 2차 저작물에 원 저작물과 동일한 라이선스를 적용해야 한다는 의미입니다.

(source: cckorea.org)



(source: wikipedia)

Common Crawl

BIG PICTURE - THE DATA - ABO

Us

We build and maintain an open repository of **crawl data** that can be **accessed and analy** **by anyone**.

Need **years of free** web page **change the world**.

はい C'est vrai! 40+ languages Si. Tak! Efectivamente. हाँ You bet.

RAW DATA
METADATA
TEXT DATA

$0

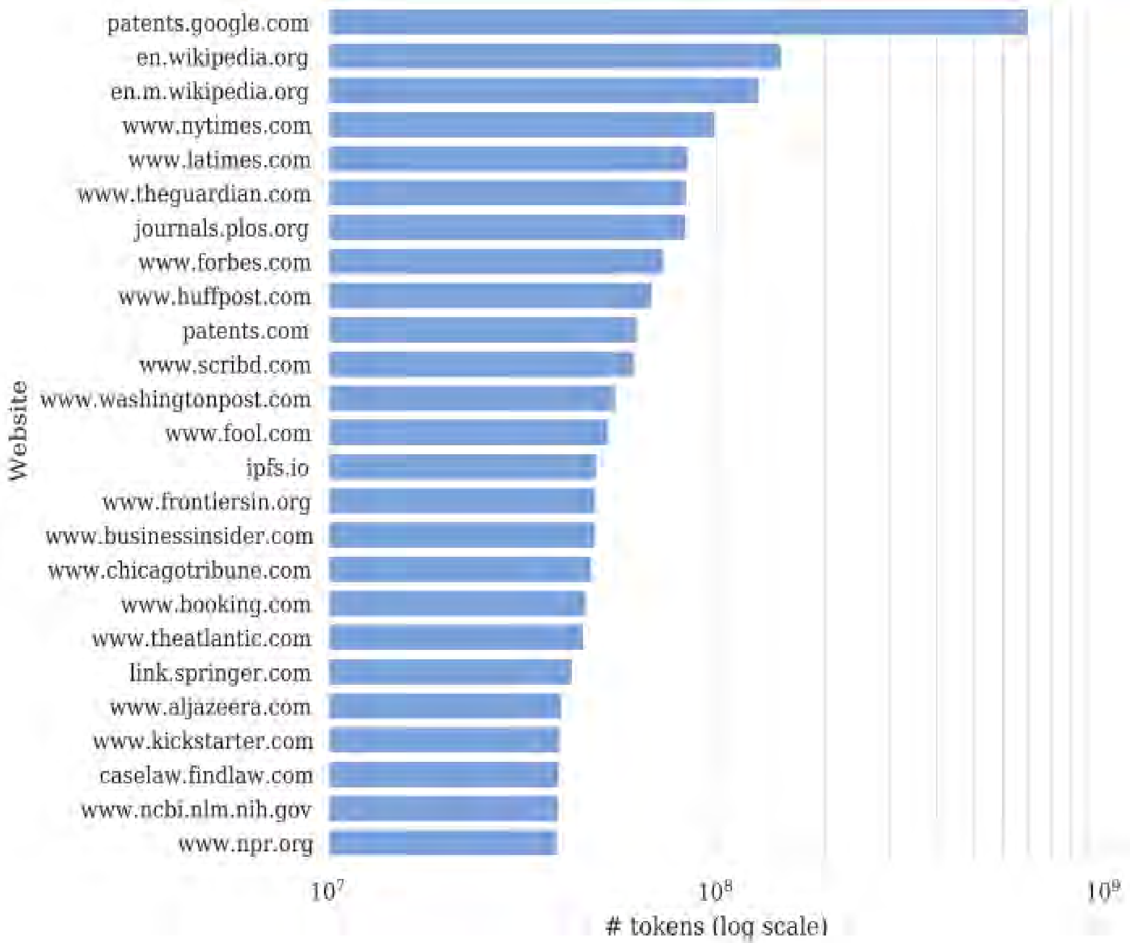
We gather it.
We aggregate it.
You utilize it.
And it's all free.

HOW WE'RE BI PETABY

bill OF

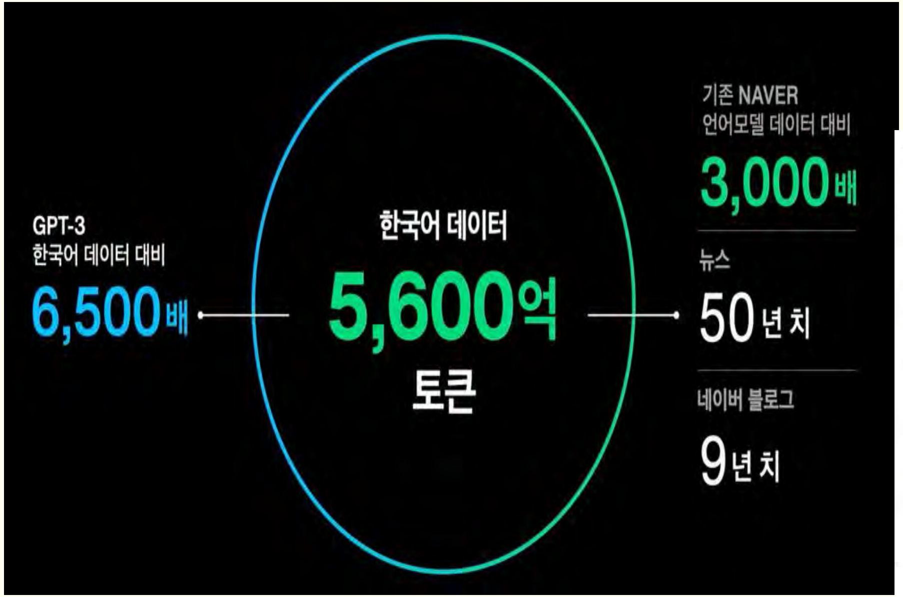
trillio OF LIN

(source: common crawl)

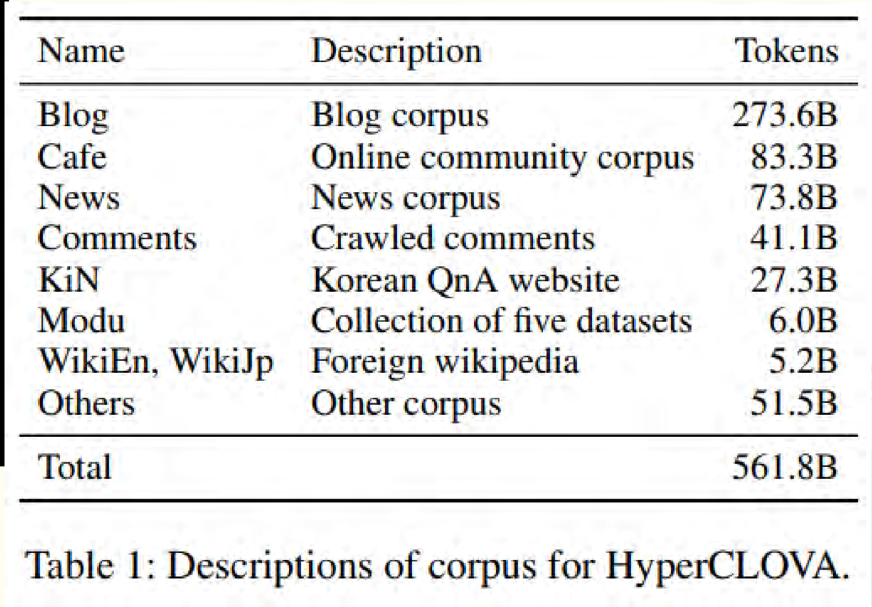


(source: arXiv:2104.08758)

HyperCLOVA 및 학습데이터

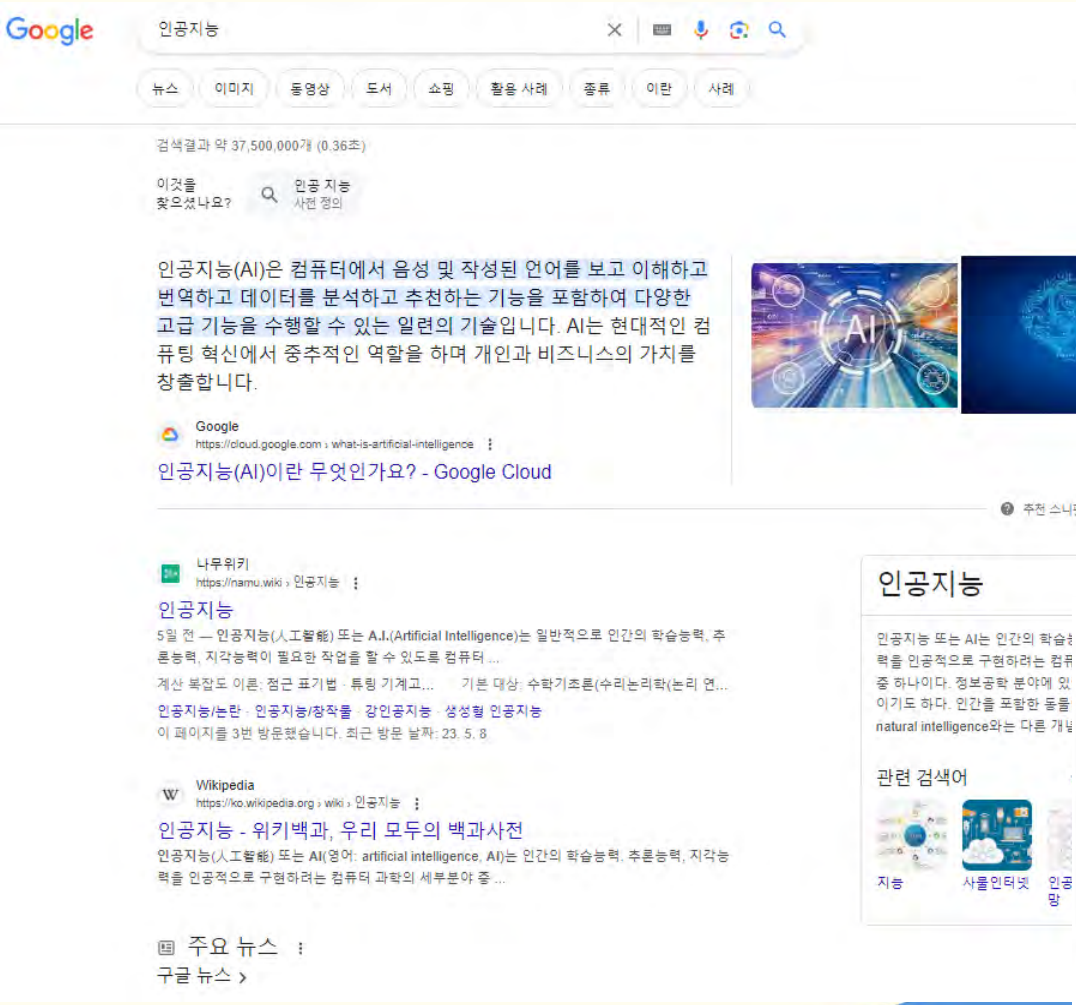


(source: naver.com)



Name	Description	Tokens
Blog	Blog corpus	273.6B
Cafe	Online community corpus	83.3B
News	News corpus	73.8B
Comments	Crawled comments	41.1B
KiN	Korean QnA website	27.3B
Modu	Collection of five datasets	6.0B
WikiEn, WikiJp	Foreign wikipedia	5.2B
Others	Other corpus	51.5B
Total		561.8B

Table 1: Descriptions of corpus for HyperCLOVA.



인공지능

검색결과 약 37,500,000개 (0.36초)

이것을 찾으셨나요? 인공 지능 사전 정의

인공지능(AI)은 컴퓨터에서 음성 및 작성된 언어를 보고 이해하고 번역하고 데이터를 분석하고 추천하는 기능을 포함하여 다양한 고급 기능을 수행할 수 있는 일련의 기술입니다. AI는 현대적인 컴퓨팅 혁신에서 중추적인 역할을 하며 개인과 비즈니스의 가치를 창출합니다.

Google
https://cloud.google.com › what-is-artificial-intelligence

인공지능(AI)이란 무엇인가요? - Google Cloud

추천 스니펫 정보 · 사용자 의견

나무위키
https://namu.wiki › 인공지능

인공지능

5일 전 — **인공지능**(人工智能) 또는 **A.I.**(Artificial Intelligence)는 일반적으로 인간의 학습능력, 추론능력, 지각능력이 필요한 작업을 할 수 있도록 컴퓨터 ...

계산 복잡도 이론: 점근 표기법 · 튜링 기계고...   기본 대상: 수학기초론(수리논리학(논리 연...

인공지능/논란 · 인공지능/창작물 · 강인공지능 · 생성형 인공지능

이 페이지를 3번 방문했습니다. 최근 방문 날짜: 23. 5. 8

Wikipedia
https://ko.wikipedia.org › wiki › 인공지능

인공지능 - 위키백과, 우리 모두의 백과사전

인공지능(人工智能) 또는 **AI**(영어: artificial intelligence, **AI**)는 인간의 학습능력, 추론능력, 지각능력을 인공적으로 구현하려는 컴퓨터 과학의 세부분야 중 ...

주요 뉴스

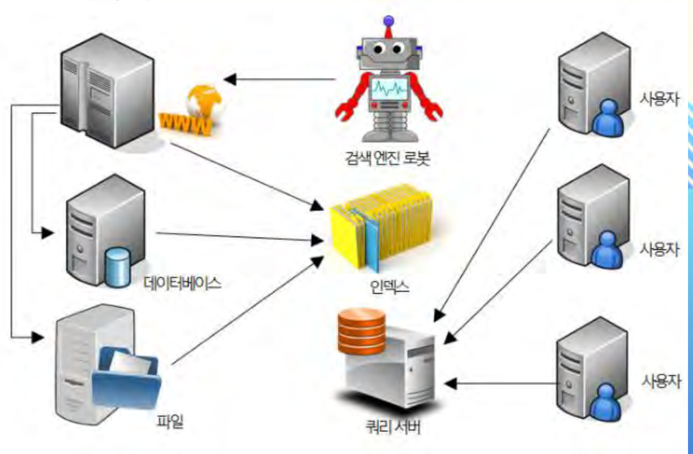
구글 뉴스 ›

인공지능

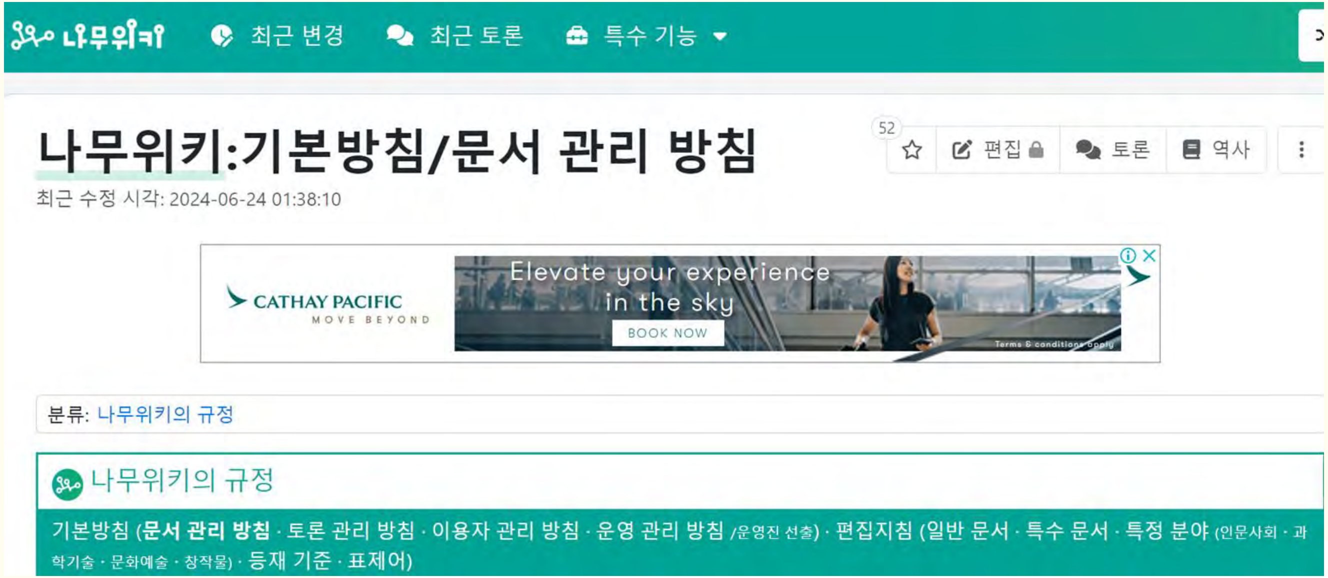
인공지능 또는 AI는 인간의 학습
력을 인공적으로 구현하려는 컴퓨
중 하나이다. 정보공학 분야에 있
이기도 하다. 인간을 포함한 동물
natural intelligence와는 다른 개념

관련 검색어

지능 · 사물인터넷 · 인공 망



나무위키 저작권 정책



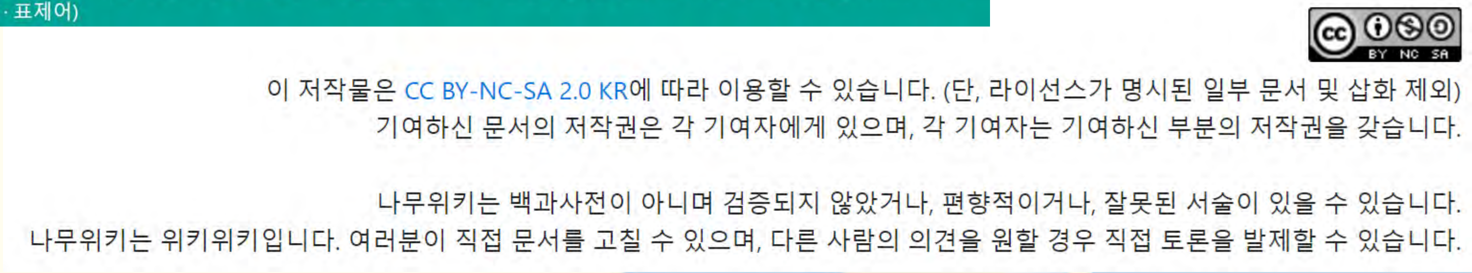
나무위키:기본방침/문서 관리 방침

최근 수정 시각: 2024-06-24 01:38:10

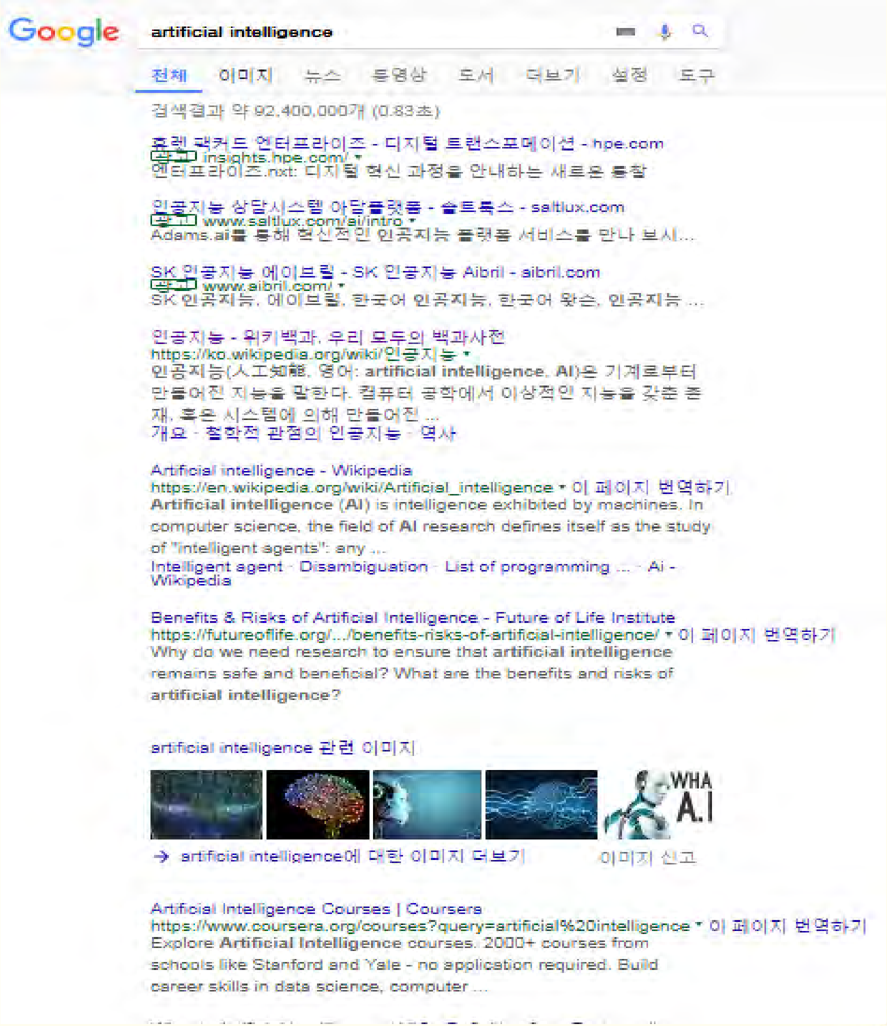
분류: 나무위키의 규정

나무위키의 규정

기본방침 (**문서 관리 방침** · 토론 관리 방침 · 이용자 관리 방침 · 운영 관리 방침 (운영진 선출) · 편집지침 (일반 문서 · 특수 문서 · 특정 분야 (인문사회 · 과학기술 · 문화예술 · 창작물) · 등재 기준 · 표제어)



나무위키는 백과사전이 아니며 검증되지 않았거나, 편향적이거나, 잘못된 서술이 있을 수 있습니다.
나무위키는 위키위키입니다. 여러분이 직접 문서를 고칠 수 있으며, 다른 사람의 의견을 원할 경우 직접 토론을 발제할 수 있습니다.



artificial intelligence

검색결과 약 92,400,000개 (0.83초)

휴렛 팩커드 엔터프라이즈 - 디지털 트랜스포메이션 - hpe.com
광고 insights.hpe.com/ ▾
엔터프라이즈.nxt: 디지털 혁신 과정을 안내하는 새로운 통찰

인공지능 상담시스템 아담플랫폼 - 솔트룩스 - saltlux.com
광고 www.saltlux.com/ai/intro ▾
Adams.ai를 통해 혁신적인 인공지능 플랫폼 서비스를 만나 보시...

SK 인공지능 에이브릴 - SK 인공지능 Aibril - aibril.com
광고 www.aibril.com/ ▾
SK 인공지능, 에이브릴, 한국어 인공지능, 한국어 왓슨, 인공지능 ...

인공지능 - 위키백과, 우리 모두의 백과사전
https://ko.wikipedia.org/wiki/인공지능 ▾
인공지능(人工知能, 영어: **artificial intelligence**, **AI**)은 기계로부터 만들어진 지능을 말한다. 컴퓨터 공학에서 이상적인 지능을 갖춘 존재, 혹은 시스템에 의해 만들어진 ...
개요 · 철학적 관점의 인공지능 · 역사

Artificial intelligence - Wikipedia
https://en.wikipedia.org/wiki/Artificial_intelligence ▾ 이 페이지 번역하기
Artificial intelligence (**AI**) is intelligence exhibited by machines. In computer science, the field of **AI** research defines itself as the study of "intelligent agents": any ...
Intelligent agent · Disambiguation · List of programming ... · Ai - Wikipedia

Benefits & Risks of Artificial Intelligence - Future of Life Institute
https://futureoflife.org/.../benefits-risks-of-artificial-intelligence/ ▾ 이 페이지 번역하기
Why do we need research to ensure that **artificial intelligence** remains safe and beneficial? What are the benefits and risks of **artificial intelligence**?

artificial intelligence 관련 이미지

→ artificial intelligence에 대한 이미지 더보기 이미지 신고

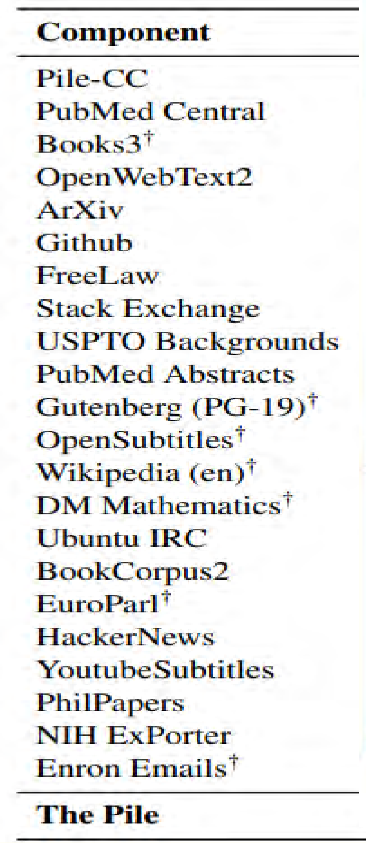
Artificial Intelligence Courses | Coursera
https://www.coursera.org/courses?query=artificial%20intelligence ▾ 이 페이지 번역하기
Explore **Artificial Intelligence** courses. 2000+ courses from schools like Stanford and Yale - no application required. Build career skills in data science, computer ...

- **Field v. Google, Inc.**
- **Perfect 10, Inc. v. Google, Inc.**
- **AFP v. Google, Inc.**
- **Viacom International Inc. v. YouTube, Inc.**
- **Authors Guild, Inc. V. Google, Inc.**

03

Open Source AI ?

자연어 모델



Component
Pile-CC
PubMed Central
Books3†
OpenWebText2
ArXiv
Github
FreeLaw
Stack Exchange
USPTO Backgrounds
PubMed Abstracts
Gutenberg (PG-19)†
OpenSubtitles†
Wikipedia (en)†
DM Mathematics†
Ubuntu IRC
BookCorpus2
EuroParl†
HackerNews
YoutubeSubtitles
PhilPapers
NIH ExPorter
Enron Emails†
The Pile

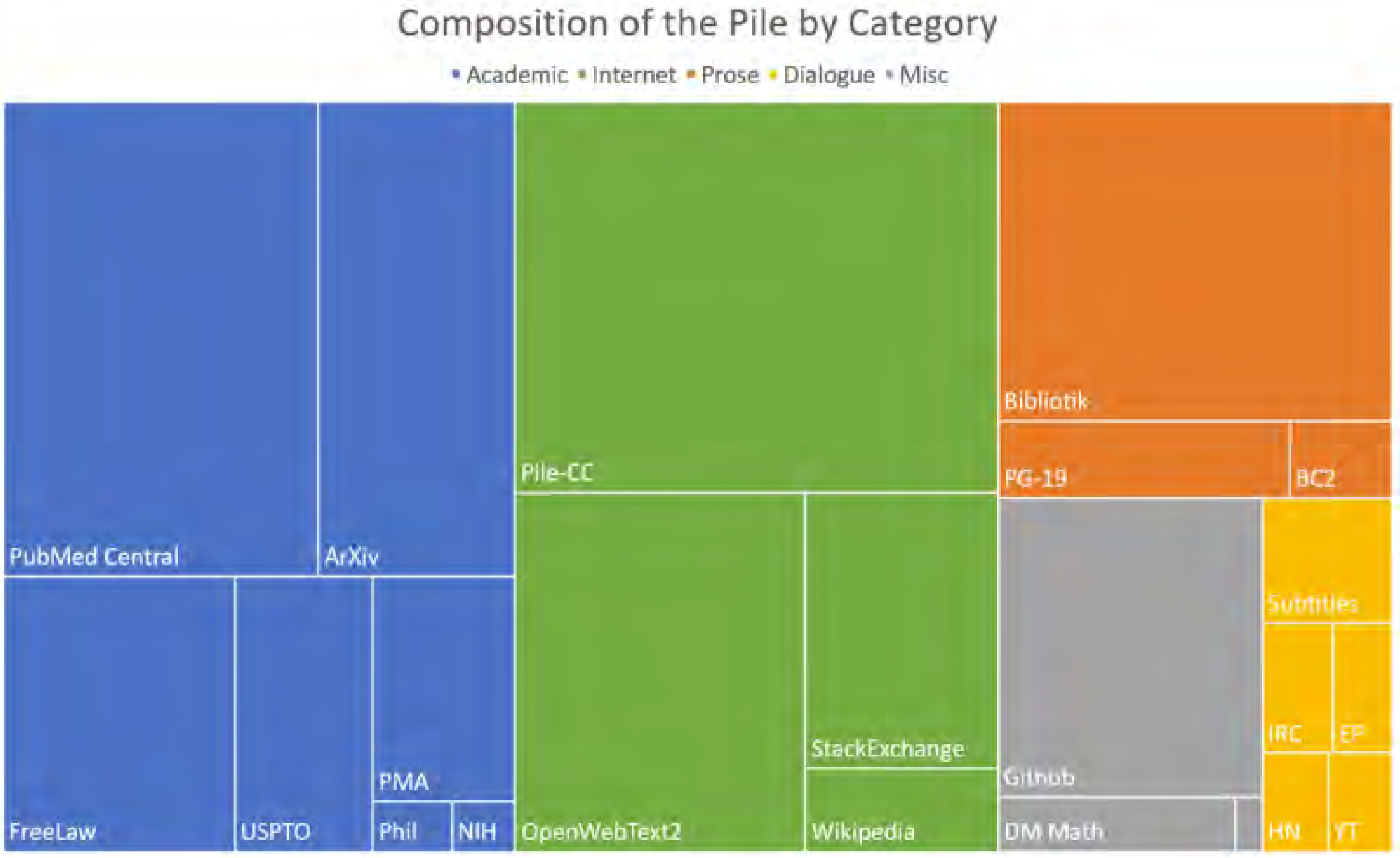
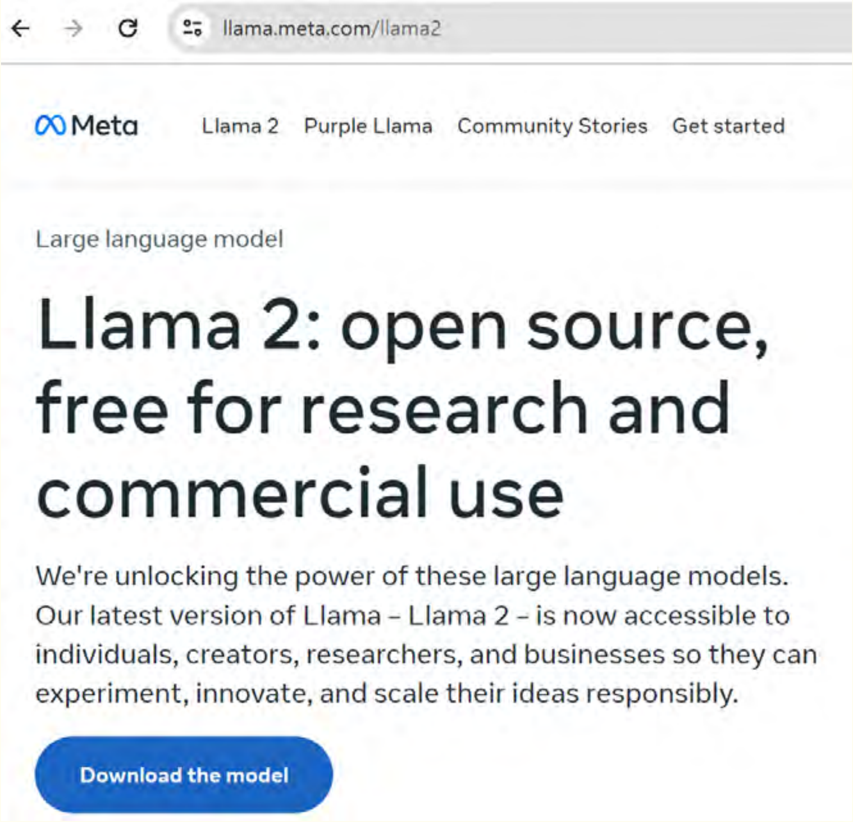


Figure 1: Treemap of Pile components by effective s

(source: arXiv:2101.00027)

LLaMA의 학습데이터

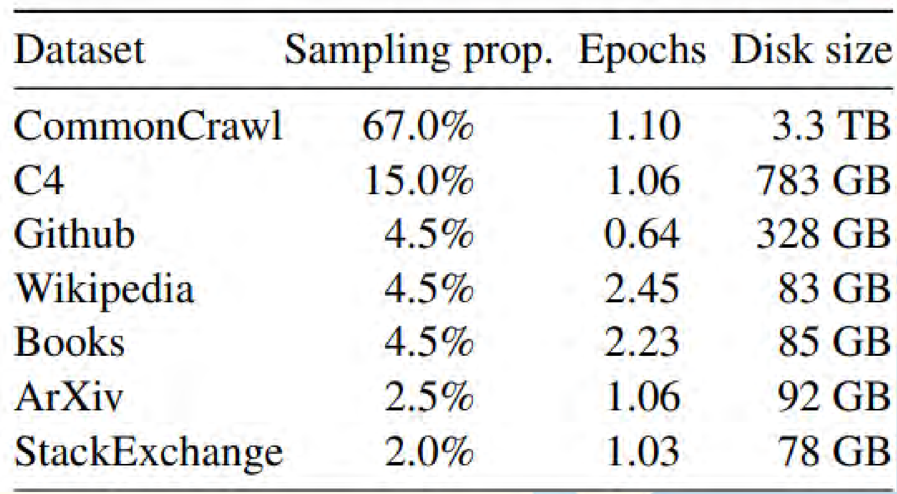


llama.meta.com/llama2

Large language model

Llama 2: open source, free for research and commercial use

We're unlocking the power of these large language models. Our latest version of Llama – Llama 2 – is now accessible to individuals, creators, researchers, and businesses so they can experiment, innovate, and scale their ideas responsibly.



Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

(source: arXiv:2104.08758)

LLAMA 3.1 COMMUNITY LICENSE AGREEMENT

1. License Rights and Redistribution. b. Redistribution and Use.

i. If you distribute or make available the Llama Materials .. you shall (A) provide a copy of this Agreement with any such Llama Materials; and (B) prominently display "Built with Llama" on a related website, user interface, blogpost, about page, or product documentation. If you use the Llama Materials or any outputs or results of the Llama Materials to create, train, fine tune, or otherwise improve an AI model, which is distributed or made available, you shall also include "Llama" at the beginning of any such AI model name. …

iv. Your use of the Llama Materials must comply with applicable laws and regulations (including trade compliance laws and regulations) and adhere to the Acceptable Use Policy for the Llama Materials (available at https://llama.com/llama3_1/use-policy), which is hereby incorporated by reference into this Agreement.

2. Additional Commercial Terms. If, on the Llama 3.1 version release date, the monthly active users of the products or services made available by or for Licensee, or Licensee's affiliates, is greater than 700 million monthly active users in the preceding calendar month, you must request a license from Meta, which Meta may grant to you in its sole discretion, and you are not authorized to exercise any of the rights under this Agreement unless or until Meta otherwise expressly grants you such rights.

The Open Source AI Definition - draft v. 0.0.8

What is Open Source AI

An Open Source AI is an AI system made available under terms that grant the freedoms to:

- Use the system for any purpose and without having to ask for permission.
- Study how the system works and inspect its components.
- Modify the system for any purpose, including to change its output.
- Share the system for others to use with or without modifications, for any purpose.

Precondition to exercise these freedoms is to have access to the preferred form to make modifications to the system.

(Source : opensource.org)

감사합니다

오픈소스/ 데이터 라이선스의 이해