15 JULIOL 2022 | De 10:00h a 11:30h

Jornada AINA en acció:

resultats, experiències i expectatives de la indústria

https://www.youtube.com/user/BSCCNS

Generalitat de Catalunya
Departament de la Vicepresidència
i de Polítiques Digitals i Territori

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Recordem

OBJECTIUS

1

Proveir el català de la **infraestructura** necessària **per al desenvolupament d'aplicacions basades en IA/TL,** (assistents de veu, traductors automàtics, agents conversacionals, etc)

2

Fer que **la inclusió del català** a les aplicacions de IA/TL sigui **rendible i atractiva per a les empreses del sector**, tant a nivell local com global.

3

Aconseguir que el ciutadà de Catalunya pugui **participar en català** en el món digital **al mateix nivell que un parlant d'una llengua global**, com ara l'anglès o el castellà.

DEFINICIÓ i FONAMENTS

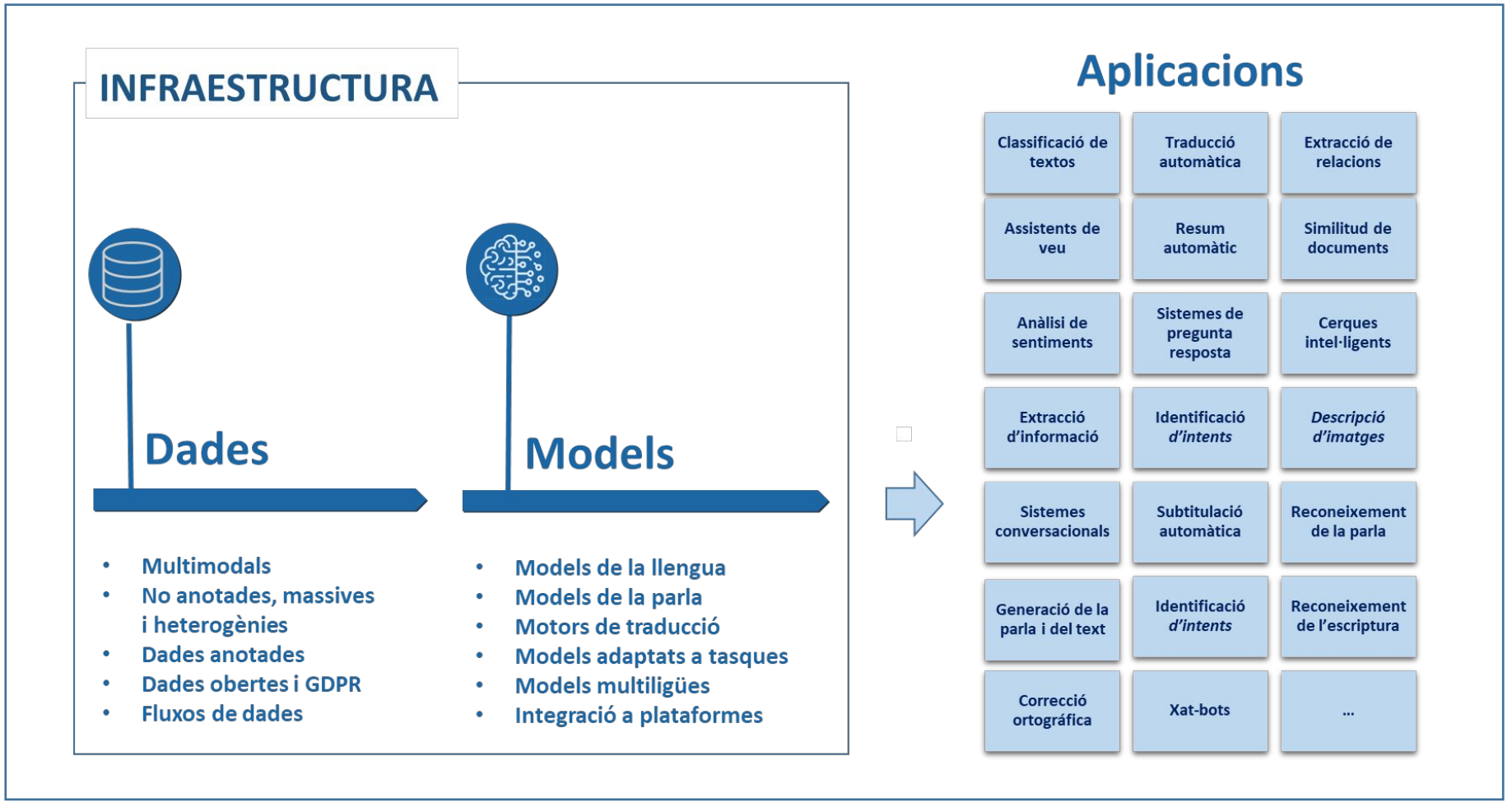
- AINA és essencialment **infraestructura** lingüística.
- El valor de les **dades**.
  - La tecnologia avança molt ràpidament però **les dades són persistents**.
  - Únicament des de la **iniciativa pública**, el català es pot garantir el subministrament de dades suficients.
  - Disposar de **dades de qualitat suficients és un actiu segur i de futur** que garanteix l'actualització de la tecnologia.

ESTRATÈGIA

- AINA implementarà una infraestructura de recollida i neteja de dades amb la **implicació de grans actors.**
- AINA reaccionarà ràpidament als **avenços tecnològics** mitjançant la vigilància tecnològica.
- AINA detectarà i donarà resposta a **noves necessitats** de les empreses i de la societat mitjançant la vigilància sectorial i de mercat.

Barcelona, 15 de febrer del 2022

Recordem


INFRAESTRUCTURA
Dades
Multimodals
No anotades, massives i heterogènies
Dades anotades
Dades obertes i GDPR
Fluxos de dades
Models
Models de la llengua
Models de la parla
Motors de traducció
Models adaptats a tasques
Models multilingües
Integració a plataformes
Aplicacions
Classificació de textos
Traducció automàtica
Extracció de relacions
Assistents de veu
Resum automàtic
Similitud de documents
Anàlisi de sentiments
Sistemes de pregunta resposta
Cerques intel·ligents
Extracció d'informació
Identificació d'intents
Descripció d'imatges
Sistemes conversacionals
Subtitulació automàtica
Reconeixement de la parla
Generació de la parla i del text
Identificació d'intents
Reconeixement de l'escriptura
Correcció ortográfica
Xat-bots
...

Dades textuals

Nou corpus del català, amb el que s'ha entrenat el nou model RoBERTa

Corpus	Size in GB	State
CaCrawlat	13.00	New
Wikipedia	1.10	Updated (04-2022)
DOGC	0.78	Updated
Catalan Open Subtitles	0.02	Updated
Catalan Oscar	4.00	Updated
CaWaC	3.60	Updated
Cat. General Crawling	2.50	Updated
Cat. Goverment Crawling	0.24	Updated
ACN	0.42	Updated
Padicat	0.63	New
RacoCatalà	8.10	New
Nació Digital	0.42	New
Vilaweb	0.06	New

9GB -> 30GB

Noves incorporacions en curs

- Corpus de **textos tècnics** , amb centenars de milers de tesis i articles de revistes en català (**RACO** i **TDX**)
- Corpus de textos administratius de diversos àmbits, procedents de l'Administració Pública

Dades anotades

LICITACIONS

Se ha dividido la **licitación en 9 lotes**:

1. Traducción de corpus de referencia multilingües.
2. Anotación y vinculación de entidades nombradas
3. Creación de un conjunto de preguntas y respuestas.
4. Anotación de polaridad, emociones y opinión.
5. Creación y anotación de un corpus de NLU para el catalán.
6. Creación y anotación de un corpus de lenguaje abusivo.
7. Creación de un corpus de resúmenes extractivos y abstractivos.
8. Adquisición y traducción de un corpus bilingüe
9. Segmentación, alineación y traducción de corpus de voz.

ANOTACIONS en CURS

Paràfrasi: 11 mil frases, cadascuna amb una frase que és paràfrasi, i un altre que no ho és.

Negació: +20 mil frases, amb exemples de negació anotades amb marcadors de la negació, i el seu focus i abast.

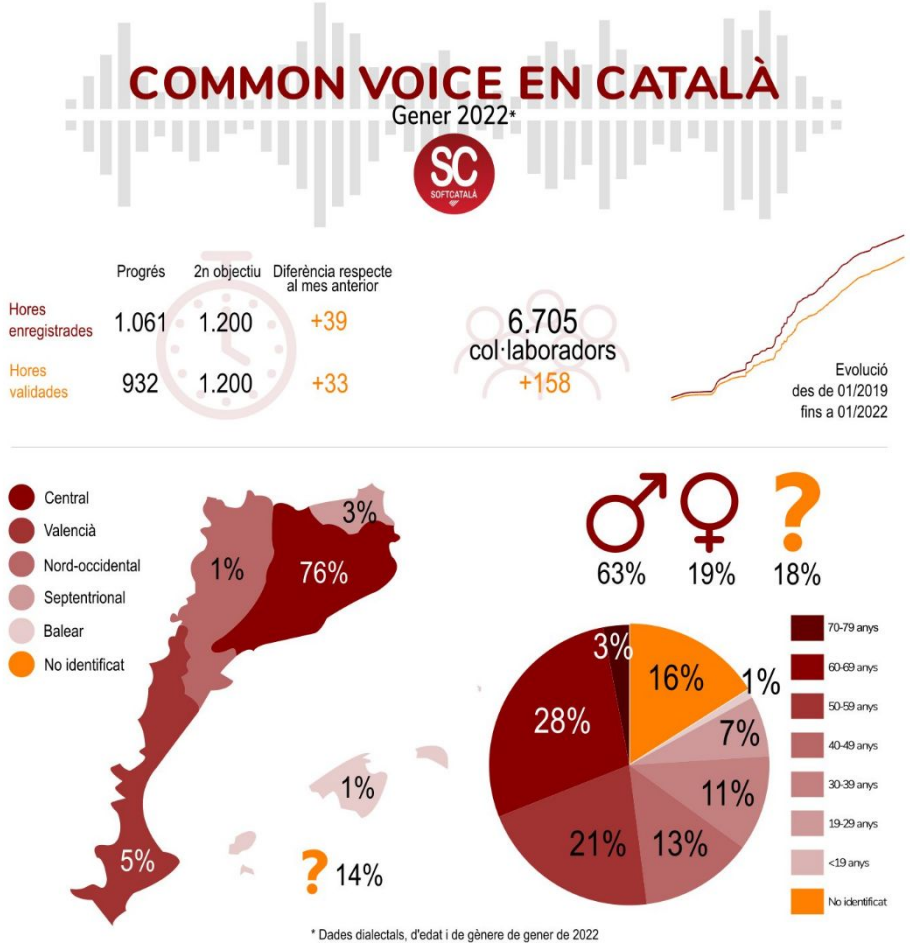
Datasets conversacionals:

> **Xit-Xat**: Corpus de converses simulades de xatbots de servei a client, de 10 dominis diferents, com ara lloguer de vehicles o habitacions, ajuntaments, transports, comerç electrònic,…. Amb anotacions d'intents i entities.
>
> Incorporació del català al més gran dataset multilingüe d'intents, el corpus **MASSIVE** , per assistents virtuals, de la mà del equip **d'Alexa a Amazon**.

Corpus de **ressenyes** de restaurants de **guiacat.cat**, amb textos d'opinió i valoracions

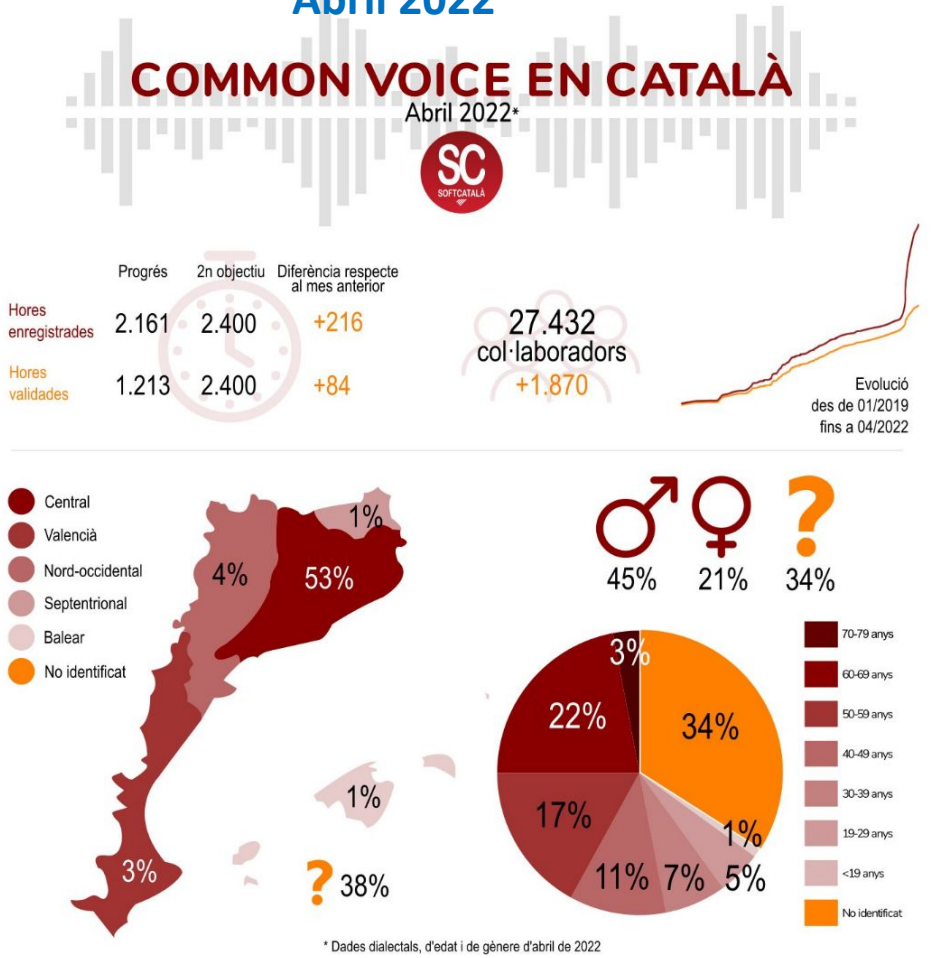
Dades de veu

Gener 2022



* Dades dialectals, d'edat i de gènere de gener de 2022

Abril 2022



* Dades dialectals, d'edat i de gènere d'abril de 2022

Dades de veu

1000h manualment validades
10.000h subtítols

Parlament API

- ParlamentParla v2, publicat el 2021:
  - https://zenodo.org/record/5541827 (+1500 descàrregues)
  - https://huggingface.co/datasets/projecte-aina/parlament_parla
- Nova RESTful API d'accés a les dades del Parlament facilitant-ne l'actualització i la descàrrega. Col·laboració amb el Parlament endpoint. Properament es farà públic s'obrirà.

CCMA

- Acord de cessió de dades signat.
- Hem identificat 37.000 hores subtitulades en català.

IB3

- Acord de cessió de dades signat
- Treballant per a compilació de dades.

Youtube datapipe:

- Eina per generar un corpus de la parla a partir dels continguts audiovisuals amb llicències obertes a la web (originalment desenvolupat per la comunitat) https://github.com/projecte-aina/datapipe
- Hem identificat 3.000 hores de contingut audiovisual amb llicència CC-BY en català

Models

Nou RoBERTa

Task	NER (F1)	POS (F1)	STS (Pearson)	TC (accuracy)	QA (ViquiQuAD) (F1/EM)	QA (XQuAD) (F1/EM)
RoBERTa-base-ca-v2	**89.84**	**99.07**	**79.98**	**83.41**	**88.04/74.65**	**71.50/53.41**
BERTa	88.13	98.97	79.73	74.16	86.97/72.29	68.89/48.87
mBERT	86.38	98.82	76.34	70.56	86.97/72.22	67.15/46.51
XLM-RoBERTa	87.66	98.89	75.40	71.68	85.50/70.47	67.10/46.42
WikiBERT-ca	77.66	97.60	77.18	73.22	85.45/70.75	65.21/36.60

Table 2: Evaluation results comparison.

Models

HUGGING FACE

Models 17 — Collapse — Sort: Most Downloads

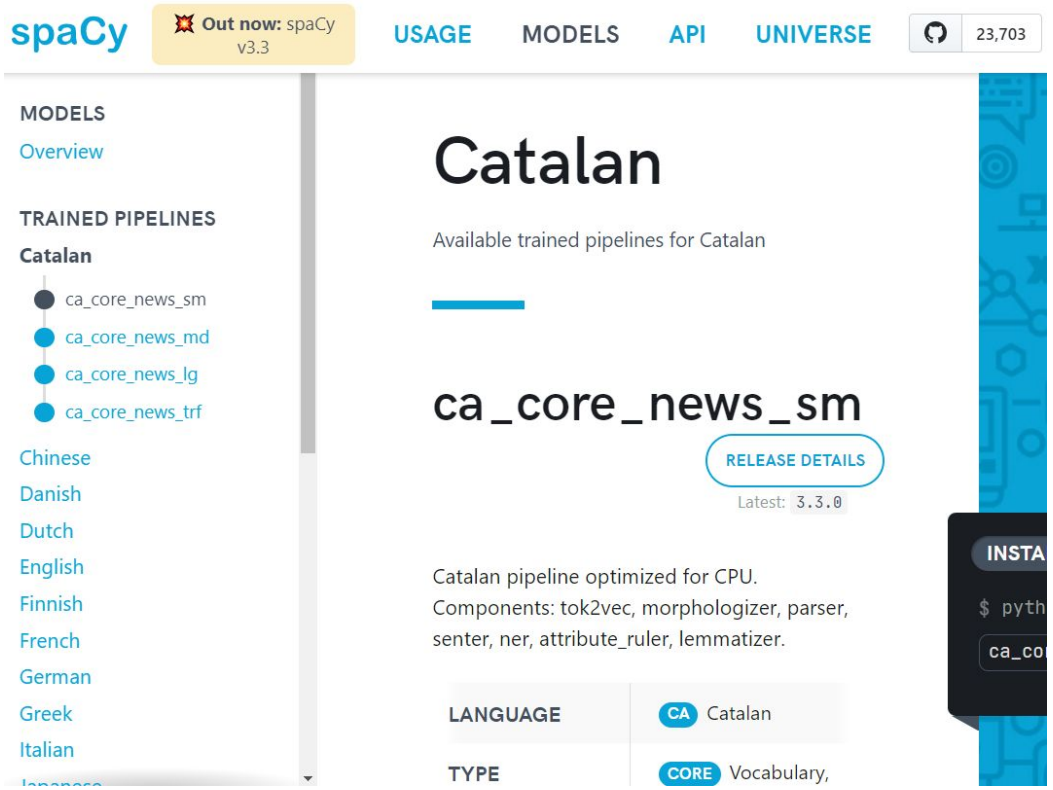
- projecte-aina/roberta-base-ca-cased-ner — Token Classification • Updated 19 days ago • ↓ 138 • ♡ 1
- projecte-aina/roberta-base-ca-cased-tc — Text Classification • Updated Feb 24 • ↓ 116 • ♡ 1
- projecte-aina/roberta-base-ca-cased-pos — Token Classification • Updated 20 days ago • ↓ 107
- projecte-aina/mbert-base-gencata — Text Classification • Updated 10 days ago • ↓ 87
- projecte-aina/roberta-base-ca-cased-sts — Text Classification • Updated 18 days ago • ↓ 71
- projecte-aina/roberta-base-ca-cased-qa — Question Answering • Updated Feb 24 • ↓ 45
- projecte-aina/bart-base-ca-casum — Summarization • Updated Feb 16 • ↓ 35
- projecte-aina/roberta-base-ca-v2 private — Fill-Mask • Updated Jun 3 • ↓ 33
- projecte-aina/roberta-base-ca-cased-te — Text Classification • Updated Feb 24 • ↓ 27
- projecte-aina/bart-base-ca — Text2Text Generation • Updated Feb 15 • ↓ 23
- projecte-aina/roberta-base-ca-v2-cased-sts private — Text Classification • Updated 4 days ago • ↓ 13
- projecte-aina/roberta-base-ca-v2-cased-qa private — Question Answering • Updated 4 days ago • ↓ 7
- projecte-aina/roberta-base-ca-v2-cased-tc private — Text Classification • Updated 4 days ago • ↓ 7
- projecte-aina/mbert-base-gencata2 private — Text Classification • Updated 10 days ago
- projecte-aina/roberta-base-ca-v2-cased-ner private — Token Classification • Updated 4 days ago
- projecte-aina/roberta-base-ca-v2-cased-te private — Text Classification • Updated 4 days ago
- projecte-aina/roberta-base-ca-v2-cased-pos private — Token Classification • Updated 4 days ago

Models

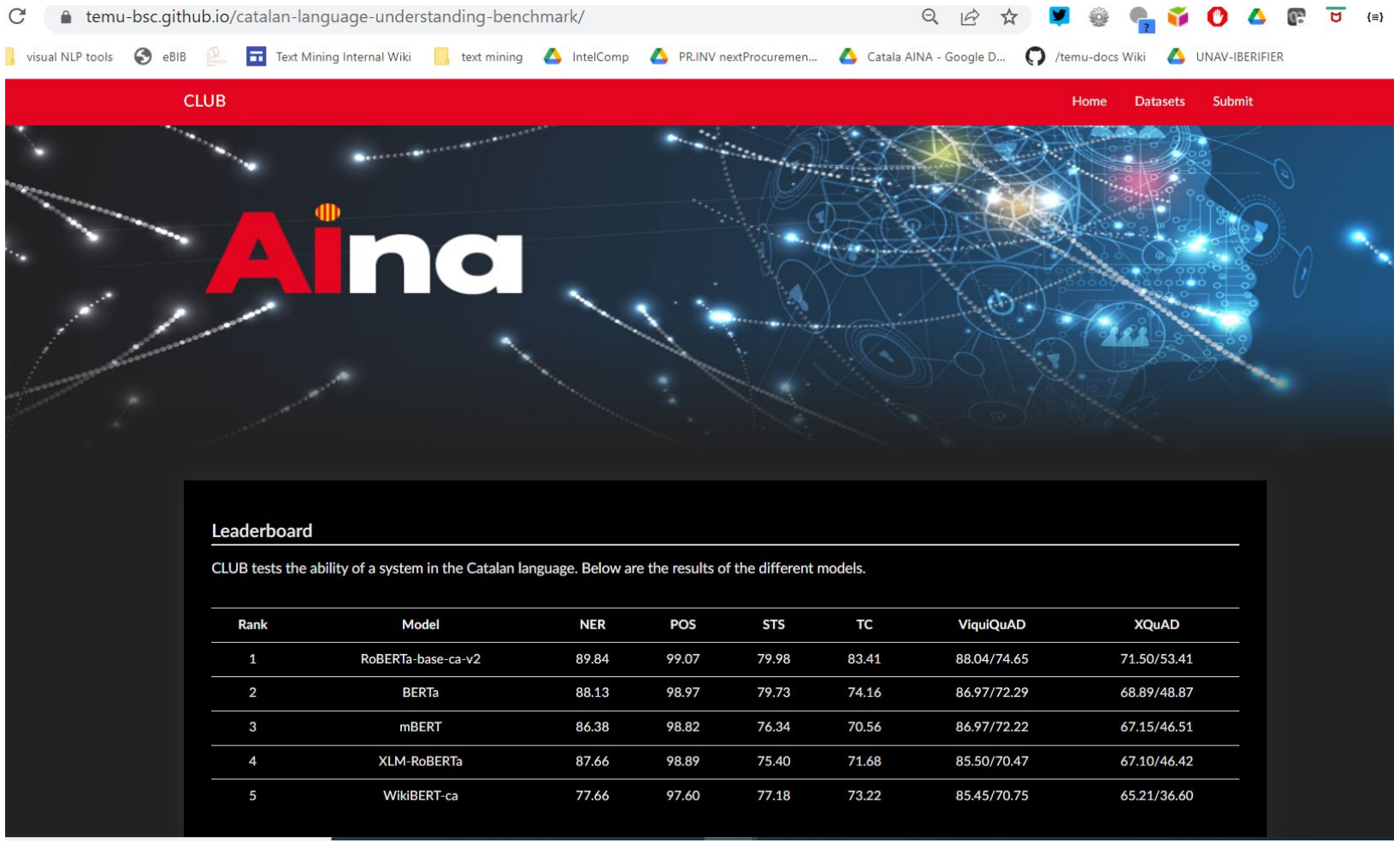
Amb models i datasets d'AINA, s'han creat dues cadenes de processament de català per a plataformes robustes d'ús industrial, i gratuïtes.

- **Spacy (https://spacy.io:)** Amb tokenització de frase i paraula, lemmatització, POS, anàlisi morfològic, detecció de sintagmes nominals i NER, i parsing de dependencies. Properament s'afegiran mòduls de classificació de textos i vectors de similitud léxica i de frase.

- **Sparknlp (https://nlp.johnsnowlabs.com)**: Amb tokenització de frase i paraula, lemmatització, reducció de dimensionalitats, POS, detecció de sintagmes nominals i NER, i vectors de similitud léxica i de frase. (https://nlp.johnsnowlabs.com/models?language=ca&q=projecte-aina)



Benchmarking



Rank	Model	NER	POS	STS	TC	ViquiQuAD	XQuAD
1	RoBERTa-base-ca-v2	89.84	99.07	79.98	83.41	88.04/74.65	71.50/53.41
2	BERTa	88.13	98.97	79.73	74.16	86.97/72.29	68.89/48.87
3	mBERT	86.38	98.82	76.34	70.56	86.97/72.22	67.15/46.51
4	XLM-RoBERTa	87.66	98.89	75.40	71.68	85.50/70.47	67.10/46.42
5	WikiBERT-ca	77.66	97.60	77.18	73.22	85.45/70.75	65.21/36.60

Dades de l'Administració Pública (campanya de recollida)

Recollida de dades textuals per al projecte AINA

L'objectiu del projecte AINA, impulsat pel Departament de Polítiques Digitals de la Generalitat de Catalunya, és crear els recursos tecnològics necessaris per facilitar la inclusió del català a les aplicacions d'intel·ligència artificial. La base d'aquests recursos són les dades lingüístiques (text o veu). Mitjançant aquest formulari, els proveïdors de continguts, del sector públic i privat, poden contribuir amb els seus corpus textuals a la generació dels recursos necessaris. Trobareu més informació del projecte AINA a https://t.co/aBscnfiRdK.
Els tipus principals de corpus que es recullen aquí són:
- conjunts de documents en català o en alguna altra llengua, en qualsevol format, incloent pdf, però preferentment word o text.
- conjunts de documents en català i la seva corresponent traducció a una altra llengua
- memòries de traducció

Per tal de contribuïr amb un o més corpus al projecte cal seguir aquestes instruccions:
- Guardeu els documents que constitueixen un corpus específic en un únic arxiu comprimit (p.e. zip, .rar, .tar, .gz, .tgz). Nota: el que defineix un corpus específic és que tots els documents tinguin el mateix format, la mateixa llicència i preferiblement siguin d'un domini concret.
- Pugeu l'arxiu comprimit al Servidor de fitxers d'AINA accedint a aquest enllaç: https://b2drop.eudat.eu/s/9KZpEfLJmBiWDJ5
- Ompliu un formulari per cada arxiu comprimit que pugeu al Servidor de fitxers d'AINA. És molt important que el nom d'arxiu que us demanem més avall correspongui amb el de l'arxiu que heu pujat al Servidor de fitxers AINA.

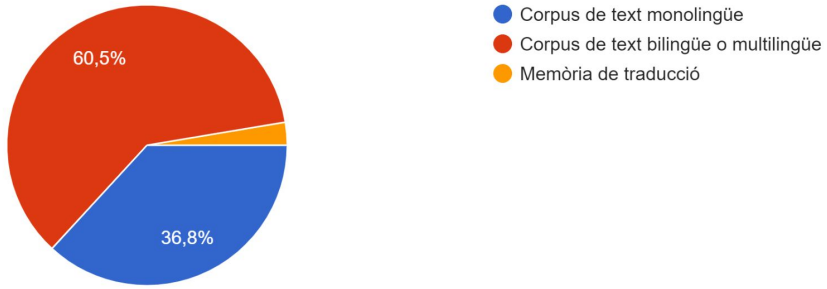
Per consultes o suggeriments, adreceu-vos a: maite.melero@bsc.es

- L'AP té grans volums de dades lingüístiques de qualitat (textuals, sonores o audiovisuals).
- La Llei sobre la Reutilització de la Informació del Sector Públic ha impulsat, des del 2015, la creació dels portals de dades obertes
- Fins ara, responsables polítics i gestors de dades obertes no coneixien el valor afegit de les dades lingüístiques i es feia poc esforç per recopilar-les i publicar-les en obert.

Dades de l'Administració Pública (campanya de recollida)

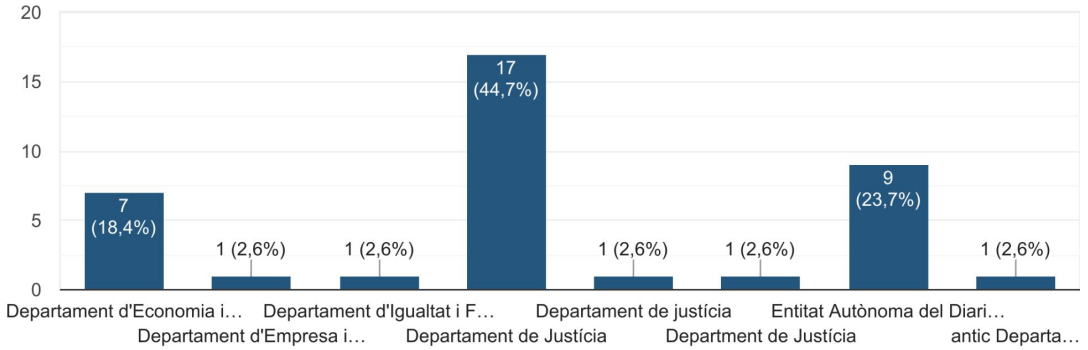
Tipus de dades textuals

38 respostes



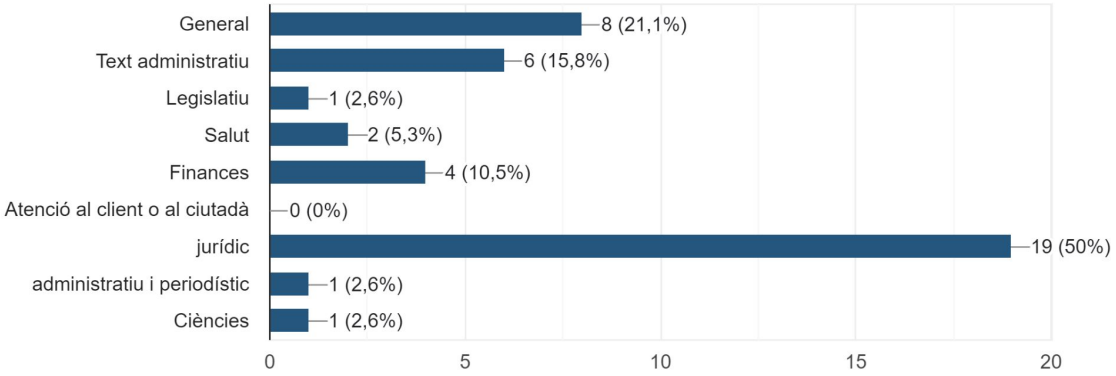
Proveïdor

38 respostes



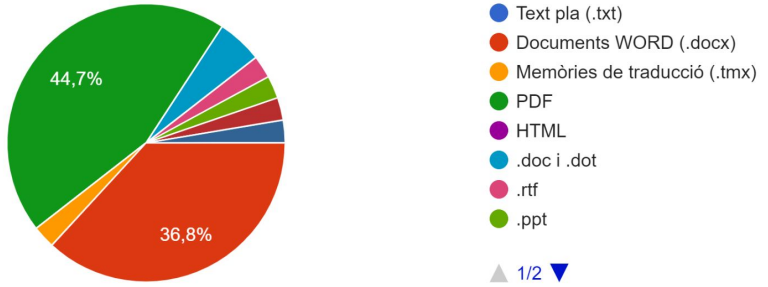
Domini(s) (si no ho sabeu marqueu General)

38 respostes



Format de les dades

38 respostes



Models de Traducció Automàtica

mBART
(*model multilingüe pre-entrenat*)

→ Adaptació a Traducció Automàtica

→ Optimització de la traducció

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

corpus bilingüe cat-ang (9,68 M)

corpus bilingüe cat-esp (18,58 M)

corpus castellà (867 M)

corpus anglès (162 M)

corpus català (73 M)

cat-xinès (M2M-100 + Wikipedia (0,299 M))
cat-alemany (M2M-100 + Softcatalà (5,21 M))

Neural Machine Translation reaches historic milestone: human parity for Chinese to English translations

Posted on March 14, 2018 by Microsoft Translator

Microsoft announced today that its researchers have developed an AI machine translation system that can translate with the same accuracy as a human from Chinese to English.

To validate the results, the researchers used an industry-standard test set of news stories (newstest2017) to compare human and machine translation results.

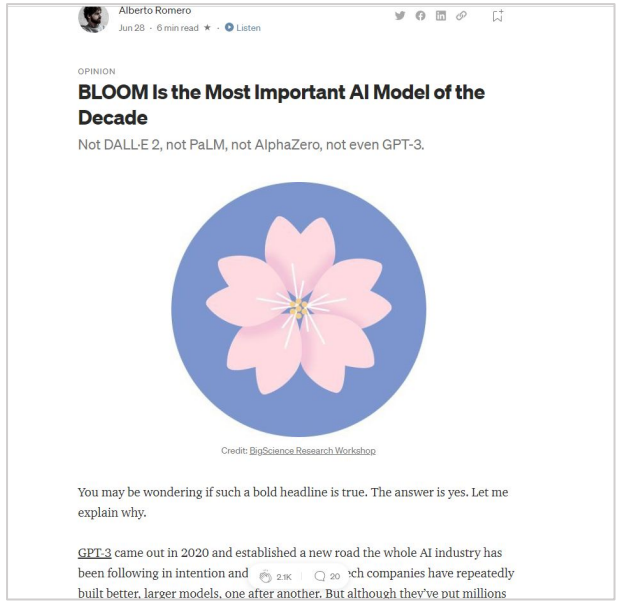
/en-us/translator/blog/wp-content/uploads/sites/13/2018/03/Translator-Graphic-1200x630.jpg

Català al BigScience

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access



Alberto Romero

Jun 28 · 6 min read ★ · Listen

OPINION

BLOOM Is the Most Important AI Model of the Decade

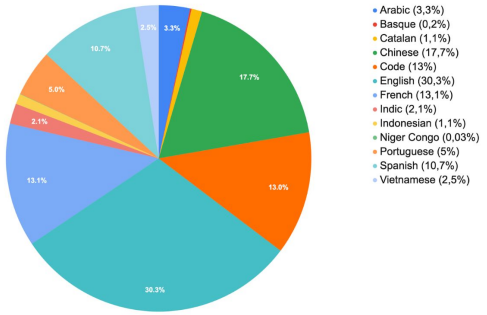
Not DALL·E 2, not PaLM, not AlphaZero, not even GPT-3.

Credit: BigScience Research Workshop

You may be wondering if such a bold headline is true. The answer is yes. Let me explain why.

GPT-3 came out in 2020 and established a new road the whole AI industry has been following in intention and 2.1K 20 ·ch companies have repeatedly built better, larger models, one after another. But although they've put millions

Objectiu: obrir les dades i el model per investigar: biaix, impacte social, capacitats, limitacions, ètica, possibles millores, rendiment en dominis específics, petjada de carboni, etc.

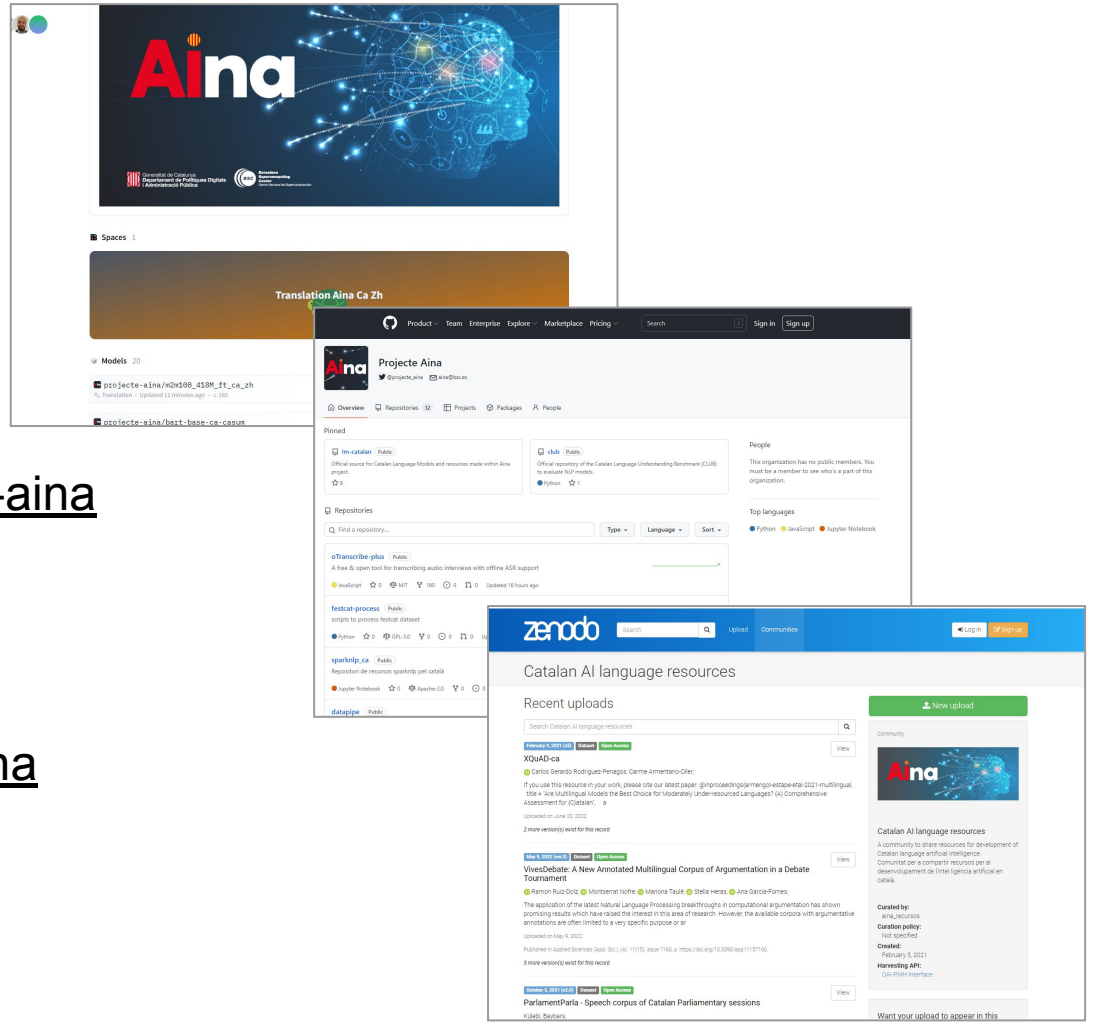


BigScience és un projecte col·laboratiu promogut per Hugging Face (+1000 investigadors, +60 països, +250 institucions)

- s'han recollit grans quantitats de dades multilingües
- S'ha creat BLOOM un **model generatiu multilingüe** entrenat durant 4 mesos al superordinador Jean Zay a París.

Gràcies a **Aina**, el català hi està inclòs

On trobar-ho?



HUGGING FACE

https://huggingface.co/projecte-aina

GitHub

https://github.com/projecte-aina

zenodo

https://zenodo.org/communities/catalan-ai

Demostradors

1. ***Natural Language Understanding***

Demostració de les capacitats de la plataforma Spacy entrenada amb models i dades d'AINA per fer **comprensió d'un text, detectant tema, entitats, relacions, etc**.

NLU és a la base de moltes aplicacions com ara xatbots.

2. **Transcripció automàtica**

oTranscribe+ és una eina de transcripció automàtica que facilita també l'edició. Transcriu els enregistraments automàticament, sense compartir els àudios amb un servei extern, i fent servir els **models de reconeixement de la parla** emmagatzemats en local.

L'eina està disponible sota llicència oberta MIT i es pot provar a https://otranscribe.bsc.es/

3. **Pregunta/Resposta a la Viquipèdia**

Demostració del **model extractiu de pregunta/resposta** "roberta-base-ca-v2-cased-qa".

Donat un tema o pàgina, i una pregunta, pot trobar el fragment on hi ha la resposta.


1million bot


nuclia


M47
LABS


parlem
telecom


PLATA
FORMA
PER LA
LLENGUA


bookline


basetis


Sabadell


PARLAMENT
DE CATALUNYA


OMNIOS