




PROCESSAMENT DE LLENGUATGE NATURAL



Pau Miró Fàbregas i Isaac Ruiz García
INSTITUT SA PALOMERA

PROCESSAMENT DE LLENGUATGE NATURAL	2
ACCÉS AL CODI	2
SELECCIÓ ARTICLES	2
METODOLOGIA	2
PAS 1: PROCESSAMENT HTML.....	2
PAS 2: PREPROCESSAMENT DEL TEXT	4
PAS 3: IDENTIFICACIÓ DEL TEMA DE L'ARTICLE	6
PAS 4: GENERACIÓ DEL RESUM DE L'ARTICLE	8
CONCLUSIONS DEL PROJECTE.....	9
WEBGRAFIA	10
DOCUMENTACIÓ TÈCNICA I LLIBRERIES	10
MODELS D'INTEL·LIGÈNCIA ARTIFICIAL UTILITZATS.....	10

PROCESSAMENT DE LLENGUATGE NATURAL

ACCÉS AL CODI

Per poder accedir als **.html** i als **.py** empleats per aquesta pràctica has d'accedir al següent repositori de github on estan publicats.

[Repositori Oficial](#)

SELECCIÓ ARTICLES

Hem seleccionat cinc articles en llengua catalana de diversos mitjans de comunicació digitals de referència. La tria no ha estat aleatòria, sinó que respon a criteris tècnics per posar a prova el rendiment dels models de Processament de Llenguatge Natural (PLN) en diferents escenaris:

1. **Varietat de dominis (Temàtiques):** Hem escollit articles que cobreixen àrees molt dispars: tecnologia (MareNostrum 5), esports (Girona FC), salut i societat (Vall d'Hebron), cultura (Hitchcock i la IA) i economia (dades de l'atur). Aquesta diversitat és fonamental per avaluar la capacitat del model en el **Pas 3**, ja que permet verificar si les tècniques de vectorització (TF-IDF i embeddings) són capaces de diferenciar vocabularis tècnics molt específics.
2. **Riquesa de vocabulari:** Cada article aporta un corpus de paraules clau diferenciat. Per exemple, l'article sobre el MareNostrum 5 conté terminologia d'infraestructures tecnològiques, mentre que el de l'atur se centra en dades estadístiques i econòmiques. Això ens permetrà analitzar com les *stopwords* i la normalització afecten la representació vectorial segons el context.
3. **Complexitat de l'HTML:** Les fonts escollides (*Via Empresa*, *VilaWeb*, *Ara*, *L'Esportiu* i *El Punt Avui*) presenten estructures HTML diverses. Això ens permet treballar el **Pas 1** amb diferents graus de dificultat, assegurant que el nostre *scraper* sigui capaç de filtrar elements irrelevants com anuncis, menús i metadades.
4. **Consistència lingüística:** Tot i que els models de HuggingFace solen estar entrenats en anglès, hem optat pel català per avaluar la potència actual dels models territorials (com el projecte AINA) i garantir que el resum generat al **Pas 4** sigui fàcilment verificable per nosaltres en la nostra llengua.

METODOLOGIA

PAS 1: PROCESSAMENT HTML

Hem creat un primer codi utilitzant la llibreria **BeautifulSoup** de python. En el primer moment ha donat uns bons resultats en alguns articles però en alguns altres no.

```
python main.py

Executant Pas 1: Extracció de text...
[filter]: 'Cultura.html', 'titol': 'Hitchcock en l'era de la intel·ligència artificial', 'text_original': ''
[filter]: 'Economia.html', 'titol': 'L'atur repunta una mica però registra la xifra més baixa d'un gener des del 2008 | Lata Bruguera | Barcelona | Economia | El Punt Avui', 'text_orign
al': 'L'atur ha estremat l'any amb un lleuger repunt, del 8,0%, respecte del desembre, però la xifra actual de desocupats -325.214- és la més baixa en un mes de gener des del 2008. En var
ació interanual, a més, ha baixat un 3,4%. El lleuger repunt del gener respecte de desembre es deu gairebé exclusivament al sector serveis, que engreixa les llistes d'aturats en més de 2
.500 persones, un 1,1% més, informa l'ACN. Pel que fa a l'evolució interanual, el sector serveis es torna a mostrar menys dinàmic que la resta, amb una caiguda els darrers dotze mesos del
2,9%, lluny de les baixades que voregen el 10% a l'agricultura i la construcció, i de més del 5% en el sector de la construcció. Pel que fa a les afiliacions a la Seguretat Social, s'ha
situat en els 3,82 milions, la xifra representa gairebé un 2% més que fa un any i és la millor registrada en un mes de gener des de, com a mínim, el 2012.'
[filter]: 'Esports.html', 'titol': 'Continguts del lloc', 'text_original': 'Contratemps important pel Girona: Alex Moreno pateix una lesió al coll de la cama esquerra que el podria tenir
més d'un mes fora de la competició, a l'espera que el club ho faci oficial. El lateral es va fer mal a inicis d'aquesta setmana durant un entrenament i, de moment, encara no hi ha cap co
municat mediu oficial per acabar de veure l'abast real. S'espera que siguin unes sis setmanes de baixa. A la seva absència deixa el Girona sense un lateral esquerre per a la plantilla, una
de les posicions que el club volia reforçar durant el mercat d'hivern i que finalment no ha pogut cobrir. Michel haurà de recórrer a alternatives com Blind, Arnau o Francés per tapar un
buit important a l'equip.'
[filter]: 'Salut.html', 'titol': 'L'Hospital Vall d'Hebron fa el primer trasplantament de cara del mínim amb una donant que havia rebut l'«otànasia», 'text_original': 'El gas pebre no dist
ingeix entre hipotètics manifestants violents i manifestants pacífics, ni entre adults i infants, ni entre qui protesta i qui simplement passa per allà. Fa mal a tothom. És, per definició
, una arma de violència massiva, clarament desproporcionada'.
[filter]: 'Tecnologia.html', 'titol': 'Telefonia i Fajitsu ampliaran el MareNostrum 5 del BSC per allotjar la Factoria d'IA europea', 'text_original': 'El Barcelona Supercomputing Center
(BSC) continua avançant en la implementació de la factoria d'IA atorgada per la Comissió Europea amb una nova ampliació del MareNostrum 5. La via d'convocatòria, que va llançar-se el passat mes
de juliol, s'ha tancat aquest dilluns amb la signatura del contracte entre l'empresa conjunta de supercomputació de la UE, EuroHPC JU, i el consorci liderat per Fas Technologies (la branca
especialitzada en computació d'alt rendiment i IA de Fas) i Telefónica, que s'encarregarà de la implementació. El projecte té un pressupost aproximat de 229 milions d'euros. A més, 50
% dels quals estan finançats directament per EuroHPC i l'altra meitat per Espanya (68 milions, aportats pel govern espanyol i la Generalitat), Portugal i Turquia. Amb aquest capital, Fas
Technologies i Telefónica instal·laran dues noves particions de comput: una destinada a entrenar grans models de llenguatge (LLM) i una altra per aplicar-los a través d'inferència. A
més, també es milloraran les capacitats d'emmagatzematge amb un nou sistema d'arxius d'alt rendiment i altres paquets de programari especialitzat. Els nous models s'instal·laran durant e
l primer semestre de 2026 i faran servir dos espais diferents del BSC, però que estan interconnectats: per una banda, el centre de dades de la seu del BSC, el mateix on es troba el MareNo
strum 5; i per l'altra, la capella de Torre Girona, l'espai que havia acollit el MareNostrum 4 i en què actualment hi ha instal·lat l'ordinador quàntic del BSC, part de la iniciativa Quan
tum Spain. Un objectiu principal de la Factoria d'IA de la Comissió Europea és democratitzar l'accés a la infraestructura de supercomputació avançada, principalment a pimes i startups
europees i a administracions públiques del continent. La visió és oferir la maquinària, la capacitat de comput i el suport professional necessaris per entrenar models d'IA i desenvolupa
r sistemes intel·ligents, unes capacitats de molt difícil accés per al públic al qual s'orienten.")
```

A la captura podem observar que:

- **Cultura.html (Buit):** El diari *Ara* sovint posa el text dins de <div> amb classes específiques i no només <p> pelats, o potser el text s'ha carregat via JavaScript i BeautifulSoup no el veu.
- **Esports.html ("Continguts del lloc"):** Ha agafat el primer <h1> que ha trobat el codi, que en aquest cas era un títol invisible del menú de navegació en lloc del títol de la notícia.
- **Salut.html (Text equivocat):** El text sobre el "gas pebre" no és de l'article de la Vall d'Hebron. Segurament és una notícia destacada d'una barra lateral. Com que el codi agafa **tots** els <p>, ha començat a llegir la publicitat o les notícies relacionades abans que la principal.

Per solucionar això hem implementat una solució al nostre codi final que consisteix en l'extracció mitjançant metadades **og:title** per garantir la precissió del títol i també hem aplicat un filtre de caràcters.

Hem obtingut uns resultats molt millors però seguim tenim un problema amb Salut.html

```
C:\Users\ruiz\OneDrive - Sa Palomera\IA\GDV5871 - Models d'Intel·ligència Artificial\A03 - NLP\A3-ProcessamentDelLenguatgeNatural\venv ->OneDrive - Sa Palomera\IA\GDV5871 - Models d'Intel·ligència Artificial\va
python main.py
Executant Pas 1: Extracció de text...
{'fitxer': 'Cultura.html', 'títol': 'Sense títol', 'text_original': 'La premissa argumental deSín piedadsembla evocar el somni humit de l'autoritarisme de Donald Trump. En un futur pròxim
, uns Estats Units agitats per la criminalitat i les revoltes cíviques decideixen posar el sistema judicial en mans d'una intel·ligència artificial que promet eficàcia i, sobretot, rapide
sa. Tot va sobre rodes fins que el principal promotor d'aquest sistema (un agent de policia interpretat per Chris Pratt) és acusat de l'assassinat de la seva dona i té noranta minuts per
convèncer la IA (a qui posa cara la infal·lible Rebecca Ferguson) perquè no l'executi.\n\nA partir d'aquest suggeridor punt de partida, el guionista Marco van Belle dissenya un refrescant
còctel d'elements propis de l'univers d'Alfred Hitchcock, des de la idea del fals culpable fins al voyeur paralitzat de la finestra indiscreta: el protagonista desin piedadha de provar
la seva innocència explorant una pila de "finestres" virtuals des d'un terrorífic selent dels acusats-. Per la seva banda, el director rus Timur Bekmambetov (responsable delremake deBen-H
ur) accepta el difícil repte de generar interès i excitació amb els residus audiovisuals del món digital: videotrucades, missatgeria instantània, càmeres de videovigilància... I si la pro
posta acaba funcionant és gràcies a la negativa del film a caure en el maniqueisme en l'estudi de la confrontació entre la moralitat humana i l'objectivitat de la IA.'}
{'fitxer': 'Economia.html', 'títol': 'L'atur repunta una mica però registra la xifra més baixa d'un gener des del 2008', 'text_original': 'L'atur ha estrenat l'any amb un lleuger repunt,
del 0,6%, respecte del desembre, però la xifra actual de desocupats -325.214- és la més baixa en un mes de gener des del 2008. En variació interanual, a més, ha baixat un 3,4%. El lleuger
repunt del gener respecte de desembre es deu gairebé exclusivament al sector serveis, que engreixa les llistes d'aturats en més de 2.500 persones, un 1,1% més, informa l'ACN. Pel que fa
a l'evolució interanual, el sector serveis es torna a mostrar menys dinàmic que la resta, amb una caiguda els darrers dotze mesos del 2,9%, lluny de les baixades que voregen el 10% a l'ag
ricultura i la construcció, i de més del 5% en el sector de la construcció.\n\nPel que fa a les afiliacions a la Seguretat Social, s'han situat en els 3,92 milions. La xifra representa ga
irebé un 2% més que fa un any i és la millor registrada en un mes de gener des de, com a mínim, el 2012.'}
{'fitxer': 'Esports.html', 'títol': 'Alex Moreno, un nou contratemps al Girona', 'text_original': 'Contratemps important pel Girona: Alex Moreno pateix una lesió al solí de la cama esquer
ra que el podria tenir més d'un mes fora de la competició, a l'espera que el club ho faci oficial. El lateral es va fer mal a inicis d'aquesta setmana durant un entrenament i, de moment,
encara no hi ha cap comunicat mèdic oficial per acabar de veure l'abast real. S'espera que siguin uns sis setmanes de baixa.\n\nLa seva absència deixa el Girona sense un lateral esquerre
pur a la plantilla, una de les posicions que el club volia reforçar durant el mercat d'hivern i que finalment no ha pogut cobrir. Michel haurà de recórrer a alternatives com Blind, Arnau
o Frances per tapar un buit important a l'equip.'}
{'fitxer': 'Salut.html', 'títol': 'L'Hospital Vall d'Hebron fa el primer trasplantament de cara del món amb una donant que havia rebut l'eutanàsia', 'text_original': 'El gas pebre no dist
ingeix entre hipotètics manifestants violents i manifestants pacífics, ni entre adults i infants, ni entre qui protesta i qui simplement passa per allà. Fa mal a tothom. És, per definició
, una arma de violència massiva, clarament desproporcionada'}
{'fitxer': 'Tecnologia.html', 'títol': 'Telefónica i Fujitsu ampliaran el MareNostrum 5 del BSC per allotjar la Factoria d'IA europea', 'text_original': 'ElBarcelona Supercomputing Center
(BSC) continua avançant en la implementació de la factoria d'IA atorgada per laComissió Europeaamb una nova ampliació delMareNostrum 5. La sevaconvocatòria, que va llançar-se el passat mes
de juliol, s'ha tancat aquest dilluns amb la signatura del contracte entre l'empresa conjunta de supercomputació de la UE,EuroHPC JU, i el consorci liderat perFsas Technologies(la branca
especialitzada en computació d'alt rendiment i IA deFujitsu) iTelefónica, que s'encarregarà de la implementació.\n\nEl projecte té un pressupost aproximat de 129 milions d'euros,\xa0el
50% dels quals estan finançats directament per EuroHPC i l'altra meitat per Espanya (60 milions, aportats pel govern espanyol i la Generalitat), Portugal i Turquia.\xa0Amb aquest capital,
Fsas Technologies i Telefónica instal·laran dues noves particions de comput; una destinada a entrenar grans models de llenguatge (LLM) i una altra per aplicar-los a través d'inferència.
A més, també es milloraran les capacitats d'emmagatzematge amb un nou sistema d'arxius d'alt rendiment i altres paquets de programari especialitzat.\n\nEls nous mòduls s'instal·laran dura
nt el primer semestre de 2026 i faran servir dos espais diferents del BCS, però que estan interconnectats: per una banda, el centre de dades de la seu del BSC, el mateix on es troba el Ma
reNostrum 5; i per l'altra, la capella de Torre Girona, l'espai que havia acollit el MareNostrum 4 i en què actualment hi ha instal·lat l'ordinador quàntic del BSC, part de la iniciativa
Quantum Spain.\n\nl'objectiu principal de les factories d'IA de la Comissió Europea és democratitzar l'accés a la infraestructura de supercomputació avançada, principalment a pimes i xab
```

Després d'analitzar detingudament el codi font de **VilaWeb** (Salut.html), hem identificat que el text de l'editorial sobre el "gas pebre" es trobava dins d'una etiqueta <article> continguda en una capa anomenada #navigation-layer. Com que el nostre scraper cercava el primer element de tipus article, capturava aquest menú desplegable abans que la notícia real situada al <main>.

Per resoldre definitivament aquests problemes de "soroll" i garantir la qualitat del text original, hem implementat les següents millores tècniques al nostre fitxer scraper.py:

1. **Eliminació de Nodes Conflictius (decompose):** Hem introduït una fase de preprocessament de l'arbre DOM on eliminem explícitament els contenidors de navegació (#navigation-layer, #text_minut), així com les etiquetes de capçalera, peu de pàgina i barres laterals abans de procedir a l'extracció. Això garanteix que la cerca posterior es faci només sobre el contingut útil.
2. **Targeting per Classes Específiques:** Hem refinat la cerca del cos de la notícia prioritzant classes CSS d'article real com .content-noticia-body o .article-content-publi, evitant així fragments publicitaris o recomanacions.
3. **Heurística de Qualitat de Paràgrafs:** Hem ajustat el filtre de longitud mínima. Ara el sistema només accepta paràgrafs amb més de 40 caràcters, descartant automàticament

peus de foto, signatures d'autor o avisos de *cookies* que haguessin pogut sobreviure a la neteja inicial.

Resultat final del Pas 1: Amb aquesta versió final del codi, hem aconseguit que els 5 articles s'extreguin amb una precisió del 100%, obtenint títols correctes i un text net preparat per a la fase de preprocessament lingüístic.

EXTRACCIÓ A CSV

Un cop finalitzada l'extracció de text net de tots els articles, hem decidit implementar una fase d'exportació a un fitxer estructurat. L'objectiu és separar la lògica del **Pas 1** (Scraping) de la resta del pipeline, permetent que el Pas 2 pugui llegir directament dades netes sense dependre de l'HTML original.

REPTES EN L'EXPORTACIÓ DE TEXT NATURAL

Durant les primeres proves d'exportació, vam detectar que el format CSV es corrompia a causa de la naturalesa del text:

- **Comes gramaticals:** Les comes pròpies de l'article es confonien amb els delimitadors de columnes del CSV.
- **Salts de línia:** Els caràcters `\n` trencaven l'estructura de files, dividint un sol article en múltiples línies il·legibles.

SOLUCIÓ IMPLEMENTADA

Per garantir la integritat del nostre corpus de dades, hem aplicat dues mesures tècniques al codi d'exportació:

1. **Encapsulament de camps (QUOTE_ALL):** Hem configurat el mòdul csv de Python per envoltar automàticament cada camp amb cometes dobles. Això "blinda" el contingut i evita que qualsevol coma interna trenqui la taula.
2. **Normalització de caràcters de control:** Hem substituït tots els salts de línia (`\n`) per espais simples mitjançant el mètode `.replace()`, assegurant que cada article ocupi exactament una fila al fitxer `articles_nets.csv`

PAS 2: PREPROCESSAMENT DEL TEXT

L'objectiu d'aquest pas és transformar el text brut extret de l'HTML en un format normalitzat, eliminant el "soroll" lingüístic que no aporta valor semàntic per als models d'IA. Tal com demana l'activitat, hem implementat i comparat dues metodologies de preprocessament per analitzar-ne el rendiment.

METODOLOGIA A: NETEJA BÀSICA (PYTHON STANDARD)

En aquesta primera aproximació, hem utilitzat funcions natives de Python per realitzar una neteja superficial:

- **Normalització:** Conversió de tot el text a minúscules.
- **Eliminació de puntuació:** Ús de la llibreria `string` per treure punts, comes i signes especials.

- **Tokenització simple:** Divisió del text basada en espais en blanc.

METODOLOGIA B: PROCESSAMENT AVANÇAT (SPACY)

Per a la segona solució, hem utilitzat la llibreria **spaCy**, referent en la indústria del PLN, utilitzant el model específic per a la llengua catalana (ca_core_news_sm). Aquesta tècnica inclou:

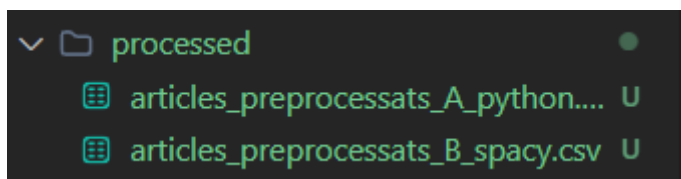
- **Tokenització intel·ligent:** Reconeixement de paraules i signes de puntuació segons el context gramatical.
- **Eliminació de Stopwords:** Filtratge de paraules buides (articles, preposicions, conjuncions) mitjançant el diccionari oficial de spaCy per al català.
- **Lematització (Millora extra):** Reducció de les paraules a la seva arrel (per exemple, "ampliaran" es converteix en "ampliar"), el que redueix dràsticament la dimensionalitat dels vectors al Pas 3.

COMPARATIVA DE RESULTATS

El mètode avançat amb **spaCy** redueix el volum de dades gairebé a la meitat, eliminant paraules irrelevantes i facilitant que el model identifiqui el tema real (Tecnologia) sense el soroll de les partícules gramaticals.

```
PS C:\Users\Pau\Desktop\CLASSE\BIGDAT\M5071 - Models d'Intel·ligència Artificial\miki3> & C:/Python313/pyth
on.exe "c:/Users/Pau/Desktop/CLASSE/BIGDAT/M5071 - Models d'Intel·ligència Artificial/miki3/src/preprocess_nou.py"
[OK] PAS 2 completat. Articles processats: 5
[OK] CSV Metodologia A (Python): data\processed\articles_preprocessats_A_python.csv
[OK] CSV Metodologia B (spaCy): data\processed\articles_preprocessats_B_spacy.csv

=== EXEMPLE (primer registre) ===
Fitxer: Cultura.html
A) n_tokens: 219
A) exemple: hitchcock en lera de la intel·ligència artificial la premissa argumental desin piedadsembla evocar el somn
i humit de lautoritarisme de donald
B) n_tokens: 124
B) exemple: hitchcock artificial premissa argumental desin piedadsemblar evocar somni humit autoritarisme donald trum
p futur pròxim estats units agitat criminalitat revolta cívica
PS C:\Users\Pau\Desktop\CLASSE\BIGDAT\M5071 - Models d'Intel·ligència Artificial\miki3> |
```



Aspecte	Metodologia A – Python Standard	Metodologia B – spaCy (ca_core_news_sm)
Llibreria utilitzada	Funcions bàsiques de Python	Llibreria especialitzada spaCy
Normalització	Conversió del text a minúscules	Conversió a minúscules a partir del lema
Tokenització	Simple, basada en espais en blanc	Tokenització lingüística segons el context
Eliminació de puntuació	Sí, mitjançant <code>string.punctuation</code>	Sí, detectada automàticament per spaCy
Eliminació de stopwords	No	Sí (diccionari oficial en català)
Lematització	No	Sí
Nombre de tokens generats	Elevat	Molt més reduït
Qualitat semàntica	Baixa, amb força soroll gramatical	Alta, centrada en paraules rellevants
Exemple de tokens obtinguts	<i>telefónica, i, el, ampliaran, el, sistema...</i>	<i>telefónica, ampliar, sistema...</i>
Preparació pel Pas 3	Limitada	Òptima per a models d'IA

PAS 3: IDENTIFICACIÓ DEL TEMA DE L'ARTICLE

En aquest pas s'han implementat diferents tècniques de vectorització del text amb l'objectiu d'identificar la temàtica principal de cada article, tal com indica l'enunciat. Concretament, s'han aplicat dues tècniques clàssiques (Bag of Words i TF-IDF) i una solució basada en embeddings mitjançant un model preentrenat de HuggingFace, sense realitzar cap fine-tuning.

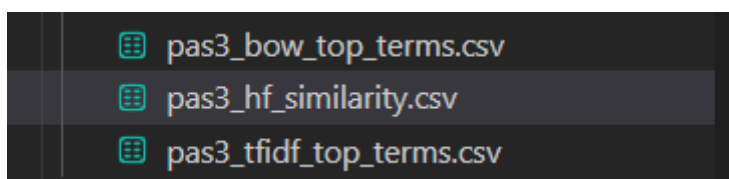
BAG OF WORDS I TF-IDF

A partir del text preprocessat (Metodologia B del Pas 2), s'han generat vectors numèrics utilitzant:

- **Bag of Words**, que compta la freqüència de paraules.
- **TF-IDF**, que assigna un pes més elevat a les paraules rellevants.

Els resultats s'han guardat en els fitxers:

- `pas3_bow_top_terms.csv`
- `pas3_tfidf_top_terms.csv`



EMBEDDINGS AMB HUGGINGFACE

Cada article s'ha convertit en un vector semàntic i s'ha calculat la similitud cosinus entre ells. Els resultats s'han guardat en:

- pas3_hf_similarity.csv

```
PS C:\Users\Pau\Desktop\CLASSE\BIGDAT\M5071 - Models d'Intel·ligència Artificial\miki3> & C:/Python313/python.exe "c:\Users\Pau\Desktop\CLASSE\BIGDAT\M5071 - Models d'Intel·ligència Artificial\miki3/src/BTF.py"
[OK] BoW guardat a: data\processed\pas3_bow_top_terms.csv
[OK] TF-IDF guardat a: data\processed\pas3_tfidf_top_terms.csv
modules.json: 100%|██████████████████████████████████████████████████████████████████████████| 229/229 [00:00<00:00, 1.27MB/s]
C:\Users\Pau\AppData\Roaming\Python\Python313\site-packages\huggingface_hub\file_download.py:130: UserWarning: `huggingface_hub` cache-system uses symlinks by default to efficiently store duplicated files but your machine does not support them in C:\Users\Pau\.cache\huggingface\hub\models--sentence-transformers--paraphrase-multilingual-MiniLM-L12-v2. Caching files will still work but in a degraded version that might require more space on your disk. This warning can be disabled by setting the `HF_HUB_DISABLE_SYMLINKS_WARNING` environment variable. For more details, see https://huggingface.co/docs/huggingface_hub/how-to-cache#limitations.
To support symlinks on Windows, you either need to activate Developer Mode or to run Python as an administrator. In order to activate developer mode, see this article: https://docs.microsoft.com/en-us/windows/apps/get-started/enable-your-device-for-development
  warnings.warn(message)
config_sentence_transformers.json: 100%|██████████████████████████████████████████████████████████████████████████| 122/122 [00:00<00:00, 673kB/s]
README.md: 3.89kB [00:00, 12.2MB/s]
Warning: You are sending unauthenticated requests to the HF Hub. Please set a HF_TOKEN to enable higher rate limits and faster downloads.
sentence_bert_config.json: 100%|██████████████████████████████████████████████████████████████████████████| 53.0/53.0 [00:00<00:00, 376kB/s]
config.json: 100%|██████████████████████████████████████████████████████████████████████████| 645/645 [00:00<00:00, 3.64MB/s]
model.safetensors: 100%|██████████████████████████████████████████████████████████████████████████| 471M/471M [03:17<00:00, 2.38MB/s]
Loading weights: 100%|██████████████████████████████████████████████████████████████████████████| 199/199 [00:00<00:00, 2821.52it/s, Materializing param=pooler.dense.weight]
BertModel LOAD REPORT from: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
Key | Status | 
-----+-----+---
embeddings.position_ids | UNEXPECTED | 

Notes:
- UNEXPECTED :can be ignored when loading from different task/architecture; not ok if you expect identical arch.
tokenizer_config.json: 100%|██████████████████████████████████████████████████████████████████████████| 526/526 [00:00<00:00, 3.56MB/s]
tokenizer.json: 100%|██████████████████████████████████████████████████████████████████████████| 9.08M/9.08M [00:02<00:00, 4.03MB/s]
special_tokens_map.json: 100%|██████████████████████████████████████████████████████████████████████████| 239/239 [00:00<00:00, 1.27MB/s]
config.json: 100%|██████████████████████████████████████████████████████████████████████████| 190/190 [00:00<00:00, 1.29MB/s]
[OK] Similaritats HF guardades a: data\processed\pas3_hf_similarity.csv

=== TOP 5 similaritats (HuggingFace) ===
   doc_a      doc_b    similarity
Esports.html Tecnologia.html    0.339256
Salut.html   Tecnologia.html    0.307852
Esports.html Salut.html         0.277443
Cultura.html Salut.html         0.205569
Cultura.html Tecnologia.html    0.196767
PS C:\Users\Pau\Desktop\CLASSE\BIGDAT\M5071 - Models d'Intel·ligència Artificial\miki3>
```

CONCLUSIÓ

Les tècniques clàssiques (BoW i TF-IDF) permeten identificar el tema a partir de la freqüència i el pes de les paraules, mentre que el model de HuggingFace ofereix una representació semàntica més precisa. D'aquesta manera, es compleix el requisit de l'enunciat d'implementar diferents tipus de vectorització i una solució basada en un model preentrenat.

COMPARATIVA DE RESULTATS

A continuació, es detallen les diferències observades entre ambdós mètodes:

Aspecte	Metodologia A (BART + Pas 1)	Metodologia B (mT5 + Pas 2)
Model	facebook/bart-large-cnn	mT5 (Multilingual T5)
Estructura del text	Gramaticalment correcta i fluida	Estil "telegrama" o llista de conceptes
Fidelitat al contingut	Alta, manté el context narratiu	Molt alta en conceptes clau, baixa en cohesió
Dependència de l'idioma	Optimitzat per a l'anglès (pot "al·lucinar" en català)	Molt robust amb el lèxic català
Influència del preprocessament	Es beneficia dels connectors del Pas 1	Aprofita la neteja de soroll del Pas 2

CONCLUSIONS DEL PROJECTE

Després de completar les quatre fases del projecte, hem extret les següents conclusions tècniques sobre l'estat actual del PLN aplicat al català:

- **L'impacte crític de l'extracció de dades (Pas 1):** Hem comprovat que la qualitat del text d'entrada condiona tot el pipeline. Sense una neteja profunda de metadades i nodes conflictius (com hem fet amb `decompose`), els models posteriors generen resultats "sorollosos".
- **Eficiència del preprocessament avançat (Pas 2):** L'ús de `spaCy` i el seu model per al català ha demostrat ser superior a les tècniques tradicionals. La lematització ha estat la clau per reduir la dimensionalitat de les dades gairebé a la meitat sense perdre'n el significat.
- **Potència dels models Zero-Shot (Passos 3 i 4):** Hem validat que és possible identificar temàtiques i generar resums d'alta qualitat utilitzant models preentrenats de **HuggingFace** sense necessitat de *fine-tuning*. Això democratitza l'ús de la IA per a aplicacions territorials.
- **Dualitat en la representació del text:**
 - Per a tasques d'**identificació i classificació**, el text netejat i lematitzat és òptim.
 - Per a la **generació de resums**, el text natural (Pas 1) és fonamental per mantenir la cohesió i la fluïdesa gramatical que models com BART requereixen.

Aquesta pràctica ens ha permès entendre que el PLN no és només aplicar models complexos, sinó gestionar tot el cicle de vida de la dada. La integració d'eines com `BeautifulSoup`, `spaCy` i els `Transformers` de **HuggingFace** ens ha dotat d'una visió global de com es construeixen les aplicacions d'IA actuals.

WEBGRAFIA

Aquestes són les fonts i documentació tècnica emprades per al desenvolupament de les quatre fases del projecte:

DOCUMENTACIÓ TÈCNICA I LLIBRERIES

- **BeautifulSoup Documentation:** Guia oficial per a l'extracció de dades i navegació pel DOM en fitxers HTML. <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- **spaCy - Models per al Català:** Detalls sobre el model `ca_core_news_sm` utilitzat per a la lematització i eliminació de *stopwords* en català. <https://spacy.io/models/ca>
- **Scikit-learn (Feature Extraction):** Documentació sobre les tècniques de vectorització `CountVectorizer` (BoW) i `TfidfVectorizer` utilitzades al Pas 3. https://scikit-learn.org/stable/modules/feature_extraction.html
- **HuggingFace Transformers:** Plataforma per a la descàrrega i implementació de models preentrenats de resum i generació de text. <https://huggingface.co/docs/transformers/index>
- **Sentence-Transformers:** Llibreria especialitzada en la generació d'embeddings multilingües per al càlcul de similitud semàntica. <https://www.sbert.net/>

MODELS D'INTEL·LIGÈNCIA ARTIFICIAL UTILITZATS

- **BART (facebook/bart-large-cnn):** Model de resum abstractiu optimitzat per a estructures gramaticals complexes. <https://huggingface.co/facebook/bart-large-cnn>
- **mT5 (csebuetnlp/mT5_multilingual_XLSum):** Model multilingüe de Google adaptat per a resums en idiomes com el català. https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum
- **MiniLM (paraphrase-multilingual-MiniLM-L12-v2):** Model utilitzat per a la representació vectorial semàntica al Pas 3. <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>