TWEETS ANALYSIS

Gianotti - Tabasso

Progetto MAADB 2020/21

CONTENUTO DELLA PRESENTAZIONE

Ecco un riassunto di cosa tratteremo oggi nella presentazione

- Ambiente di sviluppo
- Introduzione e *filosofia*
- Gestione risorse e come trattarle
- DMBS a confronto:
  - Relazionale - *MARIADB*
  - Not Only SQL - *MONGODB*
- Analisi dei risultati, tre elementi di valutazione:
  - WORD CLOUD
  - HISTOGRAMS
  - SPREADSHEET

Ambiente di Sviluppo

DOCKER

Applicazioni distribuite in container

PYTHON

`mysql-connector`

`pymongo emoji nltk`

`matplotlib wordcloud`

MONGODB

Router
Sharding
Express

MARIADB

InnoDB
utf8mb4
Adminer

“In the very beginning, people said you couldn't make relational databases fast enough to be commercially viable...”

—Larry Ellison co-fondatore e CTO della Oracle Corporation

“Users love MongoDB because it offers the fastest time to value compared to any other DBMS technology”

—Eliot Horowitz fondatore e CTO di MongoDB Inc.

01

PREPROCESS

Fase iniziale: studio del dataset e delle tecniche NLP

02

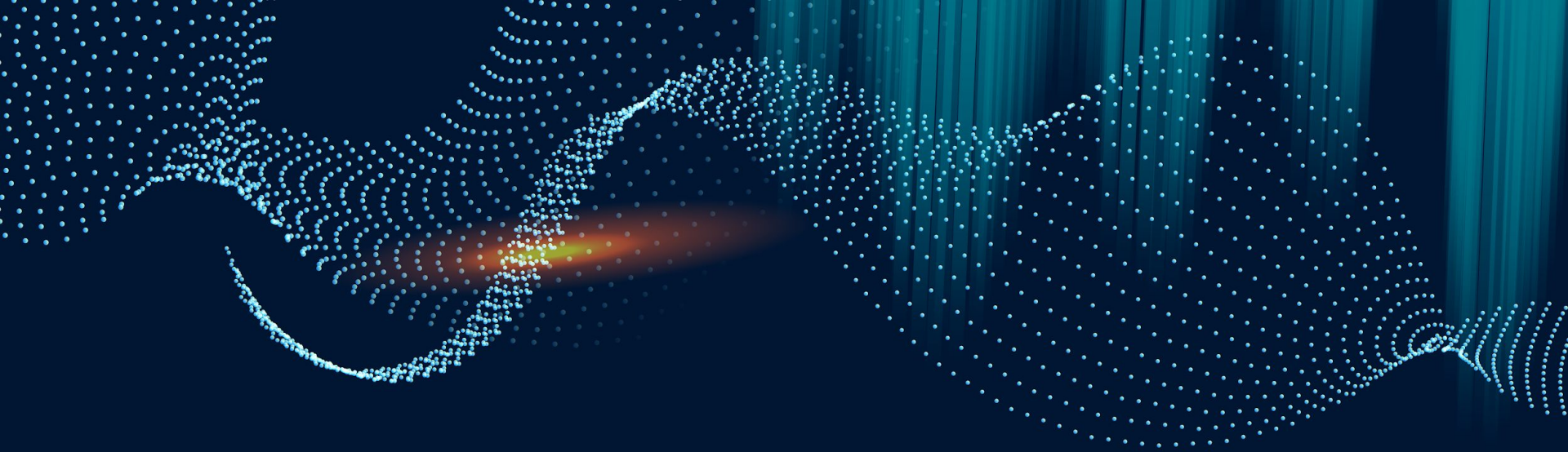
DATABASE

Progettazione e realizzazione delle due soluzioni

03

RESULT

Analisi tramite l'uso di Word cloud & istogrammi



01 Preprocess

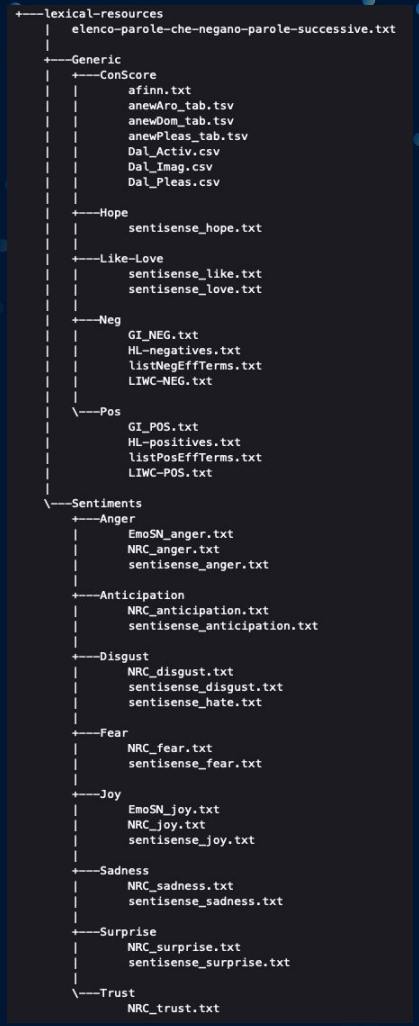
Come gestire il dataset dato?

Gestione delle risorse lessicali

Abbiamo effettuato una divisione importante tra

- **Risorse generice**
  - Caratterizzate da uno score.
  - Vengono gestite in: *manage_scoring_resources.py*

- **Sentimenti**:
  - 8 emozioni di secondo livello (Plutchik).
  - Presenza di minimo 2 risorse

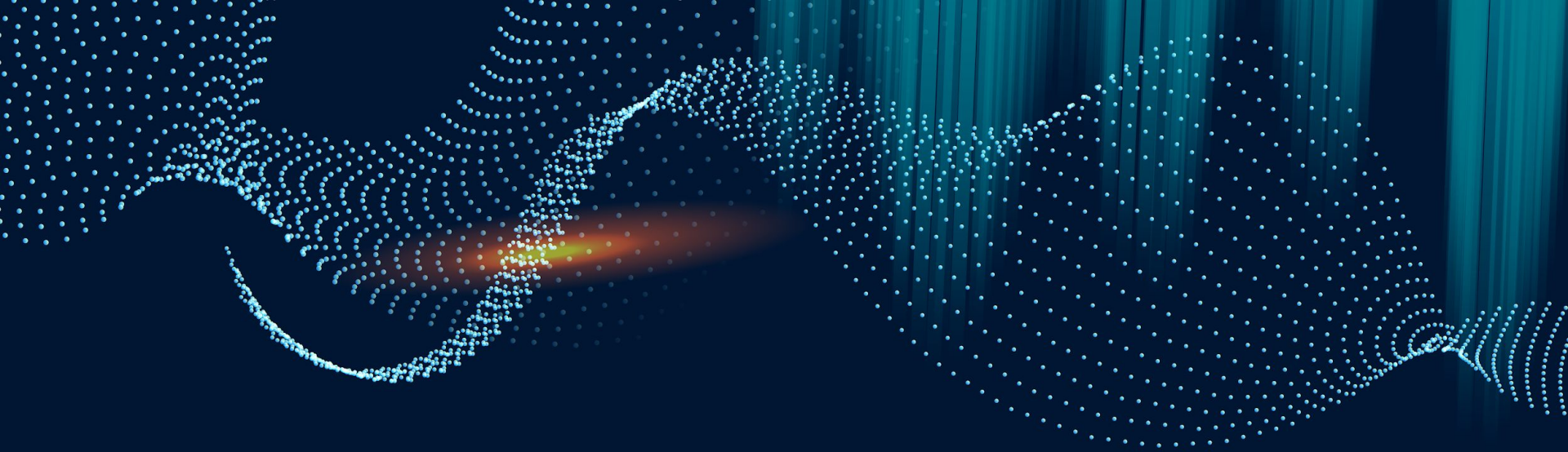


Gestione delle risorse per il preprocessamento

Per il preprocessing abbiamo:

- File diversi per diversi micro task
- Comodità del JSON mapping in python
- *noslq_preproccesing.py* e *relational_preprocessing.py*

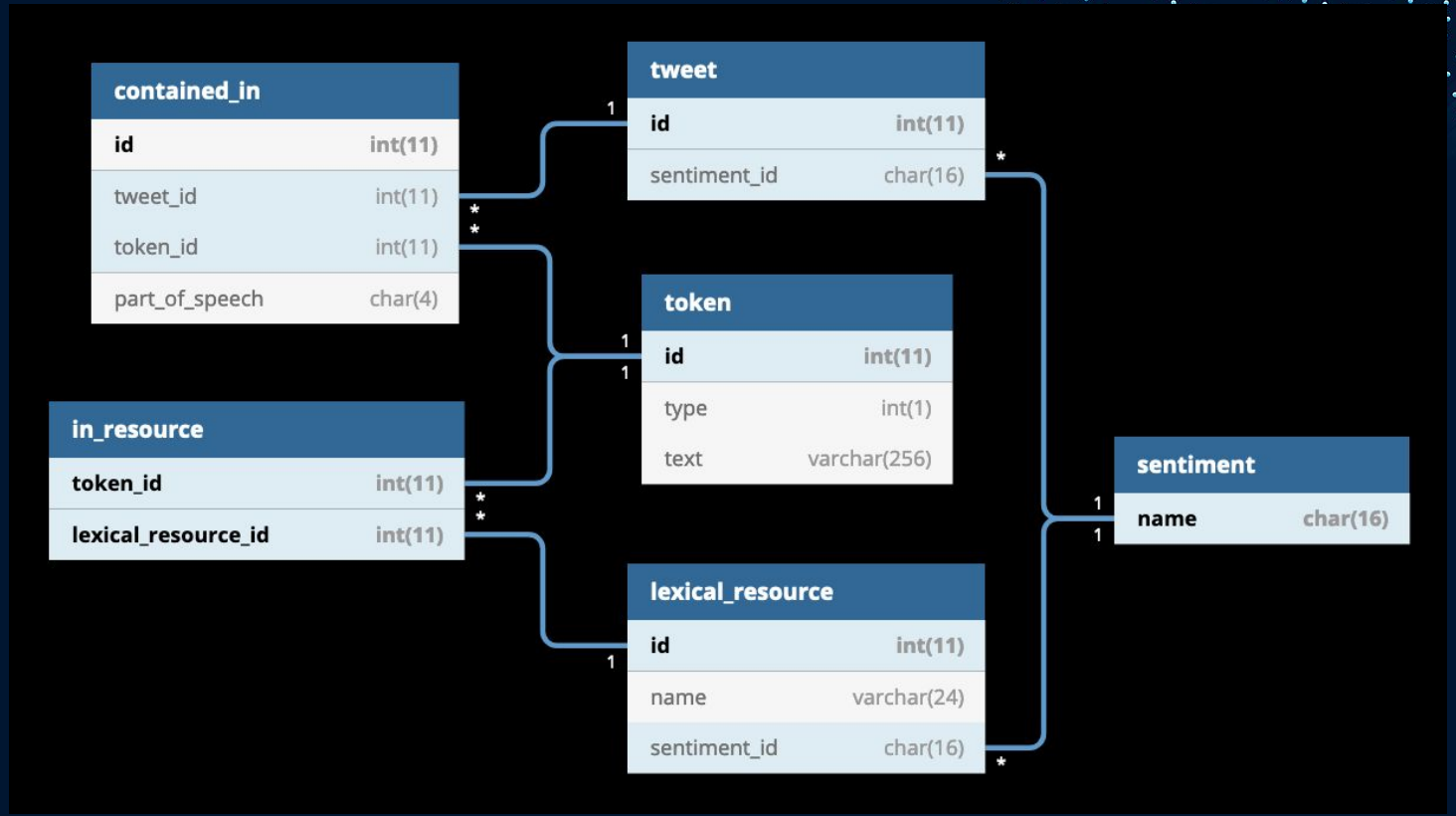
```
+---processing
|       emoji.json
|       emoticons.json
|       Penn Treebank P.O.S. Tags.html
|       Penn Treebank P.O.S. Tags.png
|       punctuation.txt
|       sentiment2emo.json
|       slang_words.json
|
\---twitter-messages
        dataset_dt_anger_60k.txt
        dataset_dt_anticipation_60k.txt
        dataset_dt_disgust_60k.txt
        dataset_dt_fear_60k.txt
        dataset_dt_joy_60k.txt
        dataset_dt_sadness_60k.txt
        dataset_dt_surprise_60k.txt
        dataset_dt_trust_60k.txt
```



02 Database

MariaDB vs MongoDB

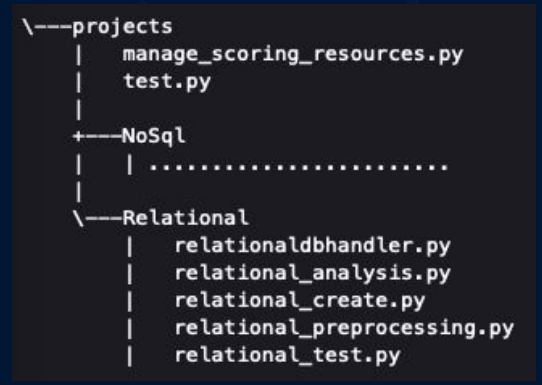
Schema ER MariaDB



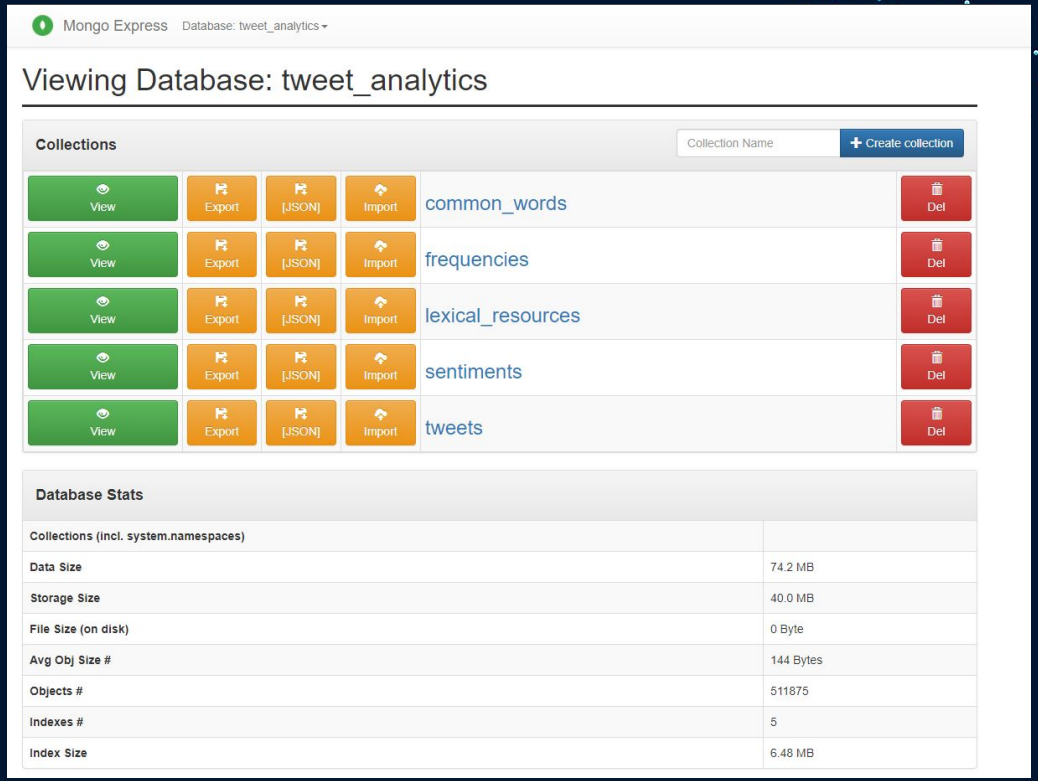
Relazionale MariaDB: struttura e compiti

Per il preprocessing abbiamo:

- Handler: interfaccia con il DBMS
- Create: inizializzazione handler
  - Caricamento risorse (lessicali e supporto)
- Test: Unit-test dell'handler
  - Classe di supporto
- Analisi: generazione statistiche e grafici



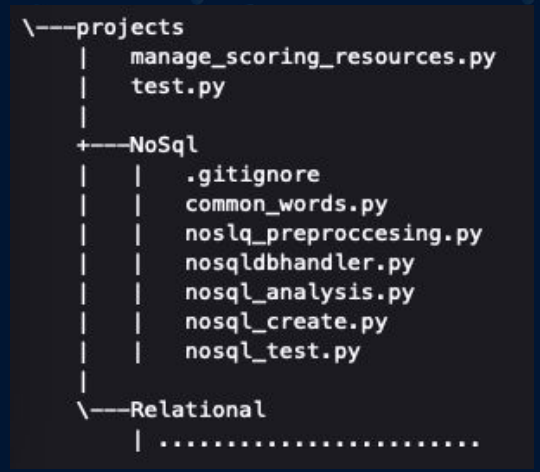
Schema collezioni MONGODB



Nosql MongoDB: struttura e compiti

Per il preprocessing abbiamo:

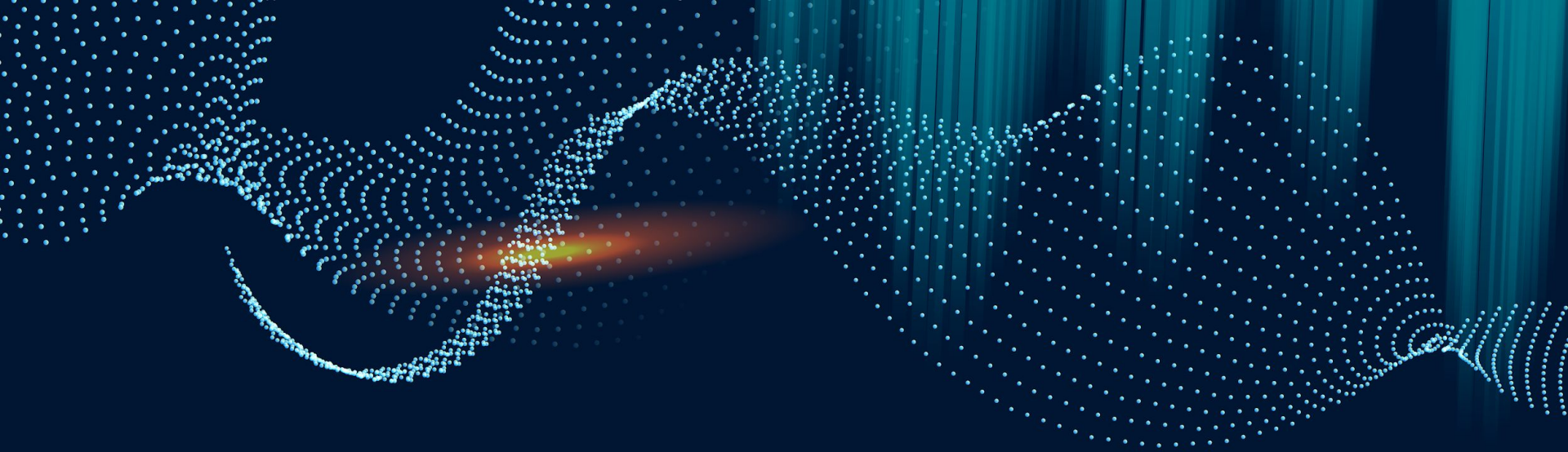
- Handler: interfaccia con il DBMS
- Create: inizializzazione handler e preprocess
  - Caricamento risorse lessicali
- Test: Unittest dell'handler
  - Classe di supporto
- Common Words: calcolo parole comuni
  - Tra le risorse lessicali e tweets
- Analisi: generazione statistiche



Elementi in comune

Considerando che le operazioni di preprocessing e di analisi sono le stesse per ambedue i database:

- Tempi di elaborazione molto diversi
- Risultati finali volto vicini
-



03 Result

Discussione e analisi

Abbiamo diverse tipologie di risultati in output

WORD CLOUD

rappresentazione visiva dei token maggiormente presenti

HISTOGRAMS

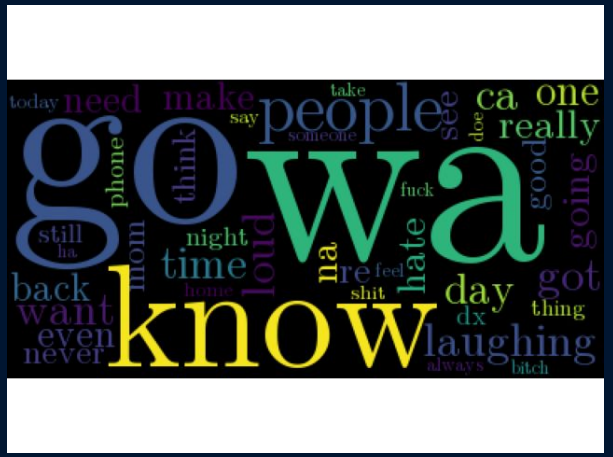
per ciascun sentimento con relative percentuali

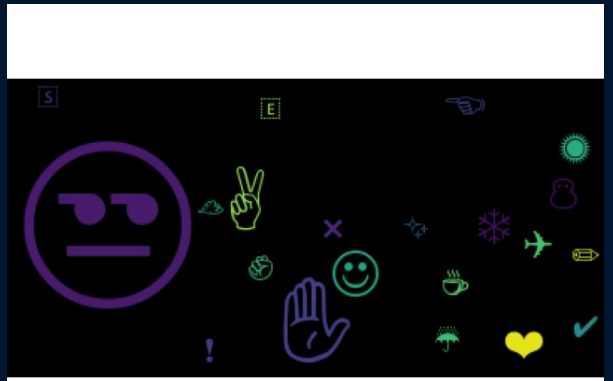
SPREADSHEET

Confronto statistico tra i due diversi approcci tramite fogli di calcolo

WORD CLOUDs

Quattro per ogni sentimento

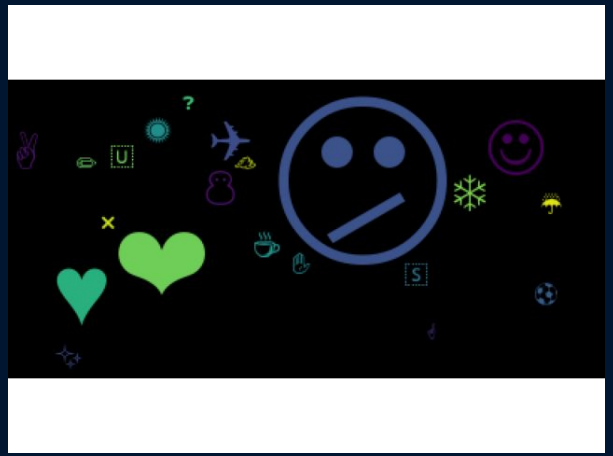




Rabbia
ANGER

Words | Emoticons

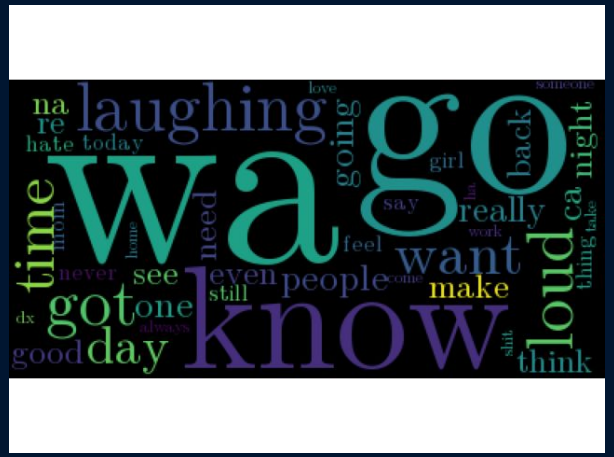
Emojis | Hashtags

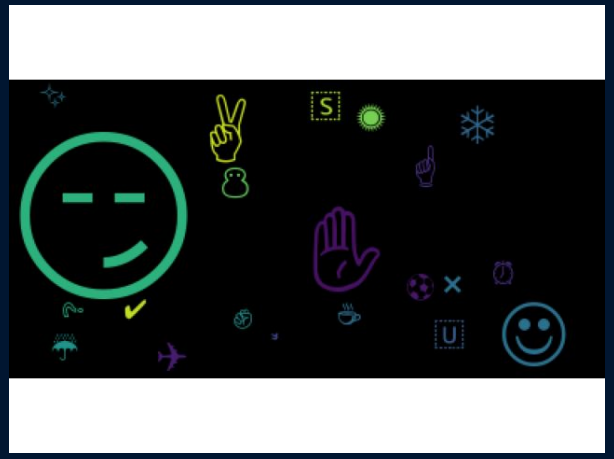


Aspettativa
ANTICIPATION

Words | Emoticons

Emojis | Hashtags

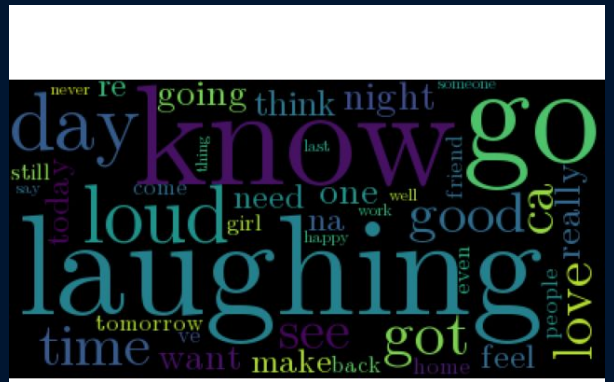


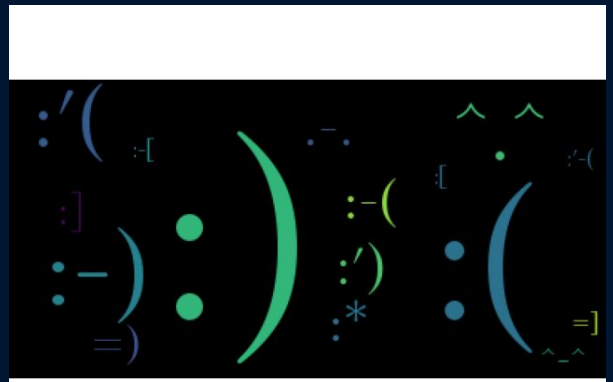


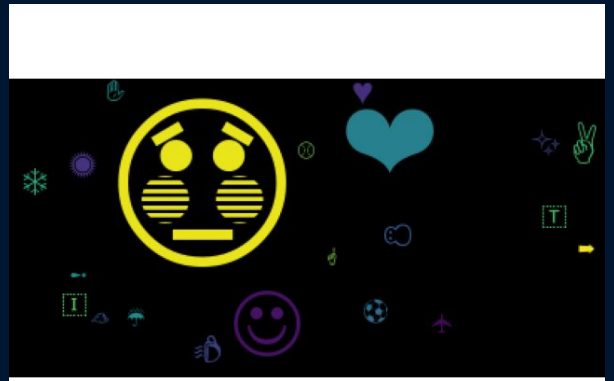
Disgusto DISGUST

Words | Emoticons

Emojis | Hashtags



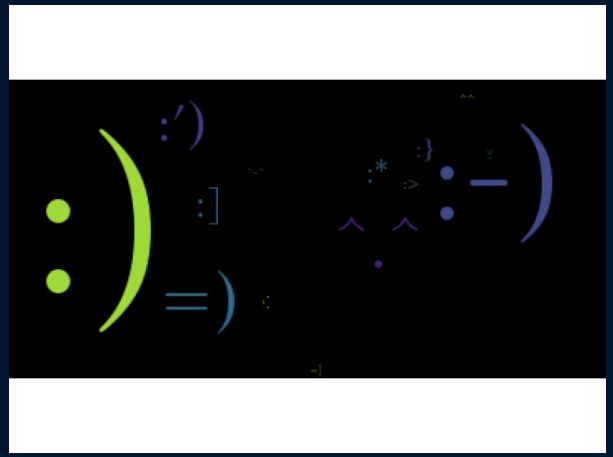


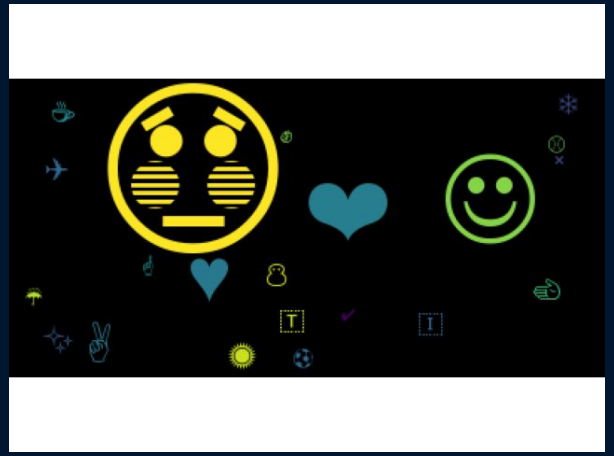


Paura
FEAR

Words | Emoticons

Emojis | Hashtags

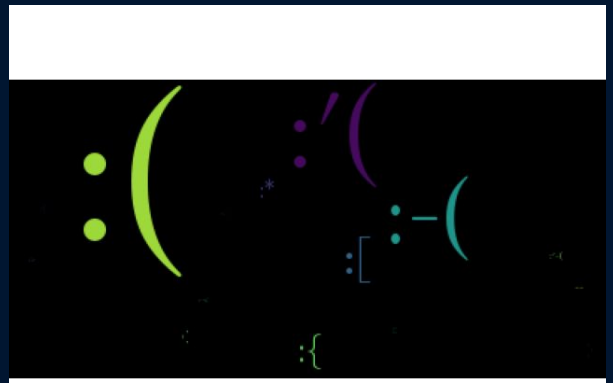


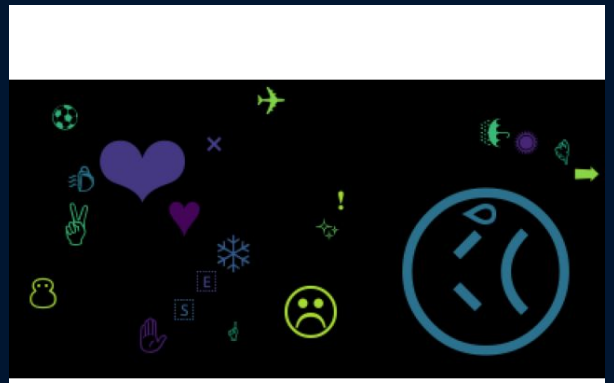


Gioia
JOY

Words | Emoticons

Emojis | Hashtags

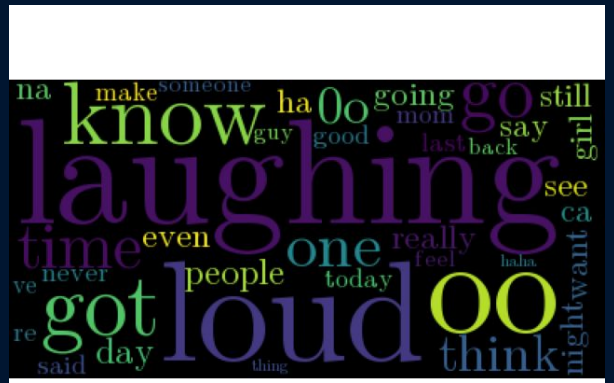


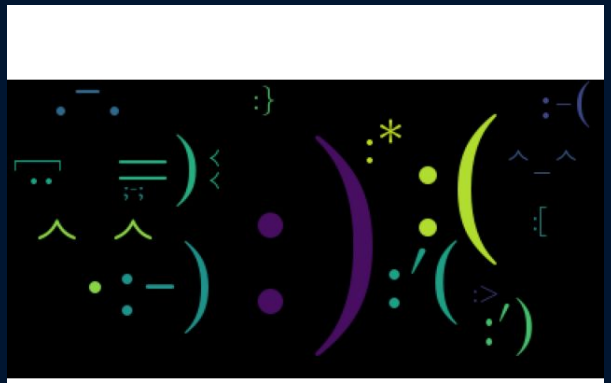


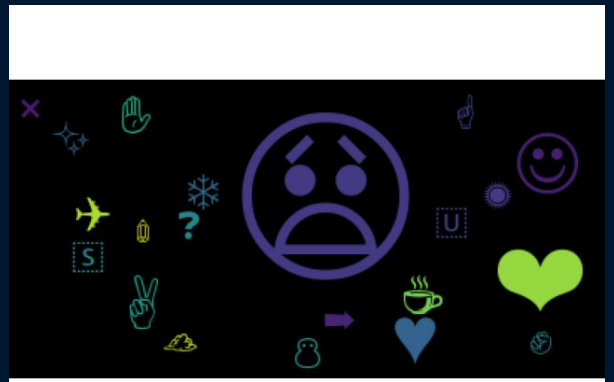
Tristezza
SADNESS

Words | Emoticons

Emojis | Hashtags



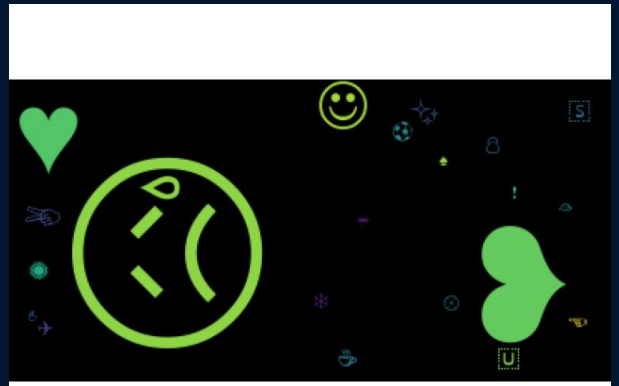


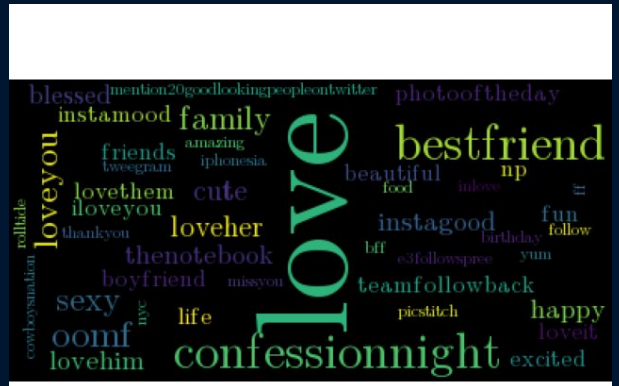


Sorpresa
SURPRISE

Words | Emoticons

Emojis | Hashtags





Fiducia
TRUST

Words | Emoticons

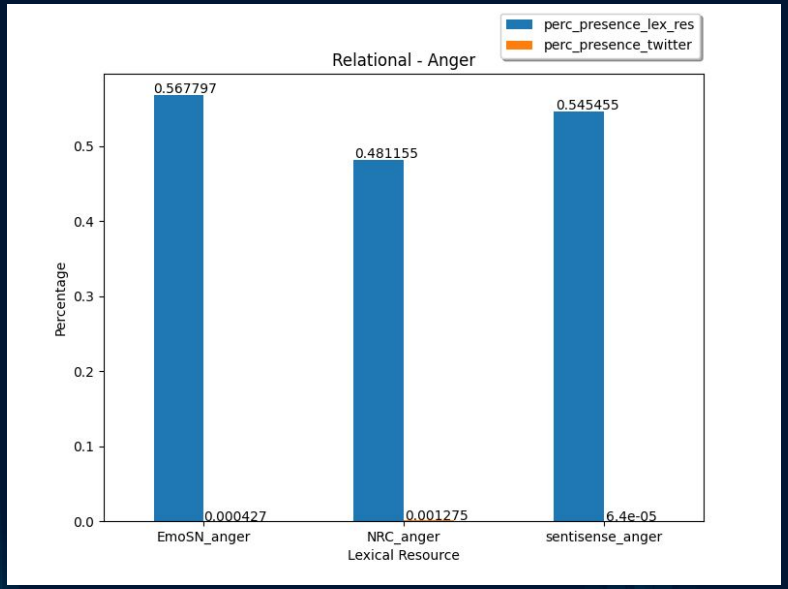
Emojis | Hashtags

HISTO GRAMMs

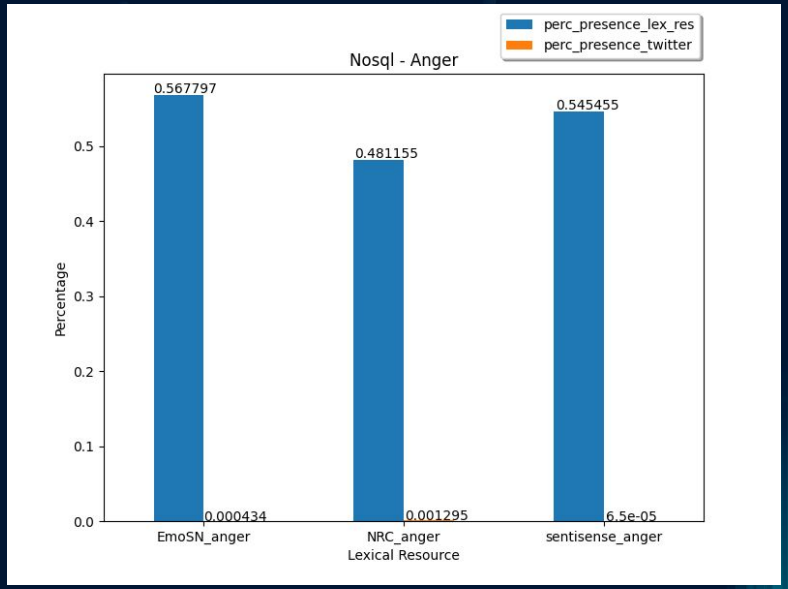
Tool per confronto tra sistemi

Anger

MARIADB

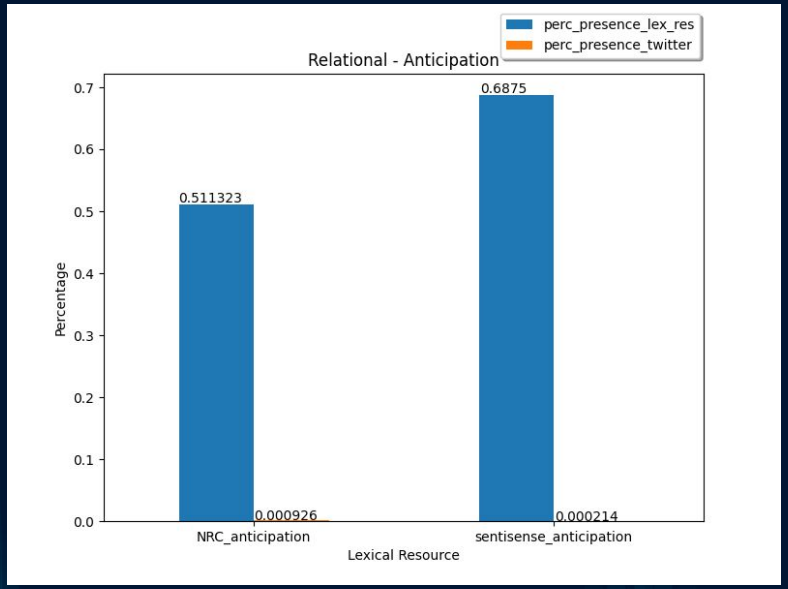


MONGODB

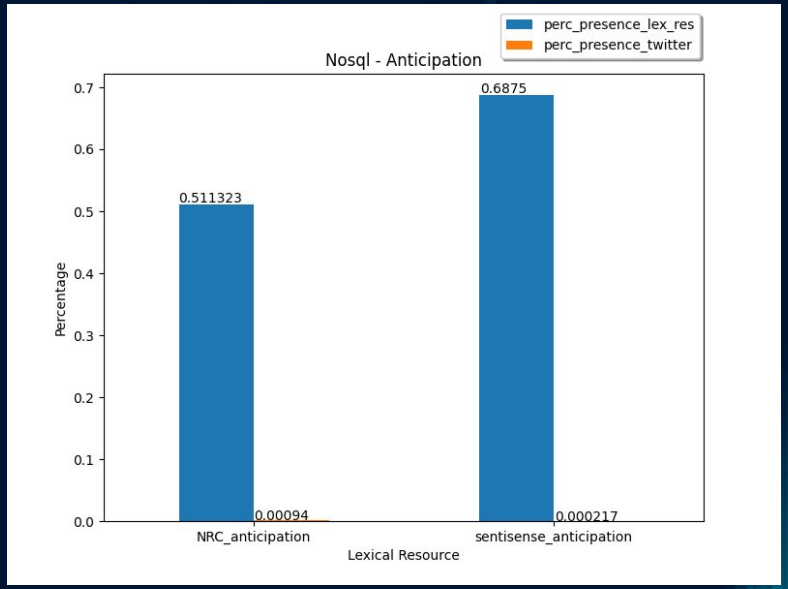


Anticipation

MARIADB

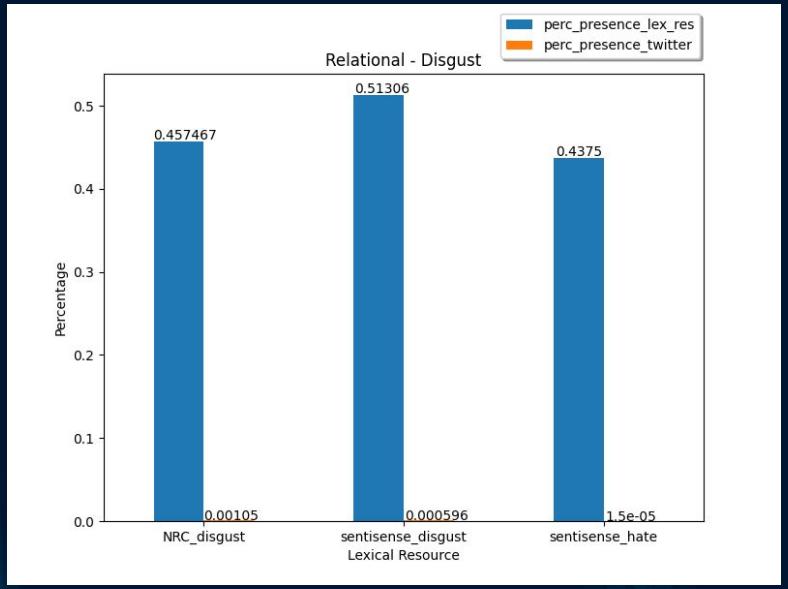


MONGODB

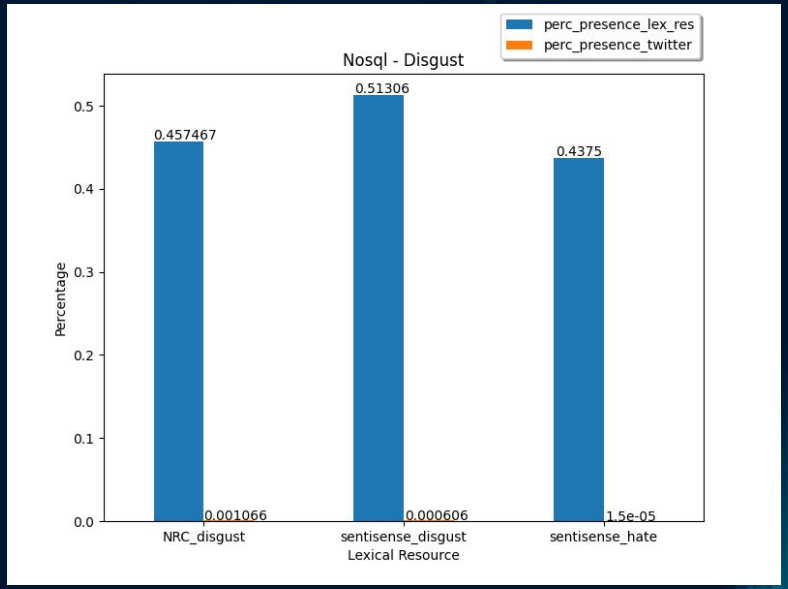


Disgust

MARIADB

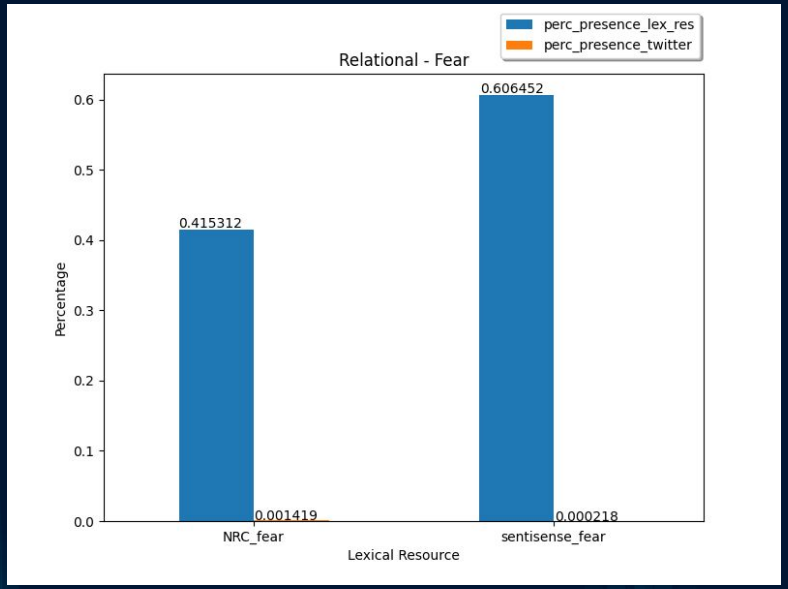


MONGODB

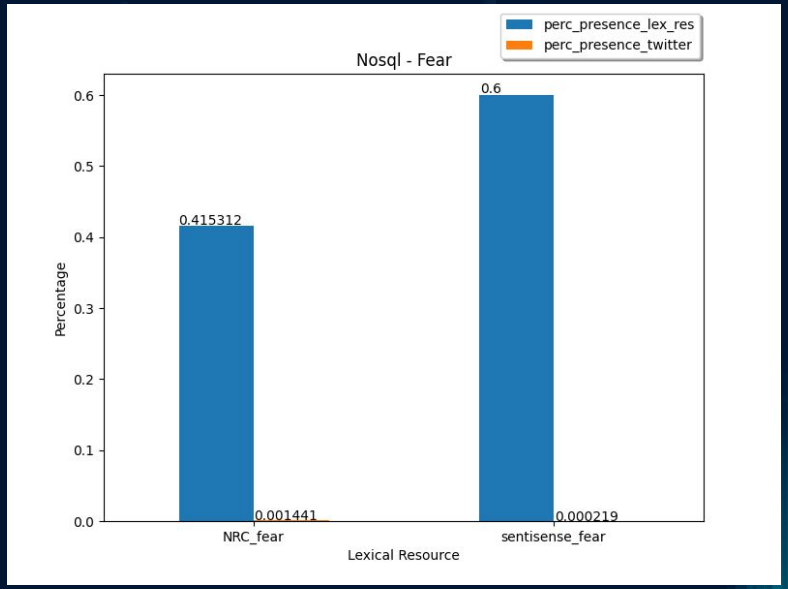


Fear

MARIADB

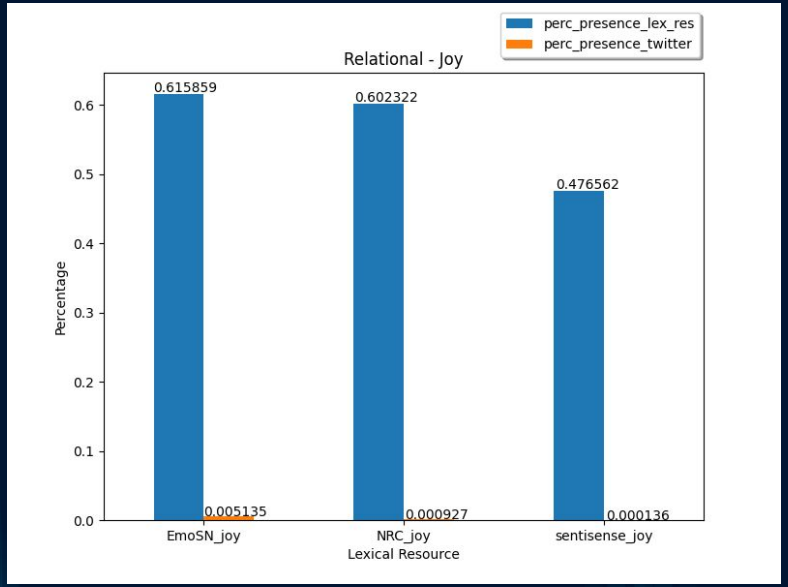


MONGODB

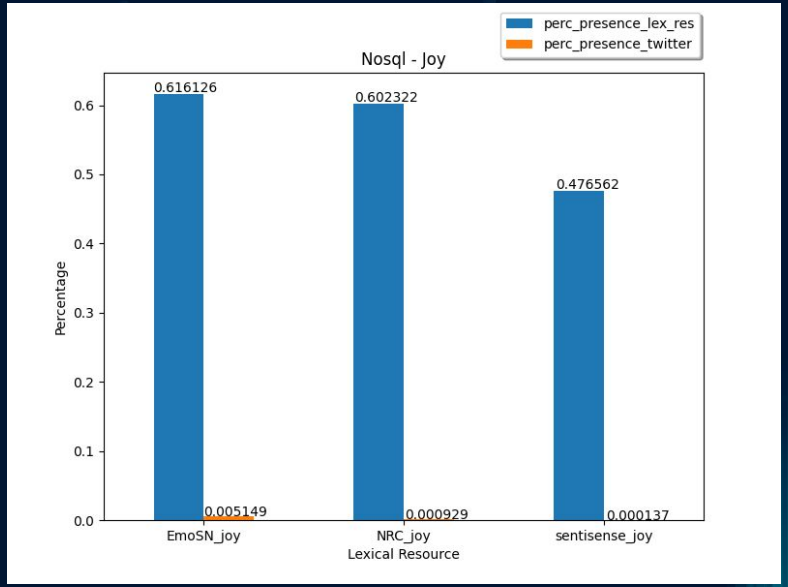


Joy

MARIADB

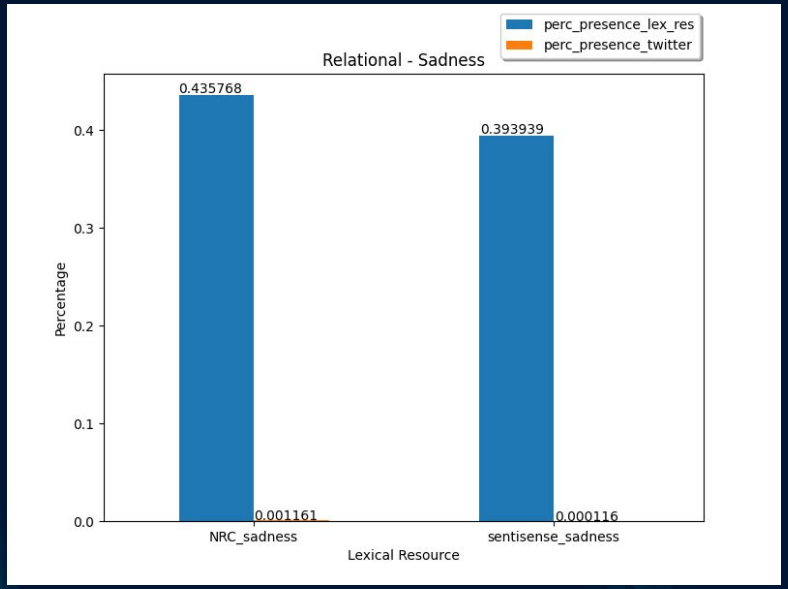


MONGODB

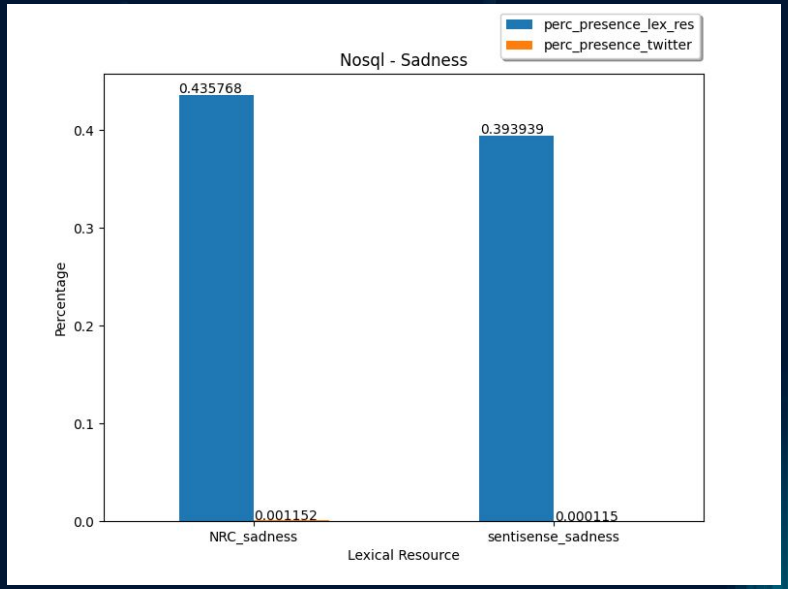


Sadness

MARIADB

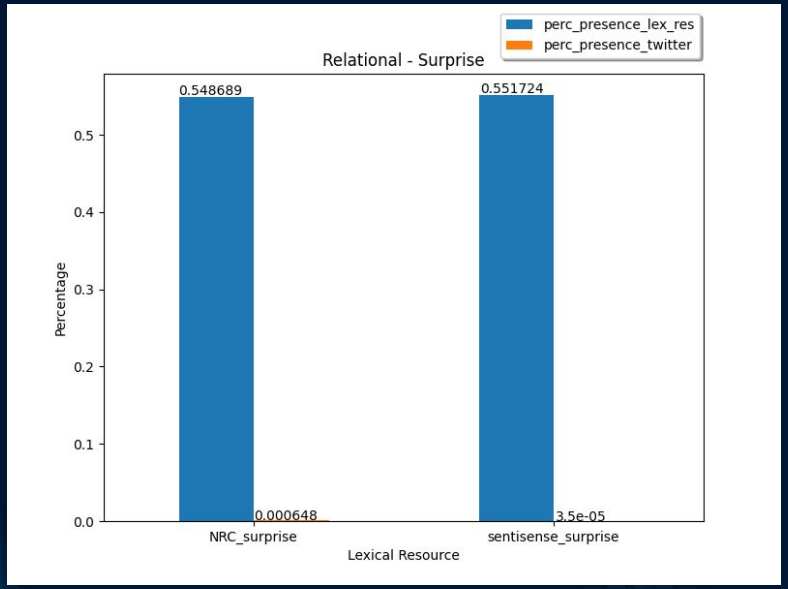


MONGODB

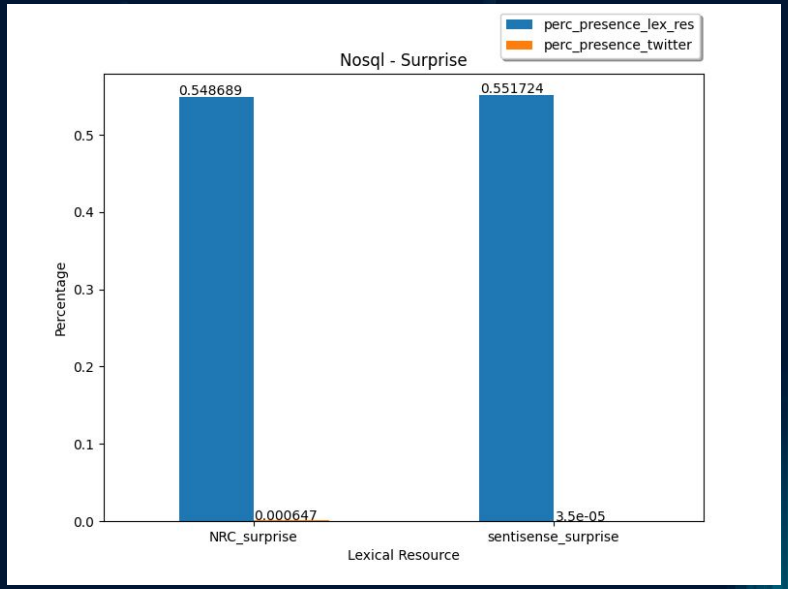


Surprise

MARIADB

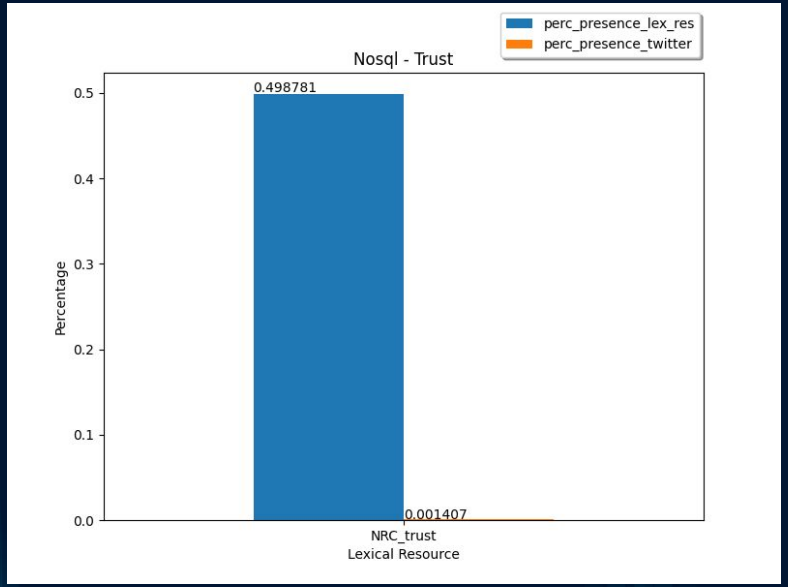


MONGODB

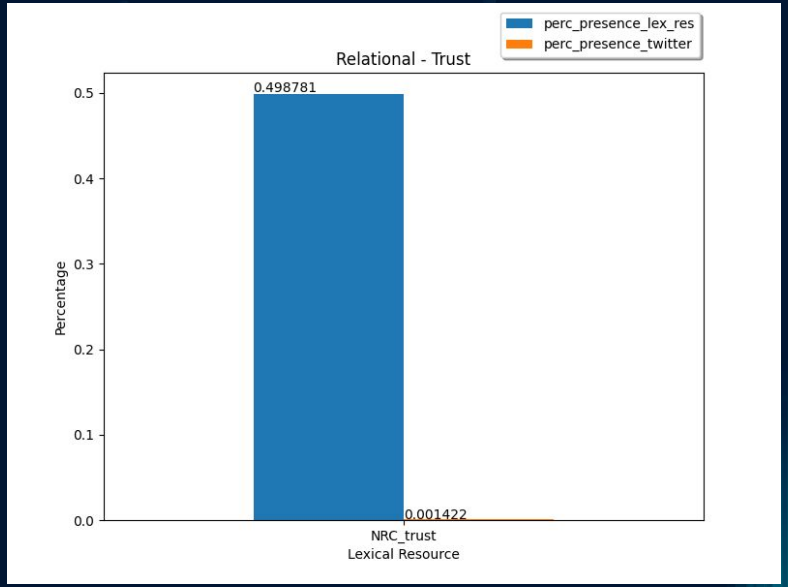


Trust

MARIADB



MONGODB



SPREAD SHEET

Statistiche & new sentiments

MARIADB foglio 1/2

sentiment	lex_resource	n_lex_words	n_twitter_words	shared_words	perc_presence_lex_res	perc_presence_twitter
anger	EmoSN_anger	354	470407	201	0,567797	0,000427
anger	NRC_anger	1247	470407	600	0,481155	0,001275
anger	sentisense_anger	55	470407	30	0,545455	0,000064
anticipation	NRC_anticipation	839	463130	429	0,511323	0,000926
anticipation	sentisense_anticipation	144	463130	99	0,6875	0,000214
disgust	NRC_disgust	1058	461080	484	0,457467	0,00105
disgust	sentisense_disgust	536	461080	275	0,51306	0,000596
disgust	sentisense_hate	16	461080	7	0,4375	0,000015

MARIADB foglio 2/2

fear	NRC_fear	1476	432025	613	0,415312	0,001419
fear	sentisense_fear	155	432025	94	0,606452	0,000218
joy	EmoSN_joy	3733	447717	2299	0,615859	0,005135
joy	NRC_joy	689	447717	415	0,602322	0,000927
joy	sentisense_joy	128	447717	61	0,476562	0,000136
sadness	NRC_sadness	1191	447110	519	0,435768	0,001161
sadness	sentisense_sadness	132	447110	52	0,393939	0,000116
surprise	NRC_surprise	534	452307	293	0,548689	0,000648
surprise	sentisense_surprise	29	452307	16	0,551724	0,000035
trust	NRC_trust	1231	431694	614	0,498781	0,001422

MONGODB foglio 1/2

sentiment	lex_resource	n_lex_words	n_twitter_words	shared_words	perc_presence_lex_res	perc_presence_twitter
anger	EmoSN_anger	354	463233	201	0,567797	0,000434
anger	NRC_anger	1247	463233	600	0,481155	0,001295
anger	sentisense_anger	55	463233	30	0,545455	0,000065
anticipation	NRC_anticipation	839	456241	429	0,511323	0,00094
anticipation	sentisense_anticipation	144	456241	99	0,6875	0,000217
disgust	NRC_disgust	1058	453955	484	0,457467	0,001066
disgust	sentisense_disgust	536	453955	275	0,51306	0,000606
disgust	sentisense_hate	16	453955	7	0,4375	0,000015

MONGODB foglio 2/2

fear	NRC_fear	1476	425352	613	0,415312	0,001441
fear	sentisense_fear	155	425352	93	0,6	0,000219
joy	EmoSN_joy	3733	446726	2300	0,616126	0,005149
joy	NRC_joy	689	446726	415	0,602322	0,000929
joy	sentisense_joy	128	446726	61	0,476562	0,000137
sadness	NRC_sadness	1191	450440	519	0,435768	0,001152
sadness	sentisense_sadness	132	450440	52	0,393939	0,000115
surprise	NRC_surprise	534	453130	293	0,548689	0,000647
surprise	sentisense_surprise	29	453130	16	0,551724	0,000035
trust	NRC_trust	1231	436284	614	0,498781	0,001407

Nuove parole legati ai sentimenti

NR	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
1	overrated	got	roll	taylor	girll	randomly	asked	randomly
2	's	one	tide	kept	n't	got	mama	got
3	spell	sitting	bandwagon	throwing	shot	really	wanted	really
4	fired	im	fan	ball	taylor	hot	beer	hot
5	youp	garage	hop	gutter	kept	wish	responded	love
6	noo	set	alabama	get	throwing	social	offered	baby
7	haha	drag	's	whatt	ball	life	drug	heart
8	roll	racing	dick	got	gutter	want	fuck	forever
9	tide	watching	one	one	get	wine	hour	best
10	bandwagon	ksl	wor	sitting	strike	celebrate	wait	wont
11	fan	getting	overrated	im	whatt	'm	grand	ever
	.....	....	....	....	....	....	....	.....
30218	michigan	-	-	-	salpointe	-	xoxoxox	biankis

Nuove parole dalle risorse

NR	anger, disgust, fear, sadness (neg)	joy, trust (pos)	anticipation, surprise (neutral)
1	abandon	abilities	activate
2	abandoned	ability	ad
3	abandons	aboard	adhere
4	abducted	absolve	adjust
5	abduction	absolved	admit
6	abductions	absolves	aggressive
7	abhor	absolving	alcove
8	abhorred	absorbed	alley
9	abhorrent	accept	aloof
10	abhors	accepted	ankle
11	absentee	accepting	antique
	…..	…..	….
15884	scary	-	-

Cosa abbiamo tratto da questa esperienza

MARIADB

- Scalabile solo verticalmente
- Ottimo con dati fortemente strutturati
- Più facile gestire query complesse e reports

MONGODB

- Scalabile sia verticalmente che orizzontalmente
- Rapido per analisi dati
- Più facile da gestire in team, da modificare radicalmente

CONCLUSIONI

Relazionale

Ideale per query complesse, transazioni multiple e analisi di routine.

NoSQL

Gestisce bene molte richieste, transazioni e attività.

GRAZIE!

Domande? Scriveteci!

damiano.gianotti@edu.unito.it
lorenzo.tabasso@edu.unito.it