

TWEETS ANALYSIS

Gianotti - Tabasso

Progetto MAADB 2020/21

CONTENUTO DELLA PRESENTAZIONE

Ecco un riassunto di cosa tratteremo oggi nella presentazione

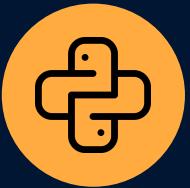
- Ambiente di sviluppo
- Introduzione e *filosofia*
- Gestione risorse e come trattarle
- DMBS a confronto:
 - Relazionale - *MARIADB*
 - Not Only SQL - *MONGODB*
- Analisi dei risultati, tre elementi di valutazione:
 - WORD CLOUD
 - HISTOGRAMS
 - SPREADSHEET

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

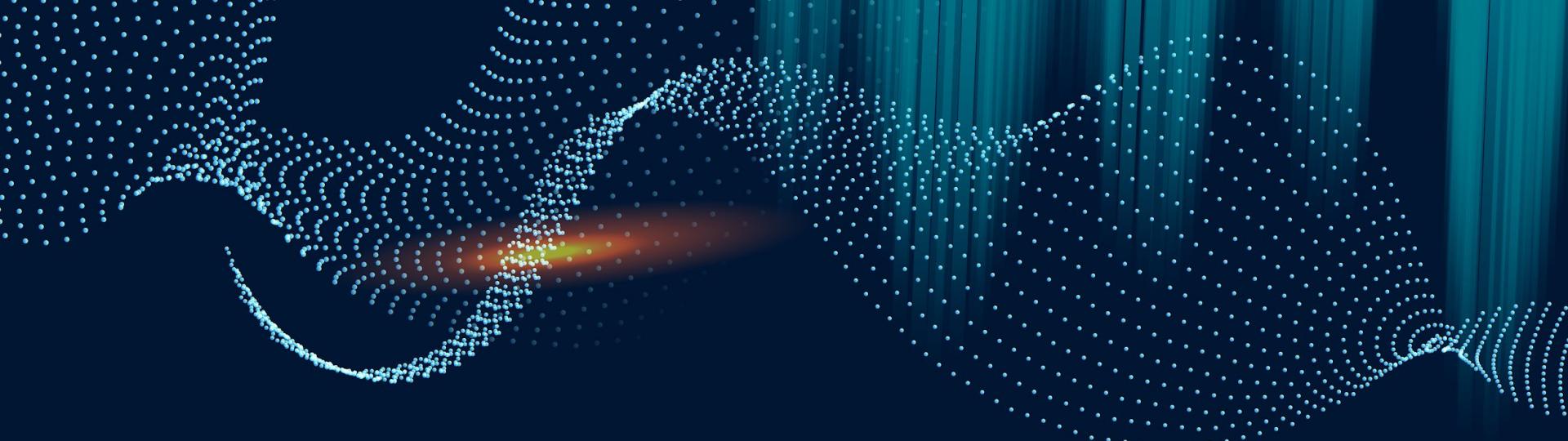
DATABASE

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 generiche**

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

- **Sentimenti:**

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

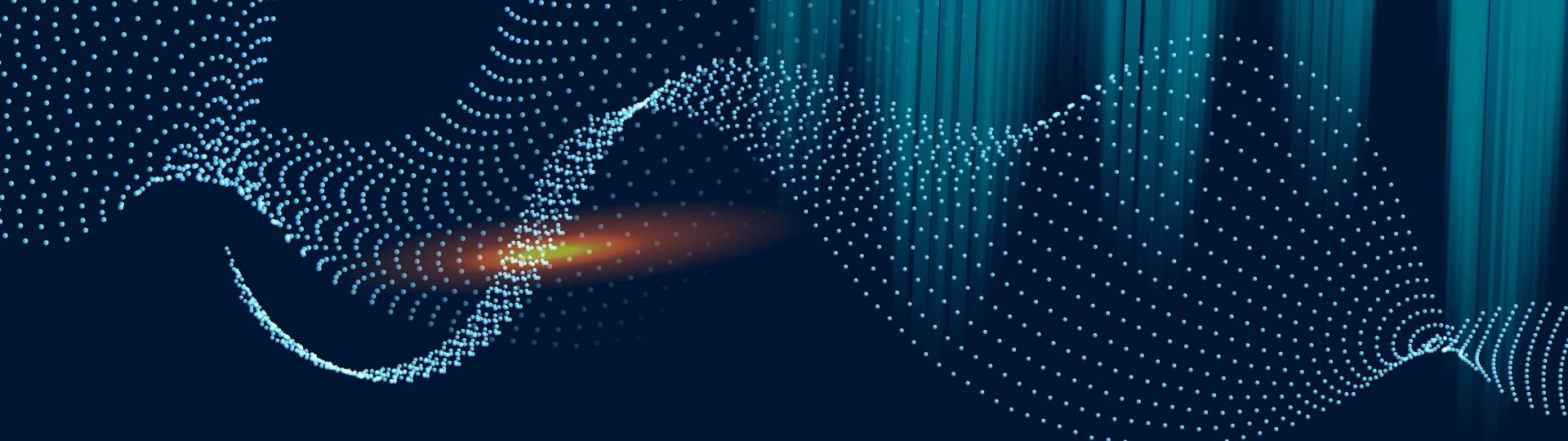
```
lexical-resources
|   elenco-parole-che-negano-parole-successive.txt
+
+--Generic
+   +--ConScore
+     afinn.txt
+     anewAro_tab.tsv
+     anewDom_tab.tsv
+     anewPleas_tab.tsv
+     Dal_Activ.csv
+     Dal_Imag.csv
+     Dal_Pleas.csv
+
+--Hope
+   sentisense_hope.txt
+
+--Like-Love
+   sentisense_like.txt
+   sentisense_love.txt
+
+--Neg
+   GI_NEG.txt
+   HL-negatives.txt
+   listNegEffTerms.txt
+   LIWC-NEG.txt
+
\---Pos
+   GI_POS.txt
+   HL-positives.txt
+   listPosEffTerms.txt
+   LIWC-POS.txt
+
\---Sentiments
+   +--Anger
+     EmoSN_angry.txt
+     NRC_angry.txt
+     sentisense_angry.txt
+
+   +--Anticipation
+     NRC_anticipation.txt
+     sentisense_anticipation.txt
+
+   +--Disgust
+     NRC_disgust.txt
+     sentisense_disgust.txt
+     sentisense_hate.txt
+
+   +--Fear
+     NRC_fear.txt
+     sentisense_fear.txt
+
+   +--Joy
+     EmoSN_joy.txt
+     NRC_joy.txt
+     sentisense_joy.txt
+
+   +--Sadness
+     NRC_sadness.txt
+     sentisense_sadness.txt
+
+   +--Surprise
+     NRC_surprise.txt
+     sentisense_surprise.txt
+
\---Trust
+   NRC_trust.txt
```

Gestione delle risorse per il preprocessamento

Per il preprocessing abbiamo:

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

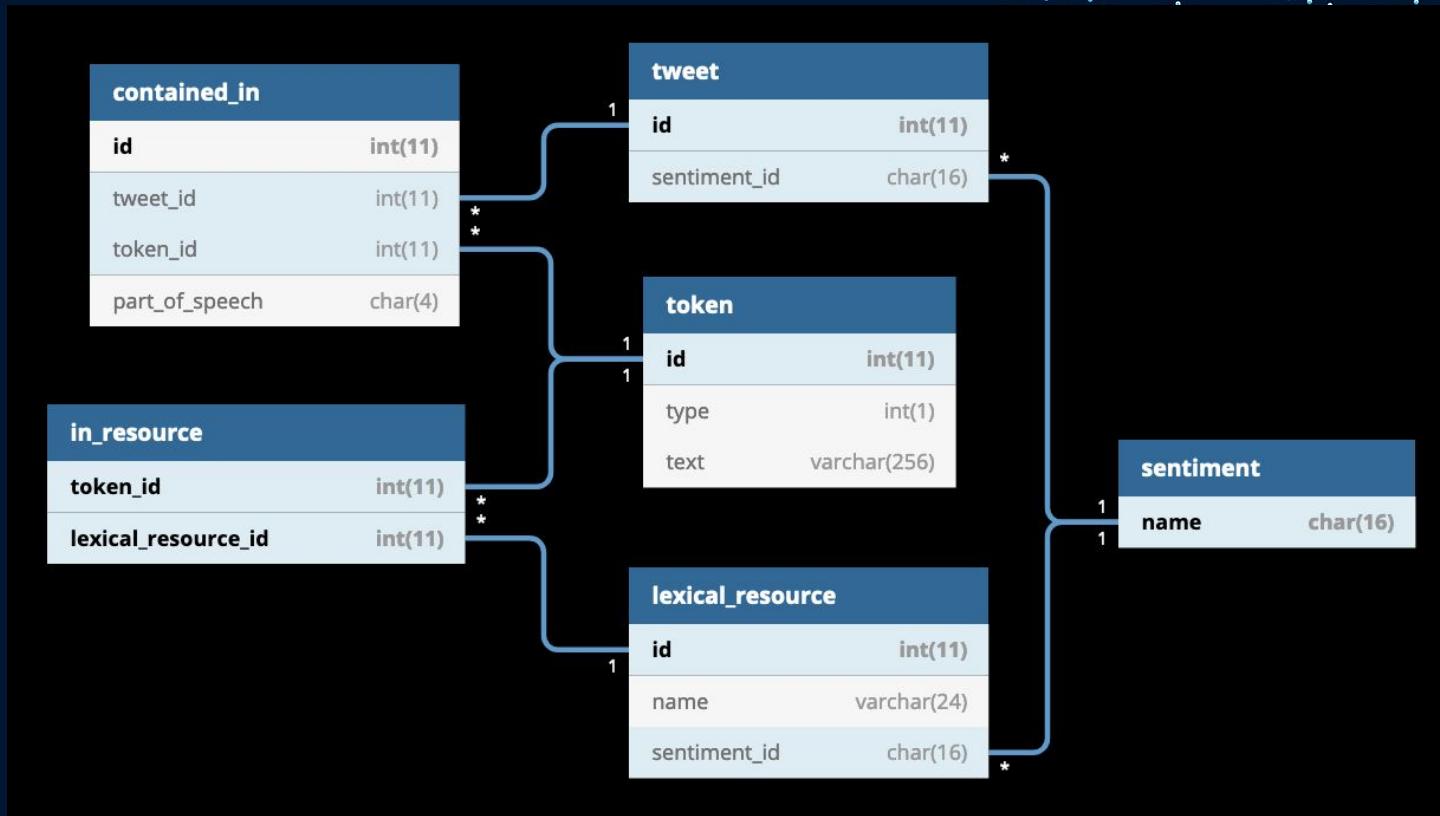
```
+--processing
|   emoji.json
|   emoticons.json
|   Penn Treebank P.O.S. Tags.html
|   Penn Treebank P.O.S. Tags.png
|   punctuation.txt
|   sentiment2emo.json
|   slang_words.json
|
\---twitter-messages
    dataset_dt_anger_60k.txt
    dataset_dt_anticipation_60k.txt
    dataset_dt_disgust_60k.txt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

```
\---projects
  |   manage_scoring_resources.py
  |   test.py
  |
  +---NoSql
  |   | .....
  |
  \---Relational
      |   relationaldbhandler.py
      |   relational_analysis.py
      |   relational_create.py
      |   relational_preprocessing.py
      |   relational_test.py
```

Schema collezioni MONGODB

Mongo Express Database: tweet_analytics

Collections

	View	Export	[JSON]	Import	Collection Name	+ Create collection
					common_words	Del
					freqencies	Del
					lexical_resources	Del
					sentiments	Del
					tweets	Del

Database Stats

Collections (incl. system.namespaces)	
Data Size	74.2 MB
Storage Size	40.0 MB
File Size (on disk)	0 Byte
Avg Obj Size #	144 Bytes
Objects #	511875
Indexes #	5
Index Size	6.48 MB

Nosql MongoDB: struttura e compiti

Per il preprocessing abbiamo:

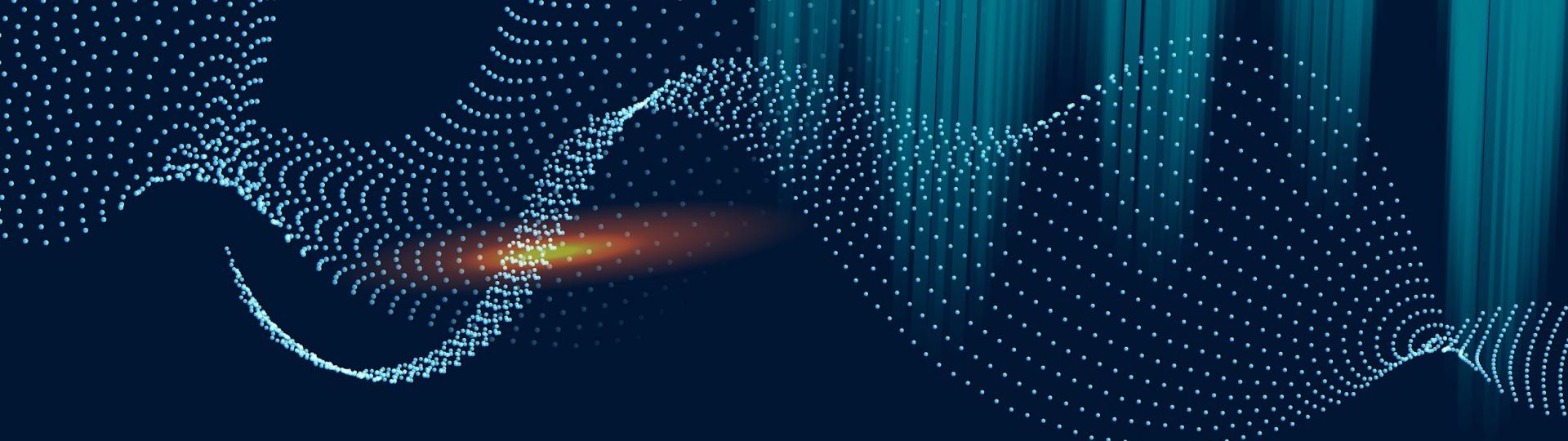
- Handler: interfaccia con il DBMS
- Create: inizializzazione handler e preprocess
 - Caricamento risorse lessicali
- Test: Unittest dell'handler
 - Classe di supporto
- Common Words: calcolo parole comuni
 - Tra le risorse lessicali e tweets
- Analisi: generazione statistiche

```
\---projects
    |   manage_scoring_resources.py
    |   test.py
    |
    +--NoSql
        |   .gitignore
        |   common_words.py
        |   noslq_preproocsesing.py
        |   nosqlbdbhandler.py
        |   nosql_analysis.py
        |   nosql_create.py
        |   nosql_test.py
        |
        \---Relational
            |   .....
```

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

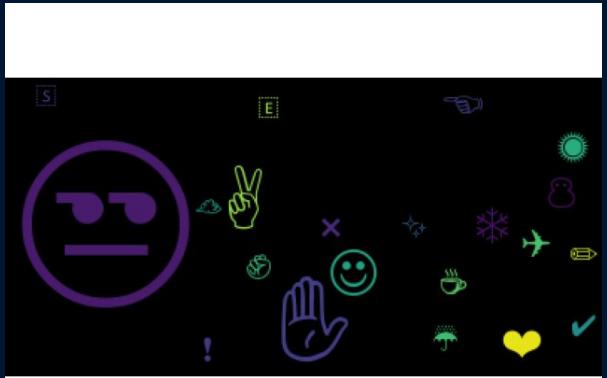
Confronto statistico
tra i due diversi
approcci tramite
fogli di calcolo

WORD CLOUDs

Quattro per ogni sentimento

today need make take say people see ca one really
gOwWa goodly going
still ha phone mom think someone see doe really
back want even never even never
want even never even never
know laughing bitch

• - •) : * [i : -) : (v : -)
:[: - (:) : - (^ ^ : -)
:^) : - (^ ^ : -)
:^) : - (^ ^ : -)



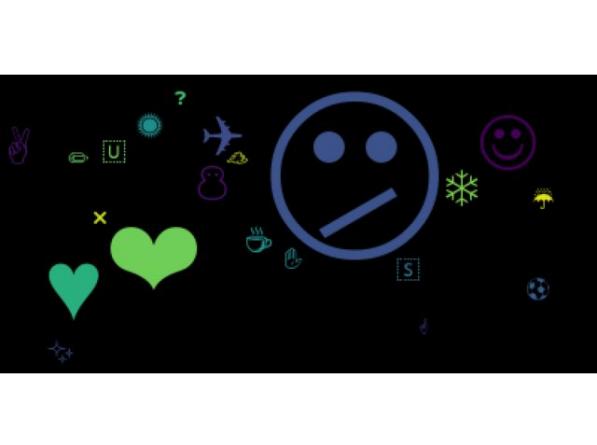
S E ↗
fuck you job annoyed
ugh pissed
annoying
suh confessionnight in7than tired goaway bored sorrynotsorry
imiddlechool shutup tweetmyjobs internship
batch seriously in7thgrade
wtf batch sadtweat fml lol ughh
sick ohwell portland
stfu stfu in7thgrade frustrated fml
oomf stop realestate te goodnight houston mad

Rabbia ANGER

Words | Emoticons

Emojis | Hashtags

make much think last time still want
know time go loud today need want
got ca miss one well na people back see
go night good im week never work
going tomorrow feel school
laughing even bad home really thing



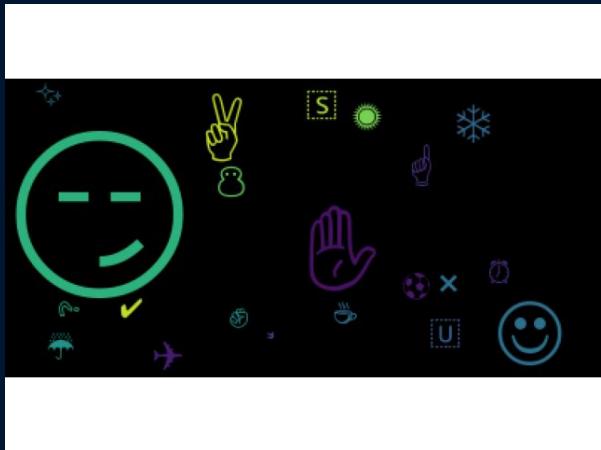
sad tweet sorry nervous exciting christmas family oomf yay love finally happy forever alone
cant wait best friend waiting ready
pumped confused sick oh well twilight ready
excited bored mission dol food the walking dead
the struggle anxious sad full black friday so pumped hurry up home
too excited confession night so excited scared lover

Aspettativa ANTICIPATION

Words | Emoticons

Emojis | Hashtags

na
re
hate today
time
W
a
g
o
time
ne
ed
ee
ve
ot
goodday
laughing
go
ing
o
back
some
girl
say
in
really
ca
night
home
need
even
still
people
make
thing
one
alwys
got
one
admys
know
loud
slut
think



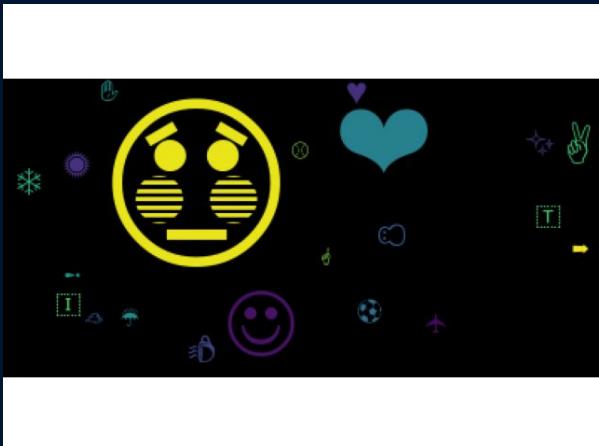
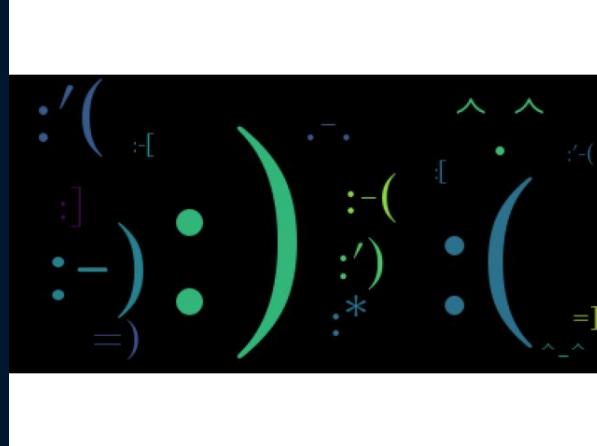
annoyed
yuck
internship
oomf
in7than
oh well
nasty
in7thgrade
boring
sorry noboby
line
hateit
going

Disgusto DISGUST

Words | Emoticons

Emojis | Hashtags

never re going think night someone
day know go
still today come thing last friend
say loud need one well
laughing even really
tomorrow see got people
time want makeback home feel
ve



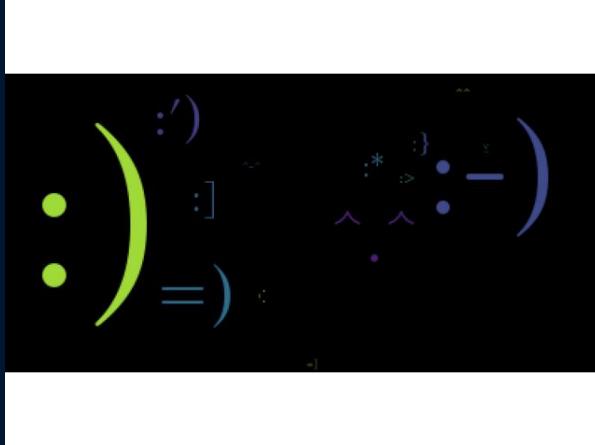
sorrynotsorry in9thgrade yay SCary nervous lol
inst9thgrade please bestfriend catfish happy happytwee
wtf excited awk stressed oomf loveit
in6thgrade fml family inmiddleschool
sad tired happygirl in7thgrade weirdthewalkingdead
awkward thatawkwardmoment oops creep
thatawkwardmomentwhen followtrain whatmakesmesmile
scared decisions help

Paura FEAR

Words | Emoticons

Emojis | Hashtags

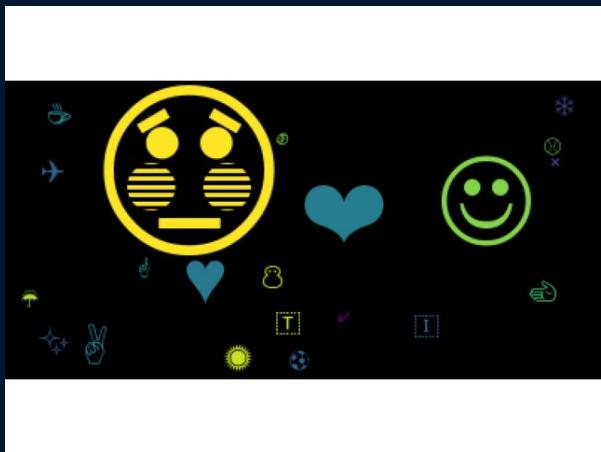
makehaha
haha
W^atch^{back}
got^{home}
one^{friend}
thanks^{come}
daynight^{ca}
really^{well}
got^{re great}
well^{need}
good^{xd}
new^{re girl}
good^{going}
see^{na}
tomorrow^{today}
want^{still}
time^{think}



Gioia JOY

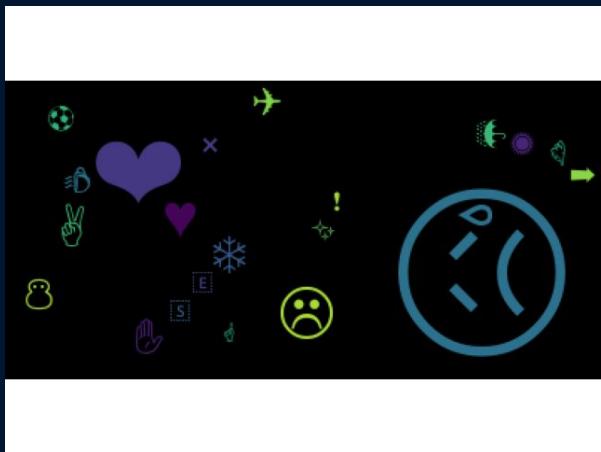
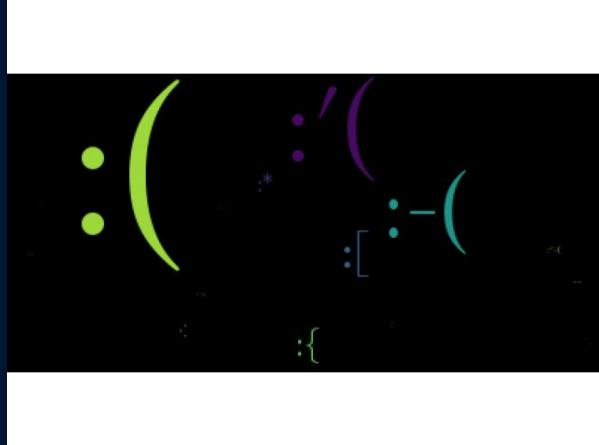
Words | Emoticons

Emojis | Hashtags



wifollowspree • beautiful
finally winning sorrynotsorry yum yay life
loveher picstitch birthday boyfriend blessed
dying orlando skyfall happy
thenotebook cute photooftheday food cantwait thewalkingdead
love loveyou iloveyou oomf excited
instagood lovehim family party
confessionnight goodnight dead bestfriend friends
goodnight dead bestfriend friends
teamfollowback

back got • laughing make nave really wish still
miss wan work
friend much one never think
some home time sad good
know today come need
W**ant** day last hate
night loud feel see going love
bad



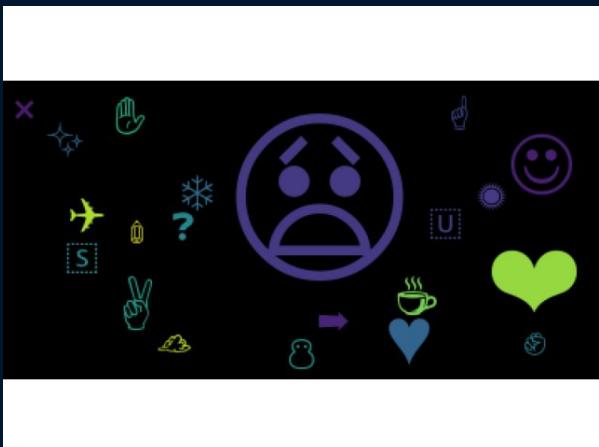
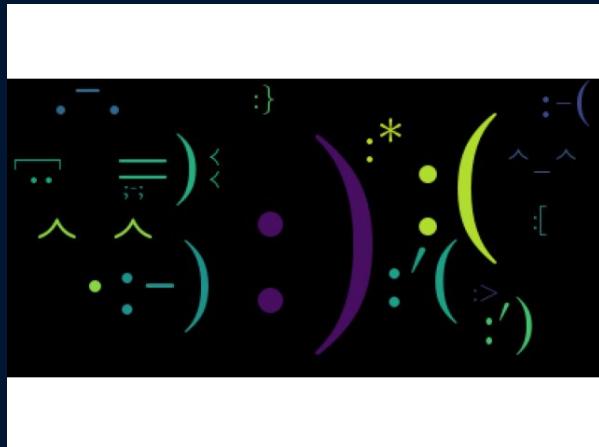
sadday wahh in8thgrade depressing
firstworldproblems thewalkingdead bored in9thgrade lonely
sadtweet ugh
ohwell heartbroken upset crying sucks rip
sosad fail depressed sick imissyou
ouch fml help forevalone
depression fail im7thgrade fuckmylife
notcool im6thgrade smh
immiddle school missyou in7thgrade boo
imforyou iwillneverunderstand thenotebook wah
imforyou wtf sorry
confessionnight wah

Tristezza SADNESS

Words | Emoticons

Emojis | Hashtags

na make someone ha 00 going go still
know guy good • mom say girl
laughing last back see ca
time even one really hadia
ve never people today feel
re got loud oo think night want
said day thing



omg funny catfish family creepy
weird seriously dying inothgrade thanksgiving
photodthday scared fml dead
lol help lmao really
smh awesome ugh scary confused
awesomeness hmlfao gross
blackfriday pissed confessionnight amazing ohwell
sad blackfriday cute random love newyork
haha weird wow crazy oomf

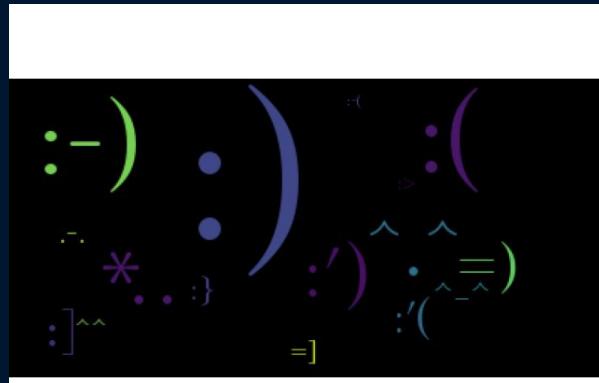
Sorpresa SURPRISE

Words | Emoticons

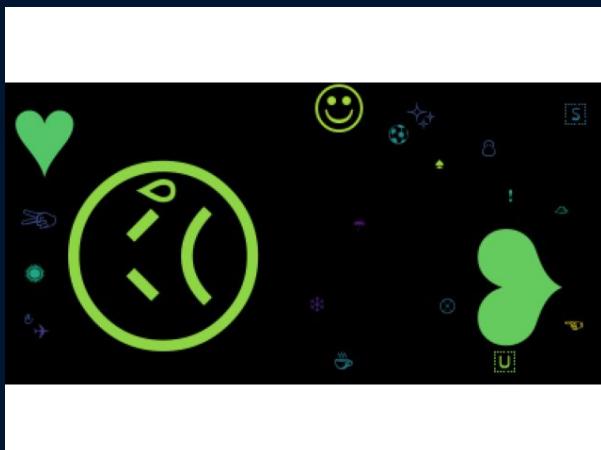
Emojis | Hashtags



want home never time make need missthink
ca friend always life tonight know come back
wait birthday best see one
got go girl go
baby much laughing
night good today loud going happy



:-) :) :-(
* :) ^ ^ =)
:] ^ ^ :)'(^ - ^)
=]



blessed mention20goodlookingpeopleontwitter photooftheday
instamood family
friends amazing tweegram iphonesia
lovethem cute
iloveyou instagood fun
thankyou loveher birthday follow
rollidole instagood bff
cowsaysonation lovehim oomf teamfollowback
nyc sexy life picstitch happy
lovehim confessionnight excited



bestfriend beautiful np
food inlove if
loveyou instagood fun
iloveyou loveher birthday follow
thankyou thenotebook bff
boyfriend missyou e3followspree yum
sexy nyc life teamfollowback
oomf lovehim confessionnight excited

Fiducia TRUST

Words | Emoticons

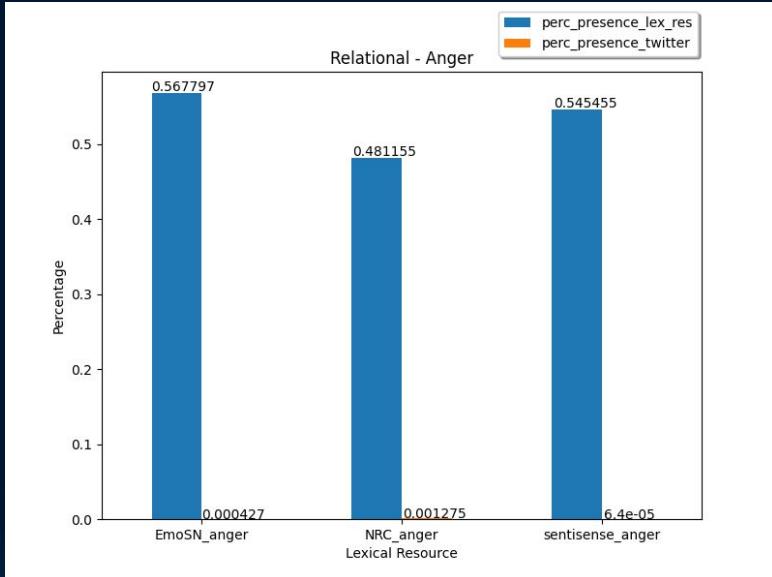
Emojis | Hashtags

HISTO GRAMMS

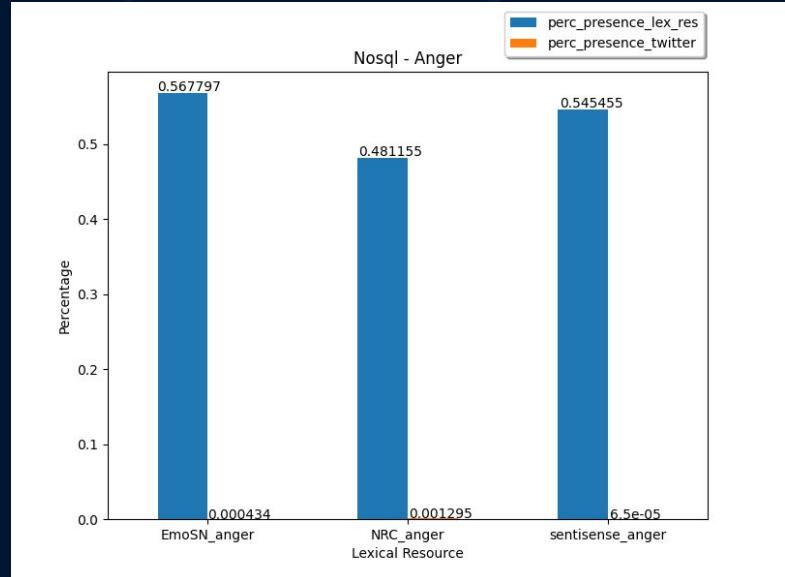
Tool per confronto tra sistemi

Anger

MARIADB

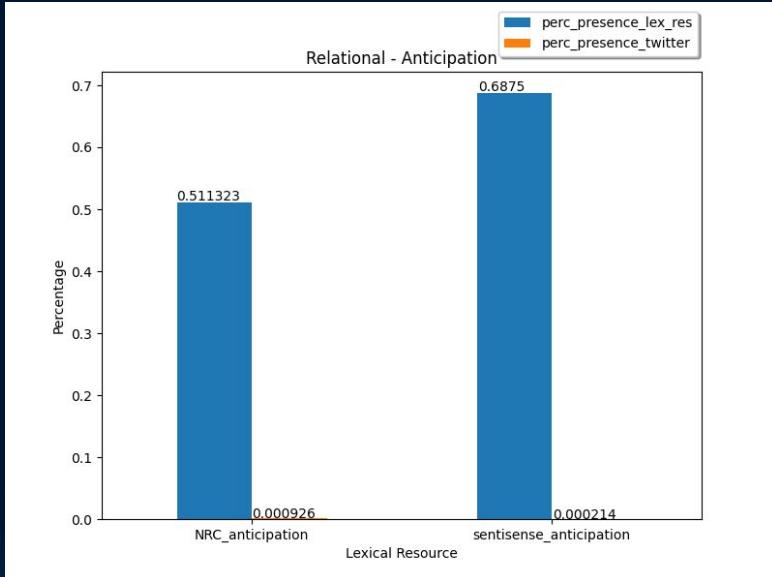


MONGODB

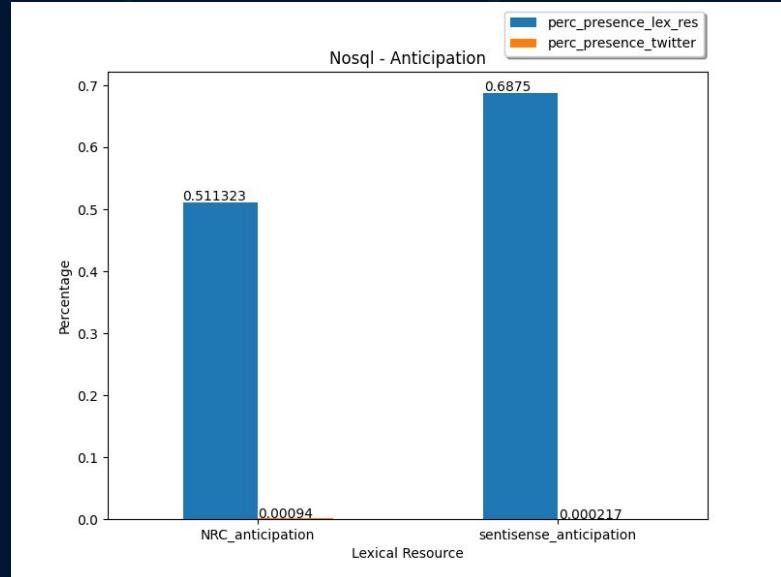


Anticipation

MARIADB

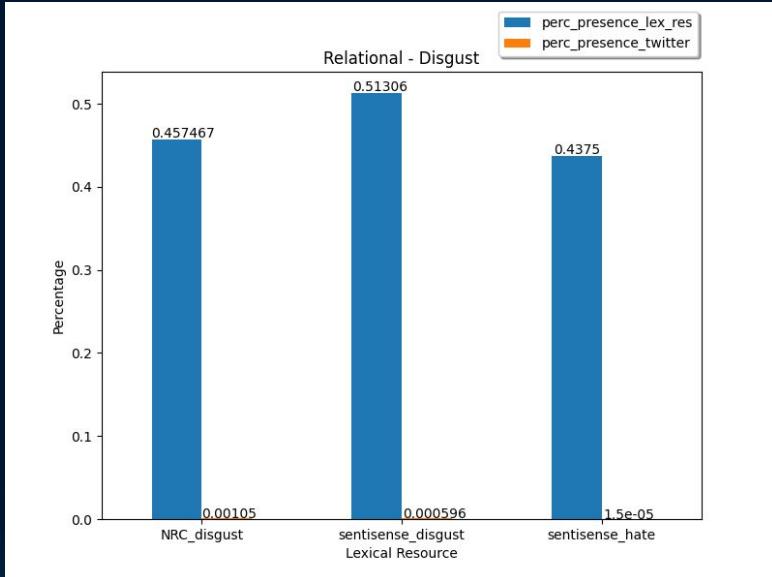


MONGODB

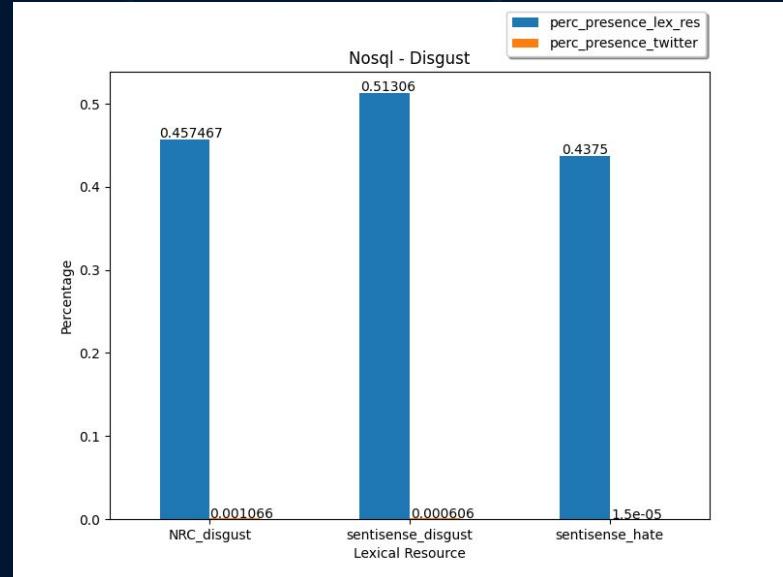


Disgust

MARIADB

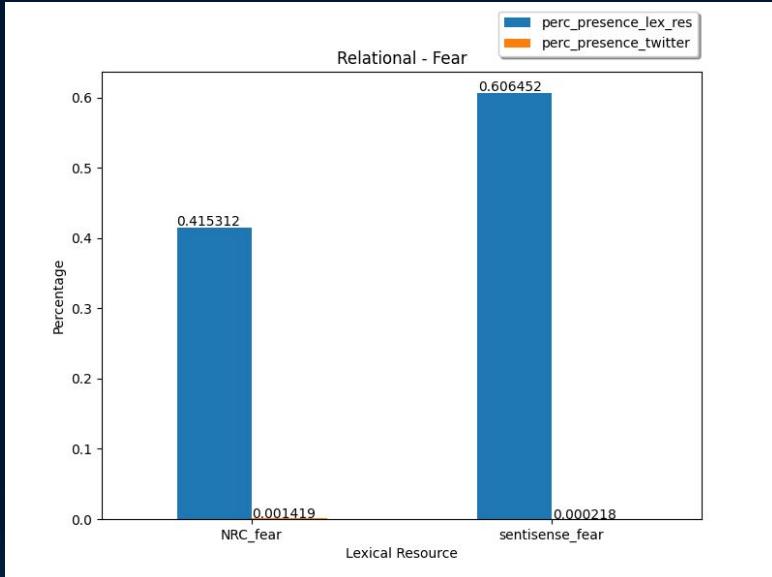


MONGODB

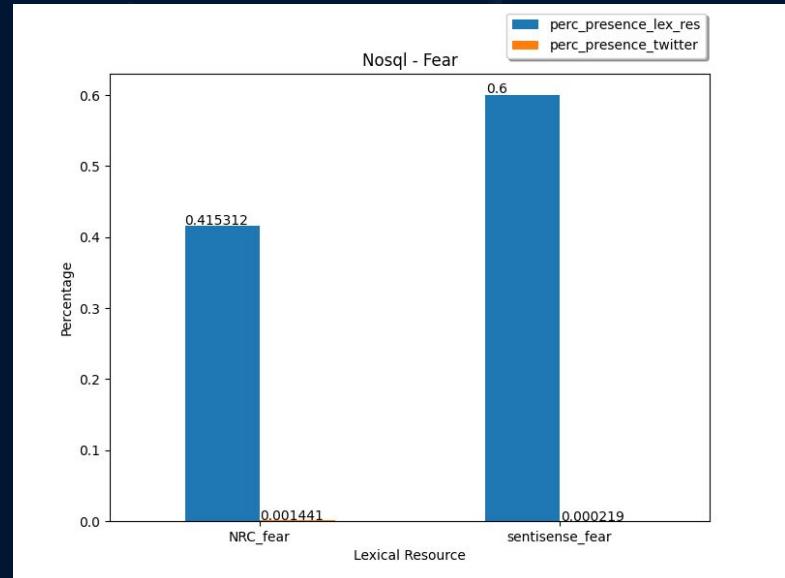


Fear

MARIADB

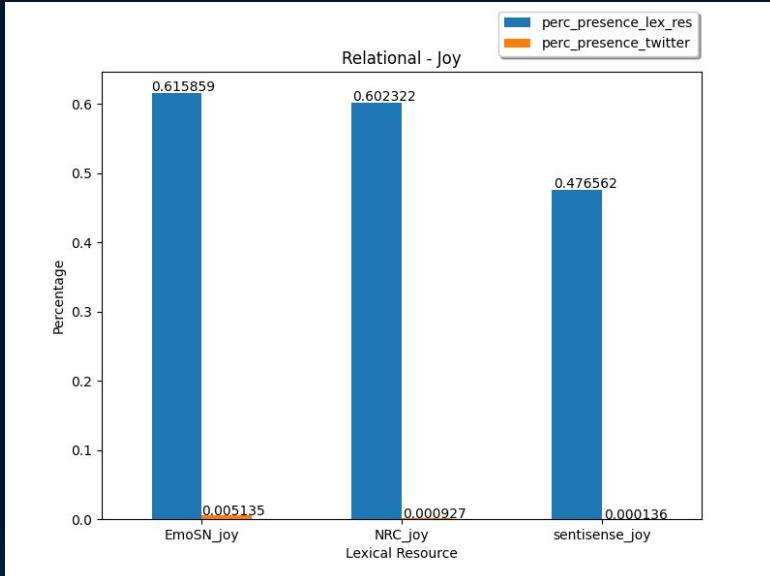


MONGODB

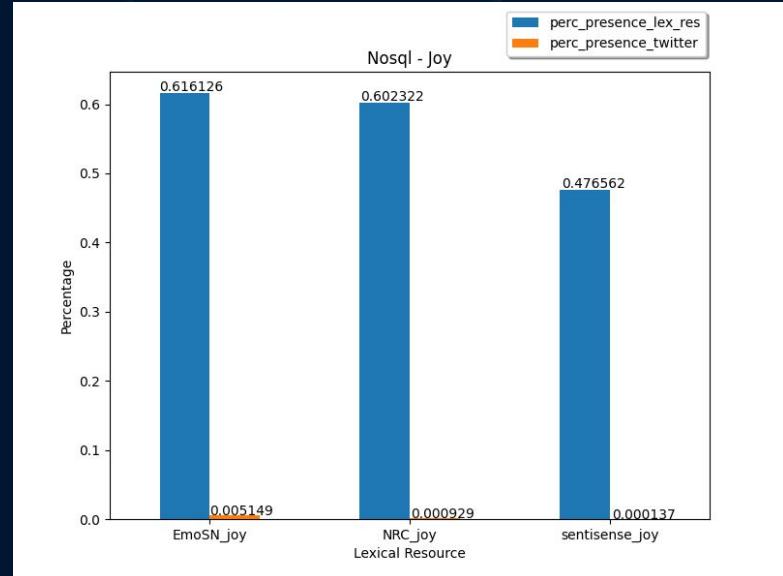


Joy

MARIADB

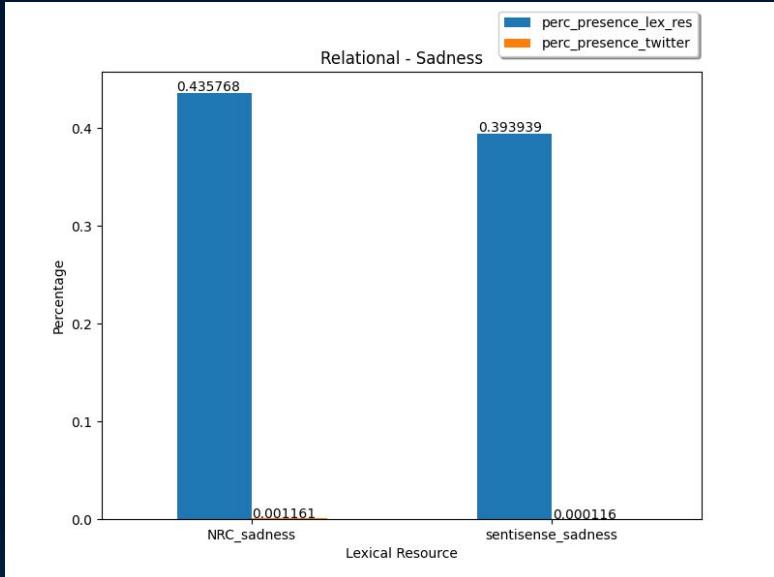


MONGODB

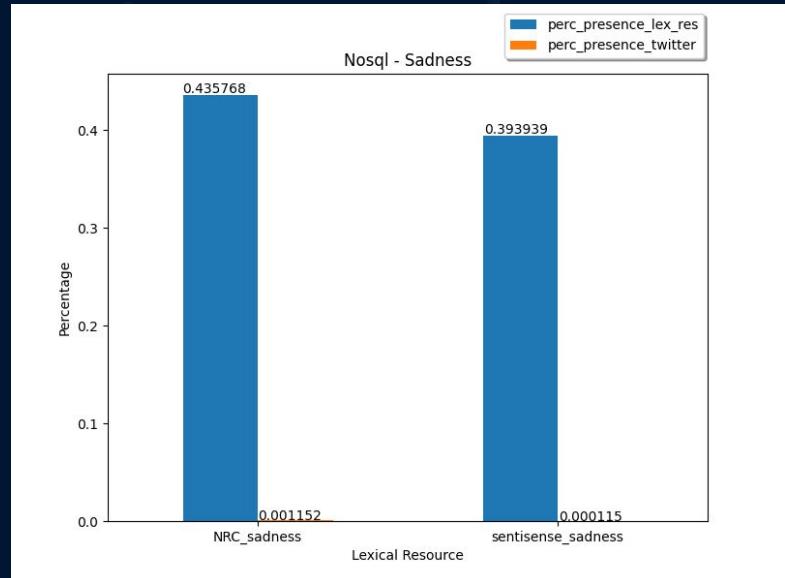


Sadness

MARIADB

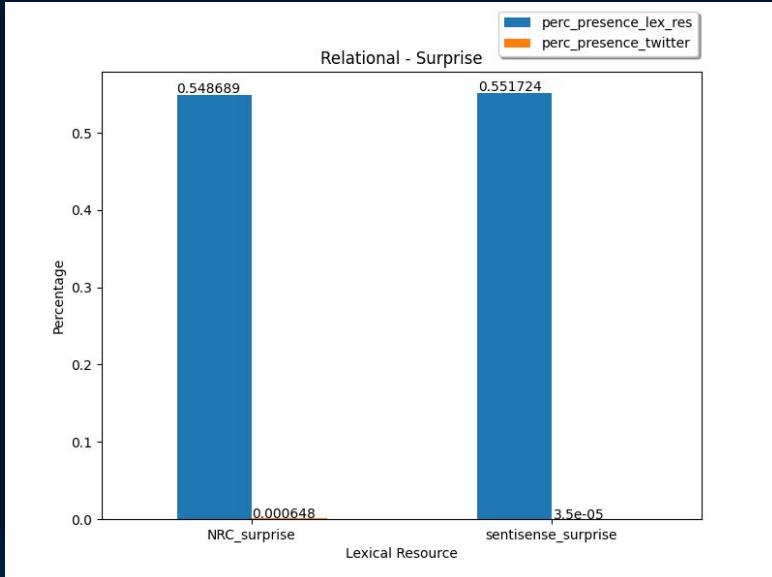


MONGODB

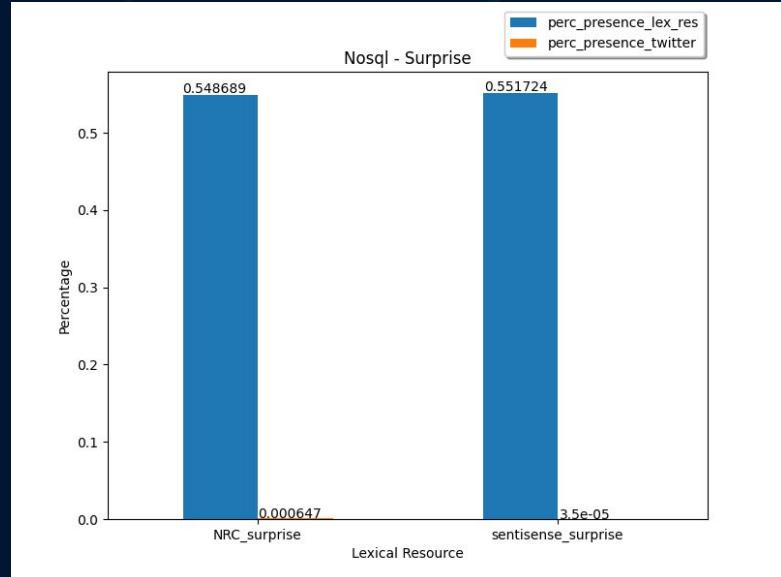


Surprise

MARIADB

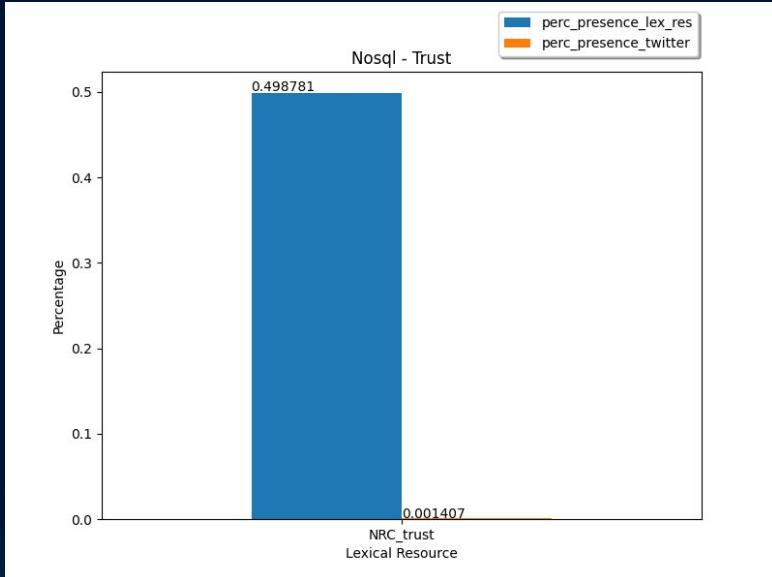


MONGODB

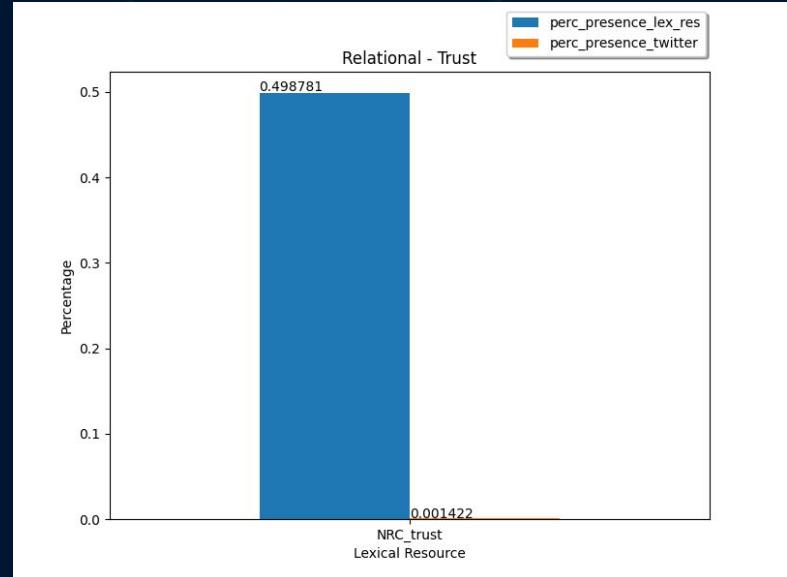


Trust

MARIADB



MONGODB



SPREAD SHEET

Statistiche & new sentiments

MARIADB foglio 1/2

sentiment	lex_resource	n_lex_words	n_twitter_words	shared_words	perc_presence_le x_res	perc_presence_tw itter
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_le x_res	perc_presence_tw itter
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	throwin g	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	bandwago n	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



CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

Please keep this slide for attribution.

