

CHALLENGE

HOW MUCH CAN YOU SCRAPE TWITTER?



Antonello Manenti

Master in Business Intelligence and Big Data
Analytics University of Milano-Bicocca

DURING THE #EURO2020 MATCHES
OF THE ITALIAN NATIONAL TEAM

920_528 TWEET

di cui buoni per l'analisi 828_635
(no doppioni, no dati importanti mancanti
no tweet < 3 parole, no lingue strane)

STREAMING: '330

220_647 TWEET
BELGIOVSITALIA

02/07/2020

STREAMING: '105

119_762 TWEET
ITALIAVSSPAGNA

06/07/2020

STREAMING: '103

580_111 TWEET
INGHILTERRAVSITALIA

11/07/2020

STREAMING: '122



STRUMENTI USATI

- ☐ Colab + GoogleDrive
- ☐ Python
 - ☐ Tweepy (scraping)
 - ☐ Pandas (cleaning, ETL, preproc.)
- ☐ Gephi

Partita Belgio-Italia

qualche dato sui giocatori in campo

1. Immobile:6812

2. Lukaku:6019

3. Insigne:5758

4. Spinazzola:5695

5. Barella:4287

TopFive

più citati
(nel testo, come @ o #)



TopFive giocatori “più amati”

**totale tweet poco gradevoli
(in lingua italiana)**

1. **Vincic 222**
2. VAR 221
3. Spinazzola 175
4. Immobile 132
5. Lukaku 130

1. VAR 37%
2. Doku 14.2%
3. Lukaku 12.2%
4. Vincic 11.2%
5. De Bruyne 10.6%

**% di tweet insolenti
(in lingua italiana sul totale)**

**% di tweet in cui si insulta la persona
specifica sul totale dei tweet in cui si
parla di quella persona**

Problemi

Problema 1: nei testi ci sono link, emoji, immagini, spazi a caso o nessuno spazio, @, #, cose senza senso e simboli imprevisti ()

Problema 2: 132 lingue diverse

Problema 3: capire in che modo il giocare è insultato

Problema 4: capire a quale giocatore riferire il contenuto del tweet

**Di Lorenzo non se lo caga nessuno.
É stato citato 33 volte e di questi in
soli 5 tweet in lingua italiana**

Le lingue più utilizzate sono state:
francese: 49812 tweet
inglese: 49455 tweet
italiano: 44825 tweet (26,4% dei tweet)

Seselwa: 1

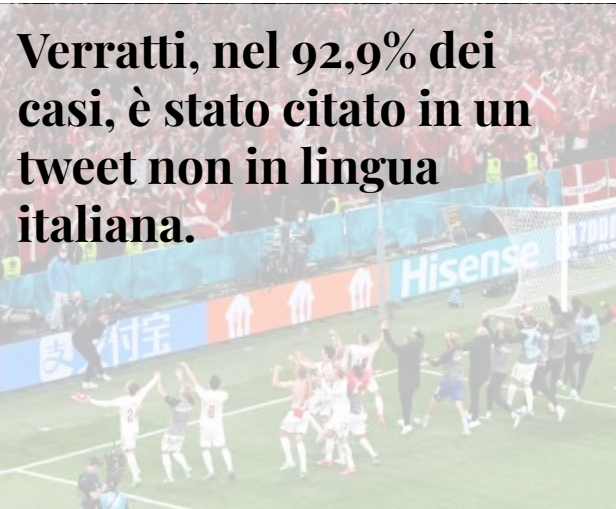
Shona: 1

Uyghur: 1


**tra il 6,3% e il 6,6% dei tweet in
lingua italiana contenevano
insulti a giocatori o arbitri**

Soluzioni


1. pd + emoji
2. polyglot (+PyICU+pycld2)
3. pattern di insulti da una lista creata a mano e `.str.contains(pattern)`
creare un dizionario di insulti multilingua che interpreti anche le emoji e le immagini in un sistema che parametri tutto in base ad elementi contestuali, semantici ed espressivi = impossibile
4. a mano con robe del tipo `p.loc[p['text_demoji'].str.contains('vialli')]`



Verratti, nel 92,9% dei casi, è stato citato in un tweet non in lingua italiana.



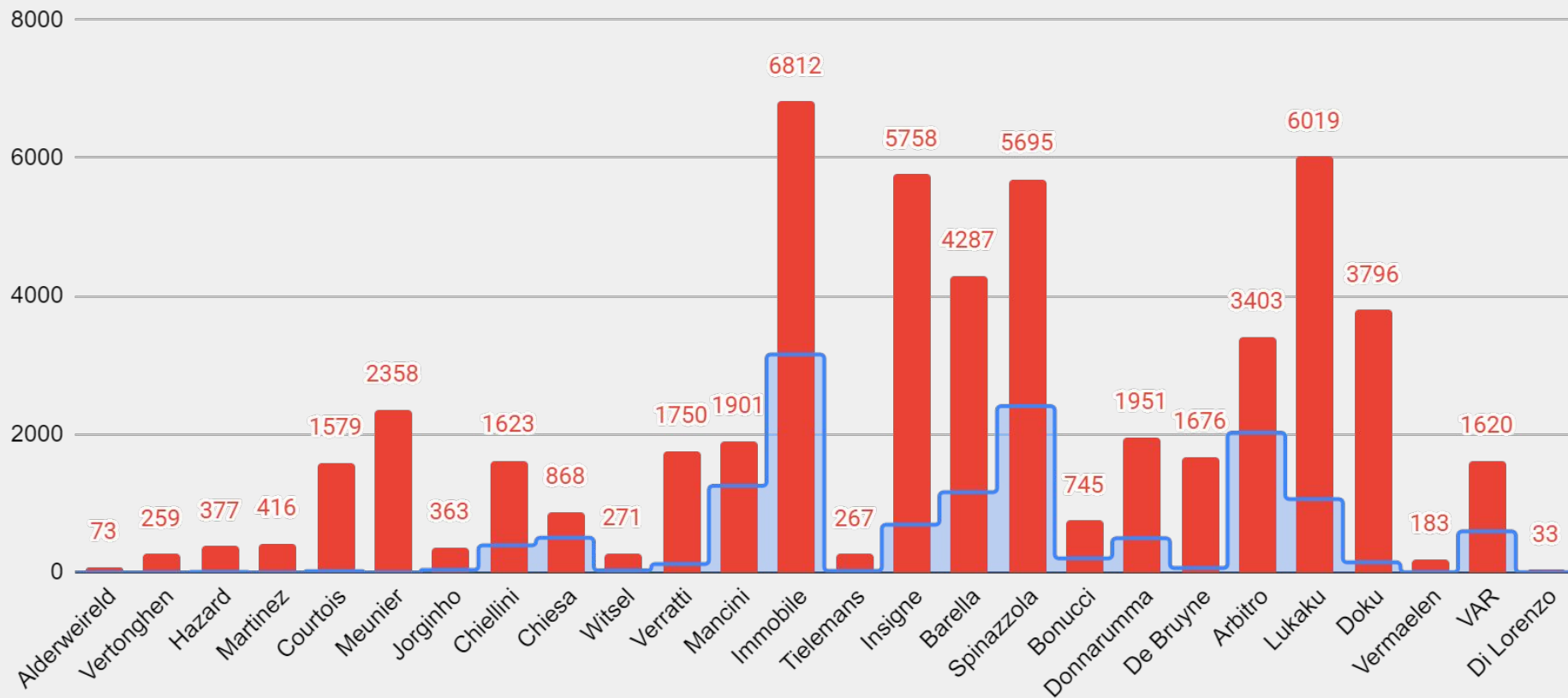
Mancini è stato citato nel 66% dei casi in tweet di lingua italiana. Vinvic nel 59,4%, Chiesa nel 58,1%



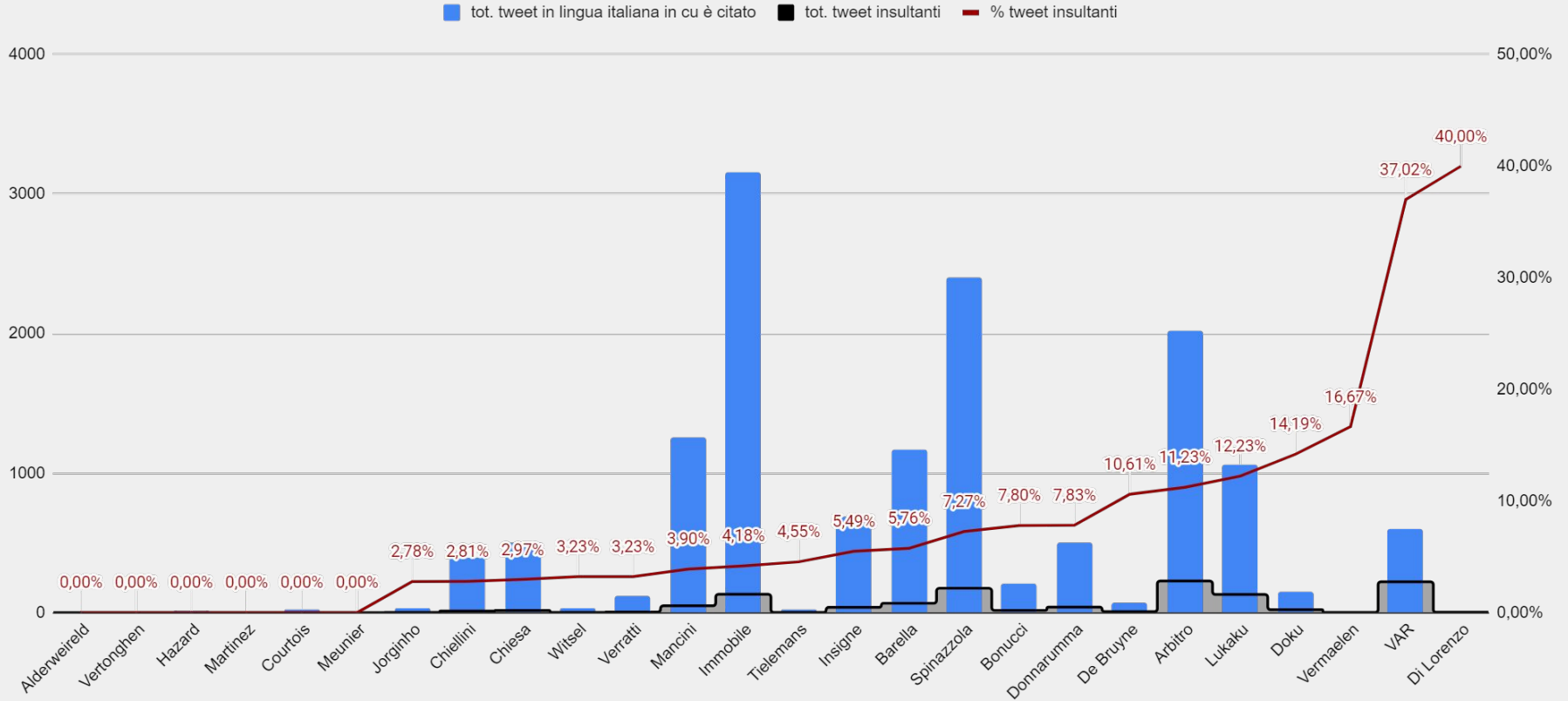
Jorginho e Chiellini sono stati insultati nel 2,8% dei tweet che li riguardano. É la percentuale più bassa tra i giocatori italiani.

Grafico_1

■ tot. tweet in cui è citato ■ tot. tweet in lingua italiana in cui è citato



Grafico_2



Che altro si potrebbe fare?

Conoscendo per ogni tweet se ci sono insulti, in che lingua è scritto e a quale/i giocatore/i si riferisce il tweet si potrebbe:

1. vedere la frequenza dei tweet nel corso di una partita e sapere cosa suscita maggiore interesse su specifici giocatori per ogni contesto linguistico;
2. capire per ogni lingua le abitudini all'insulto, cioè chi è più bersagliato, in cosa e perchè per ogni contesto linguistico;
3. ipotizzare il grado di competenza in altre lingue o indagare le interazioni tra appartenenti a gruppi linguistici diversi attraverso l'analisi delle interazioni tra gli utenti di diverse lingue

Es: attraverso quali lingue, per ogni partita, c'è maggiore interazione (retweet o altre interazioni) con tweet di lingue diverse?

Quanti e quali utenti di una specifica lingua interagiscono con altre lingue?

Ecc.

UPDATE

avendo già un dataset pulito, perchè
non provare a sbatterlo in Gephi
e vedere cosa ne esce?

Che succede se

vieni messo/a in una stanza
con 1000 persone
di 50 lingue diverse
e sei obbligato/a a comunicare?

Opzioni

1. Parlo nella mia lingua con chi conosce la mia lingua
2. Parlo con chi non conosce la mia lingua usando la loro lingua
3. Parlo con chi non conosce la mia lingua usando una lingua che entrambi sappiamo
4. Comunico in qualche modo, anche a gesti

Table2Net for Gephi

Bipartite network
undirected
two type of nodes

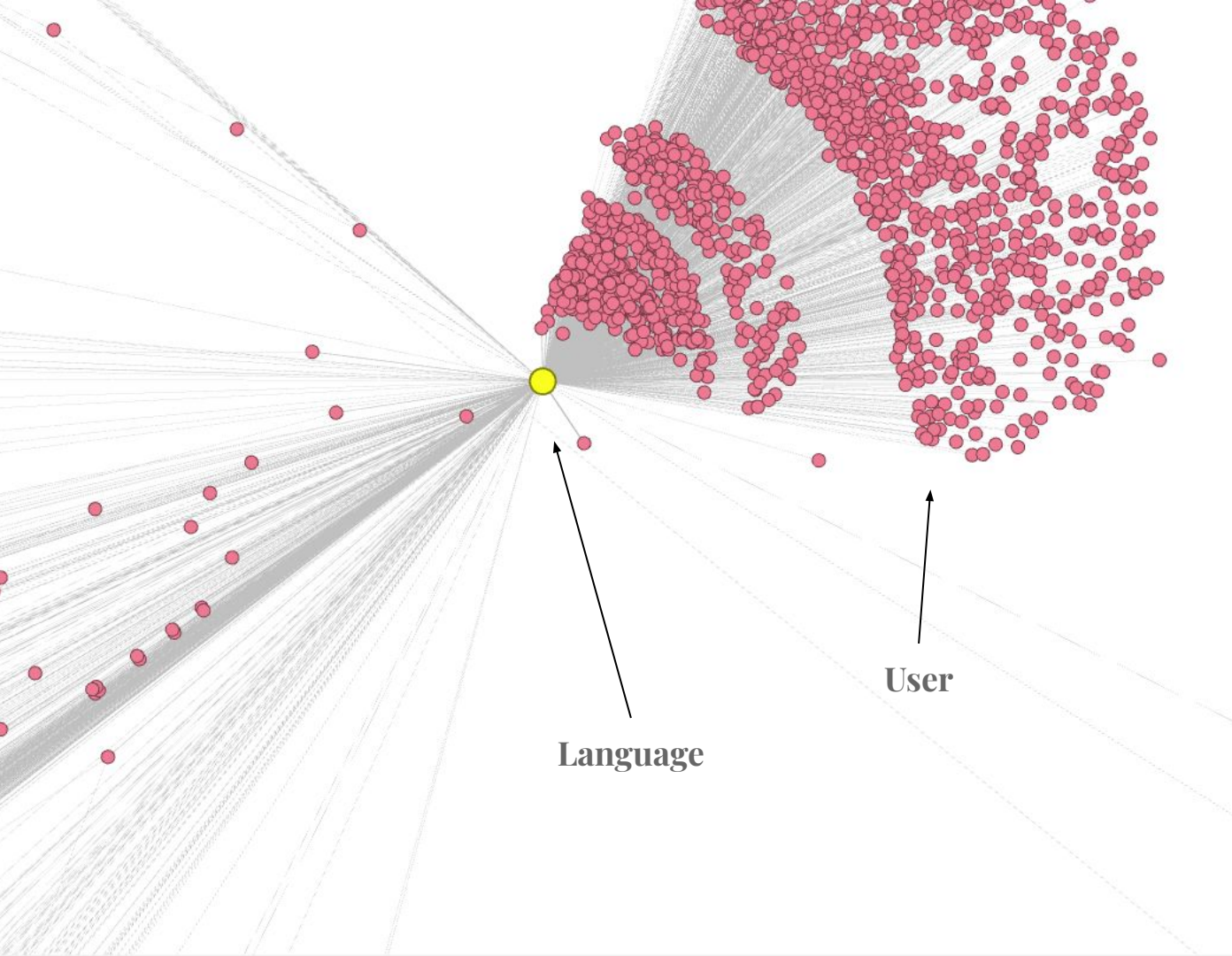
First node: **user**
Attributes: **language**,
hashtag, **mentions**.

Second node: **language**
Attributes: **language**,
hashtag, **mentions**.



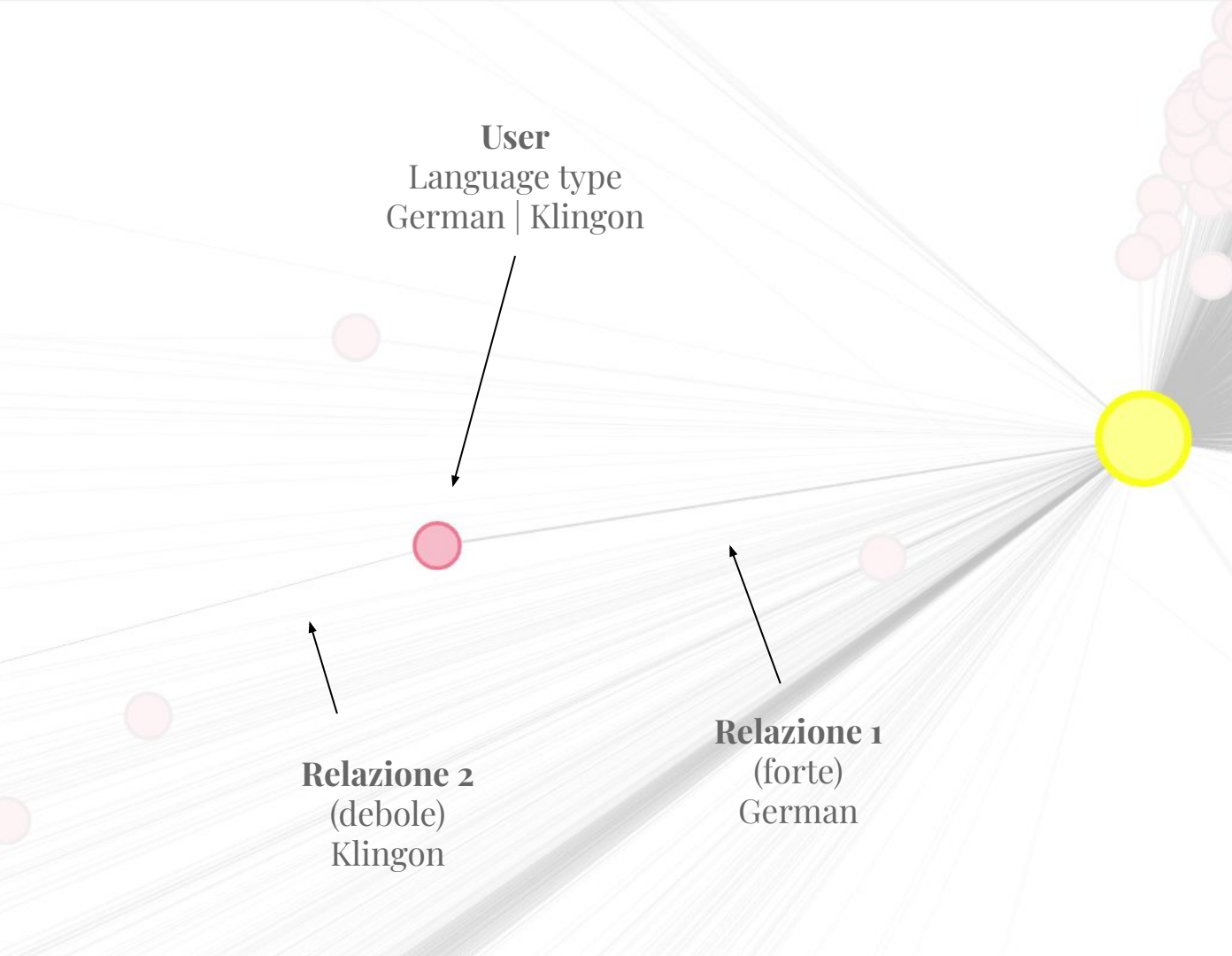
Problemi

1. Gephi crasha di continuo
2. La spazializzazione ha richiesto 3 ore



Soluzioni

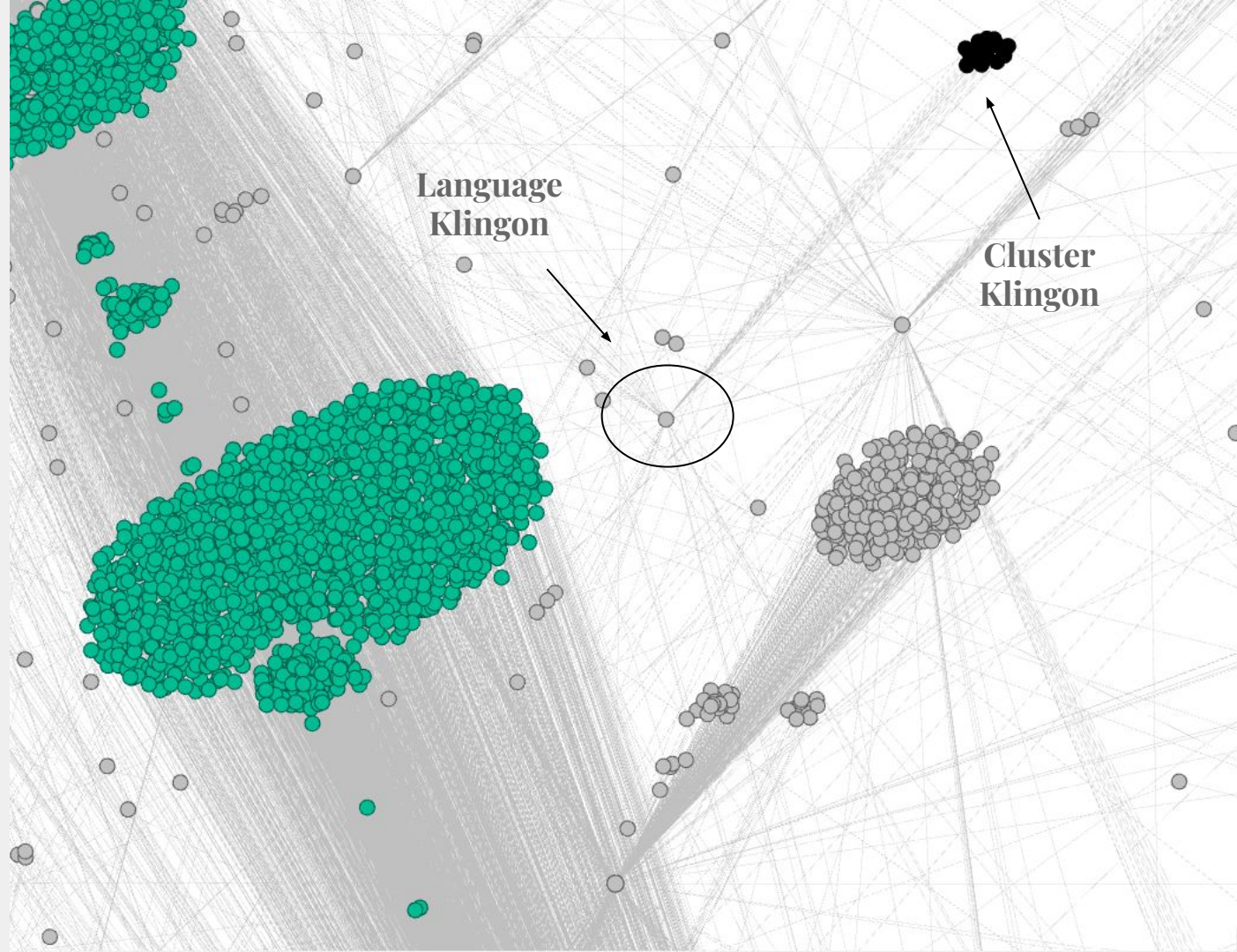
1. Salvare ogni 3 secondi
2. Gephi lavora in locale, è ora di cambiare PC



Cluster Klingon

users

an5ore
anninortheast
chris_a_tye
epaaasuka
fian_1927
jim_na_jim
jimsjimbo
lesanchonl
luvbokie
mehffi
salomonbreezy
tparada_
yesaniser
zapletysya



Cluster Klingon

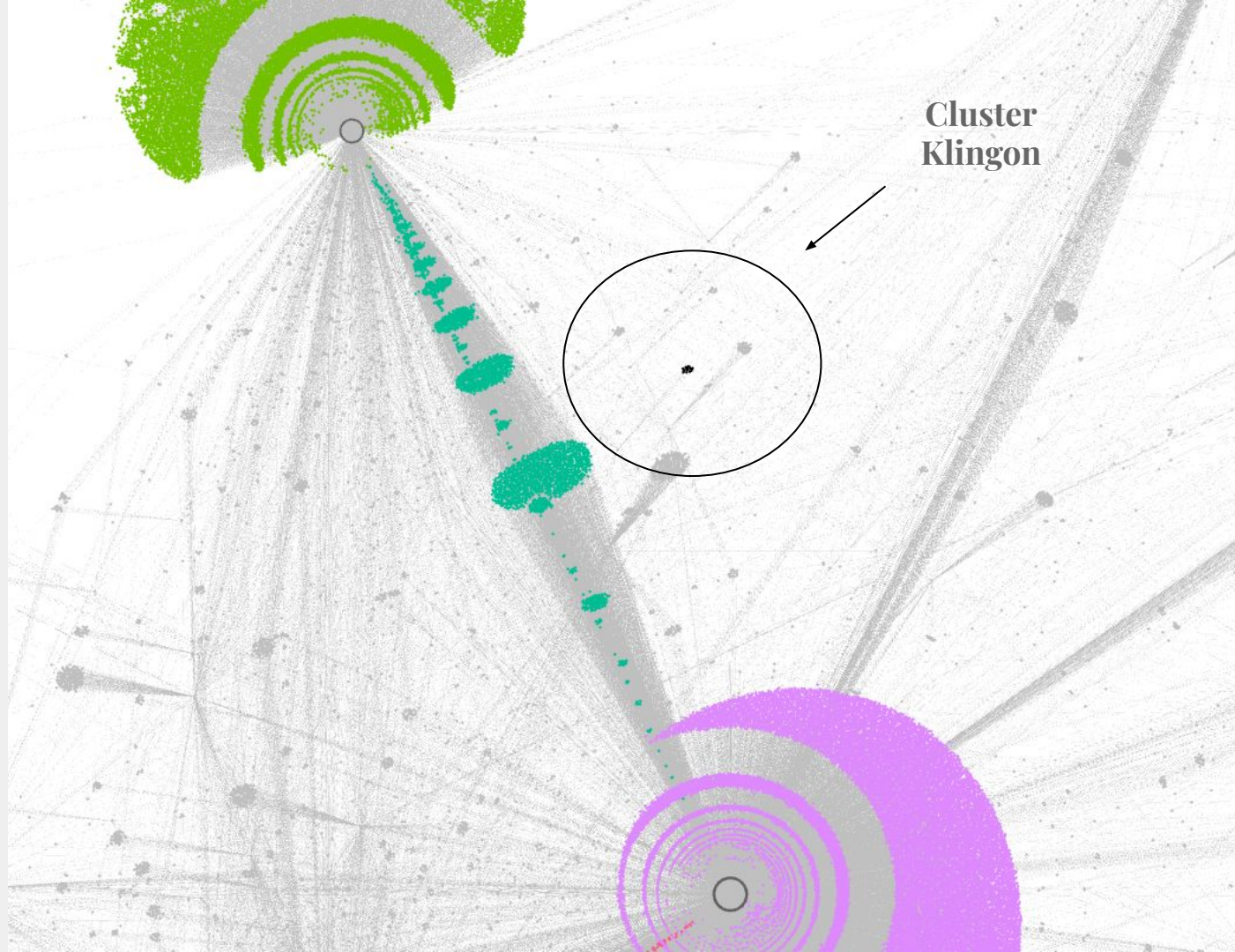
Essendo pochi utenti li ho controllati uno a uno.

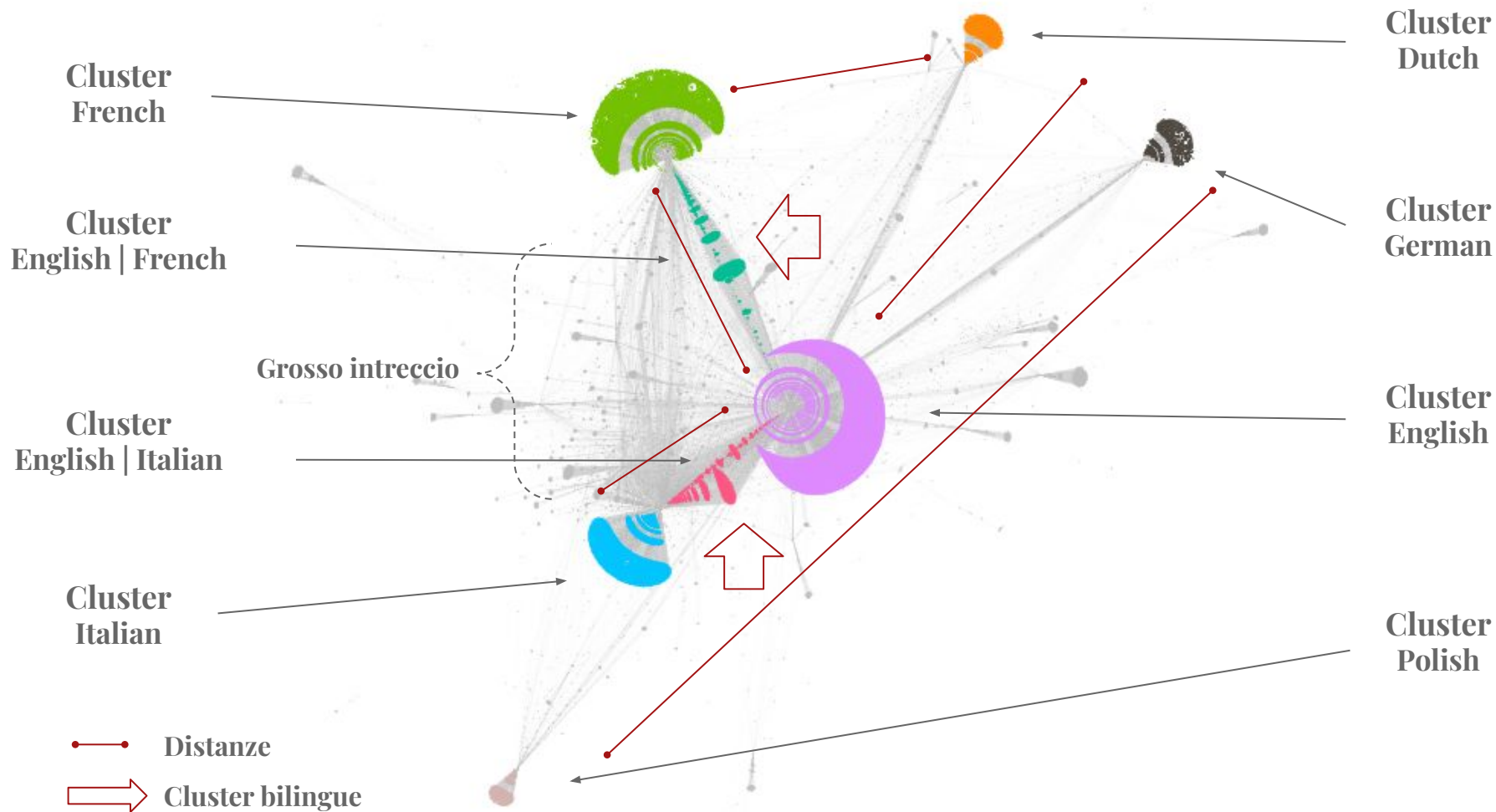
Speravo di trovare gli Sheldon Cooper e i Leonard Hofstadter di turno.

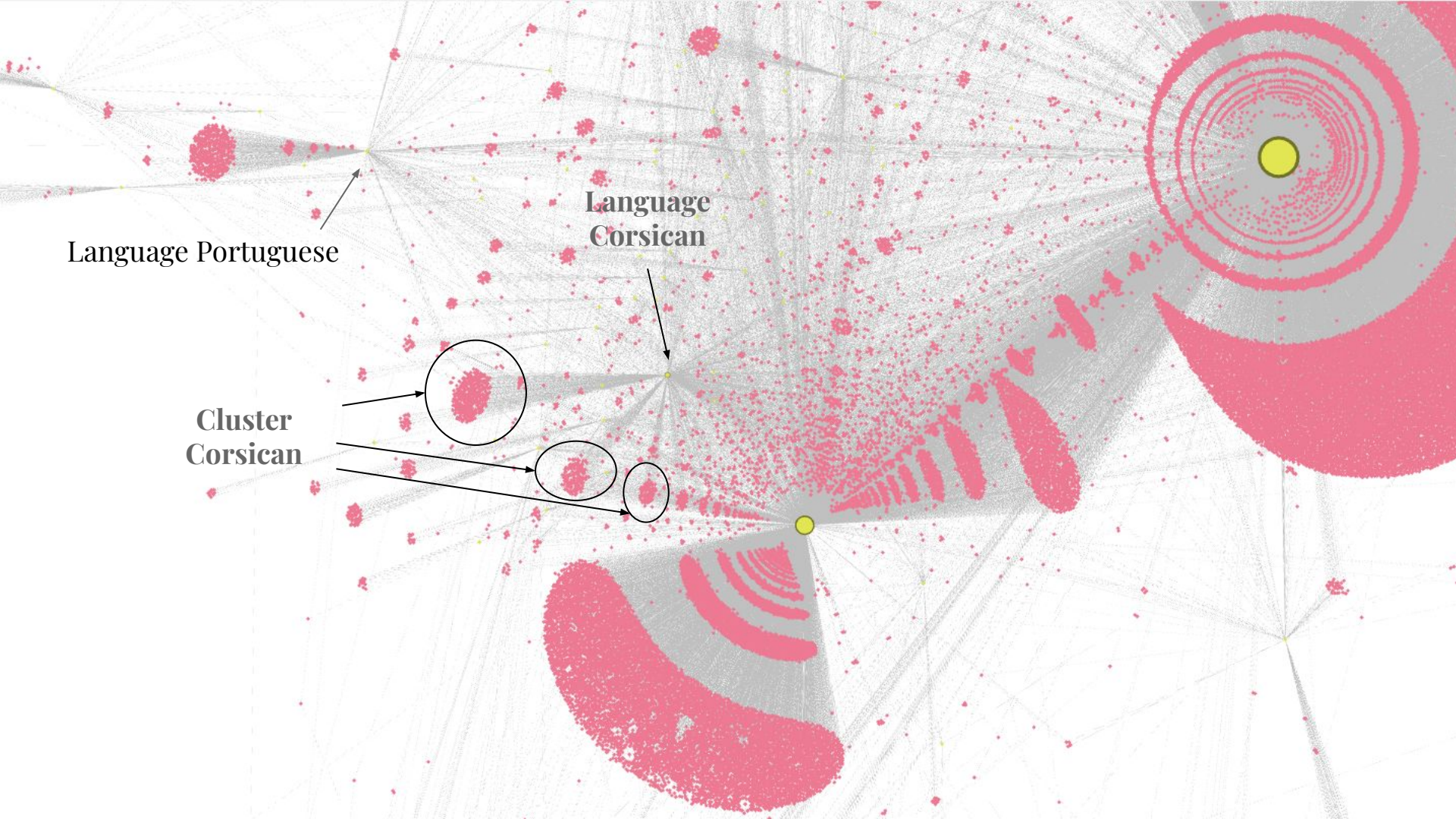
E invece...

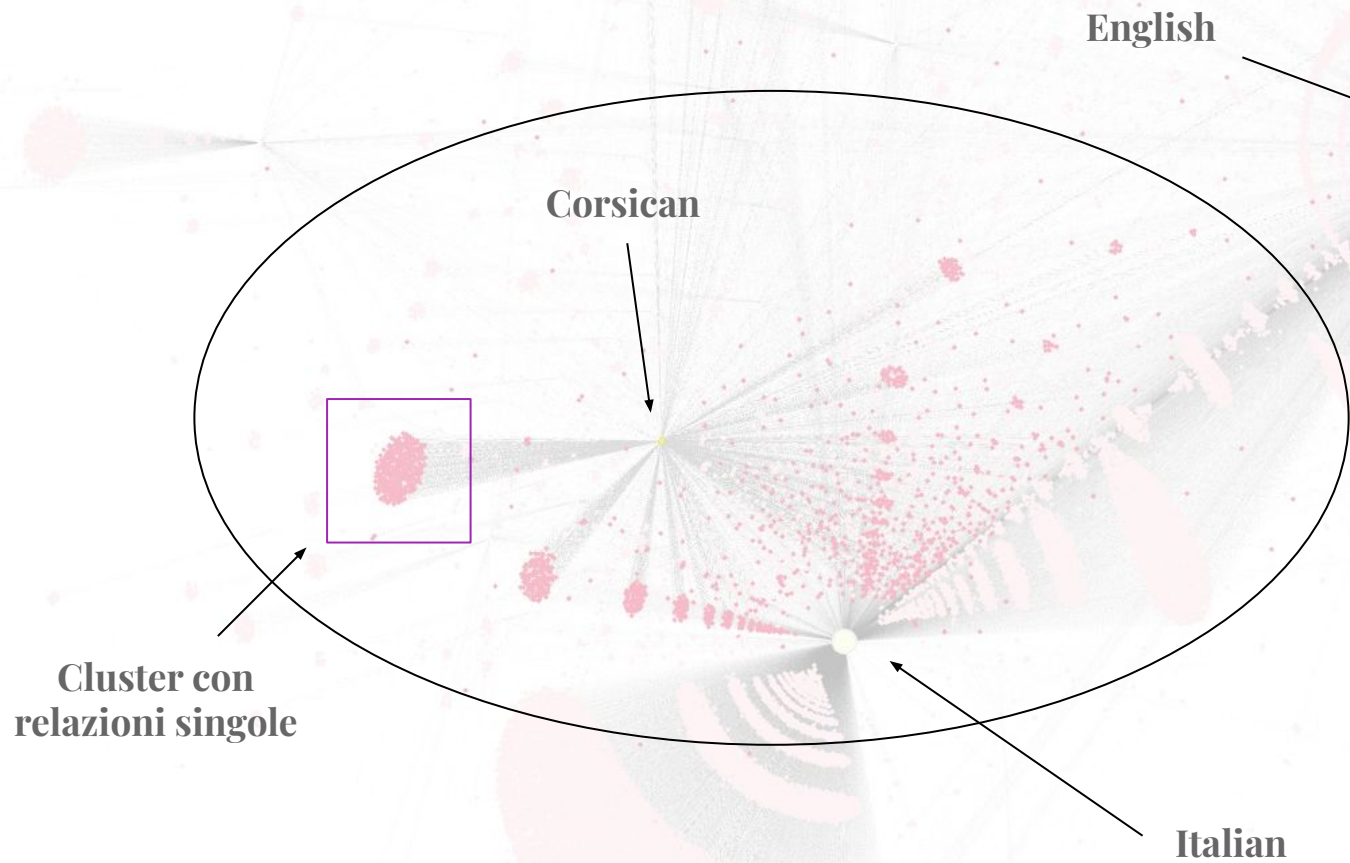
Gente che ha tweettato così male che Polyglot ha pensato fosse Klingon.

Resta da chiedersi perchè chi ha scritto la libreria Polyglot abbia inserito il Klingon.









game over

Github

https://github.com/AntonelloManenti/challenge_master_2021

Contact

antonellomanenti@gmail.com
[linkedin.com/antonello-manenti](https://www.linkedin.com/antonello-manenti)
github.com/AntonelloManenti

Colab

https://drive.google.com/file/d/1fU-xJSgHgQ_OxWRDIzEZprhot060XM6n/view?usp=sharing

Dataset (raw - 213.6 mb)

https://drive.google.com/file/d/1-CbMKWnTCTPvp8FWcZMqot_oGQ8i1hgq/view?usp=sharing

Gephi file

<https://drive.google.com/file/d/1NUWhXWUNbGx6oNlsYfxyomIBp3dUuXid/view?usp=sharing>