



Progetto 5 (Spark + Scala)

Colantonio Viviana 224473

Costa Cristian Giuseppe 227507

II dataset

```
{
  "server": "173",
  "notes": [],
  "publicFlag": true,
  "placeId": "FphPyURWU7ux6h4",
  "description": "St Peter's church in Rome",
  "secret": "b96b1fece0",
  "originalFormat": "jpg",
  "media": "photo",
  "title": "Basilica di San Pietro",

  "iconServer": "",
  "urls": [],
  "farm": "1",
  "id": "430793876",
  "hasPeople": false,
  "datePosted": "Mar 22, 2007 11:58:08 PM",
  "views": 60,
  "originalSecret": "",
```

```

  "owner": {
    "photosCount": 0,
    "admin": false,
    "revFamily": false,
    "pro": false,
    "iconServer": 0,
    "iconFarm": 0,
    "revContact": false,
    "filesizeMax": 0,
    "bandwidthUsed": 0,
    "bandwidthMax": 0,
    "id": "91071733@N00",
    "revFriend": false,
    "username": "swashford"
  },
  "comments": 0,
  "originalHeight": 0,
  "familyFlag": false,
  "rotation": -1,
  "mediaStatus": "ready",
  "geoData": {
    "latitude": 41.90245,
    "accuracy": 16,
    "longitude": 12.456661
  },
  "friendFlag": false,
  "url": "https://flickr.com/photos/91071733@N00/430793876",
  "originalWidth": 0,
```

```

  "originalWidth": 0,
  "tags": [
    {
      "count": 0,
      "value": "holiday"
    },
    {
      "count": 0,
      "value": "vatican"
    },
    {
      "count": 0,
      "value": "stpeters"
    },
    {
      "count": 0,
      "value": "rome"
    }
  ],
  "license": "",
  "iconFarm": "",
  "lastUpdate": "Dec 10, 2014 1:09:09 AM",
  "favorite": false,
  "dateTaken": "Jan 1, 0001 12:00:00 AM",
  "primary": false,
  "pathAlias": ""
```

Sommario delle analisi qualitative effettuate

analisi relative
all'**utilizzo del
social** da parte
degli utenti

analisi relative
all'**utilizzo dei tag**
all'interno dei post

analisi di **trajectory
mining** sulla base
dei geotag dei post

analisi di
clusterizzazione
degli utenti

analisi che
sfruttano concetti
di **machine
learning**

Analisi di utilizzo del social
da parte degli utenti

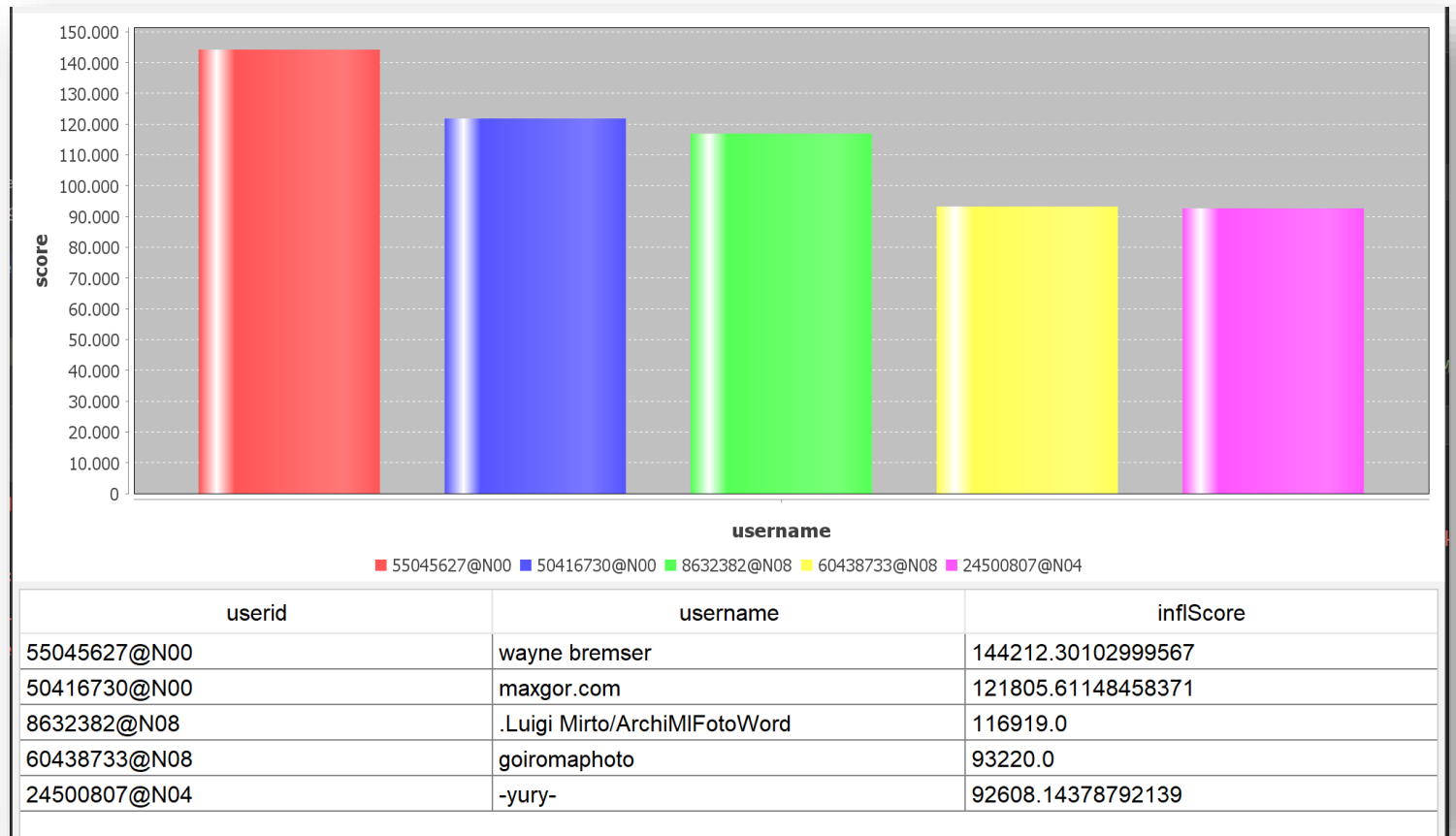


Utenti più influenti

Classifica sulla base di:

- Numero di post pubblicati
- Totale di visualizzazioni ottenute

$$score = \frac{views}{count} + \log(count)$$



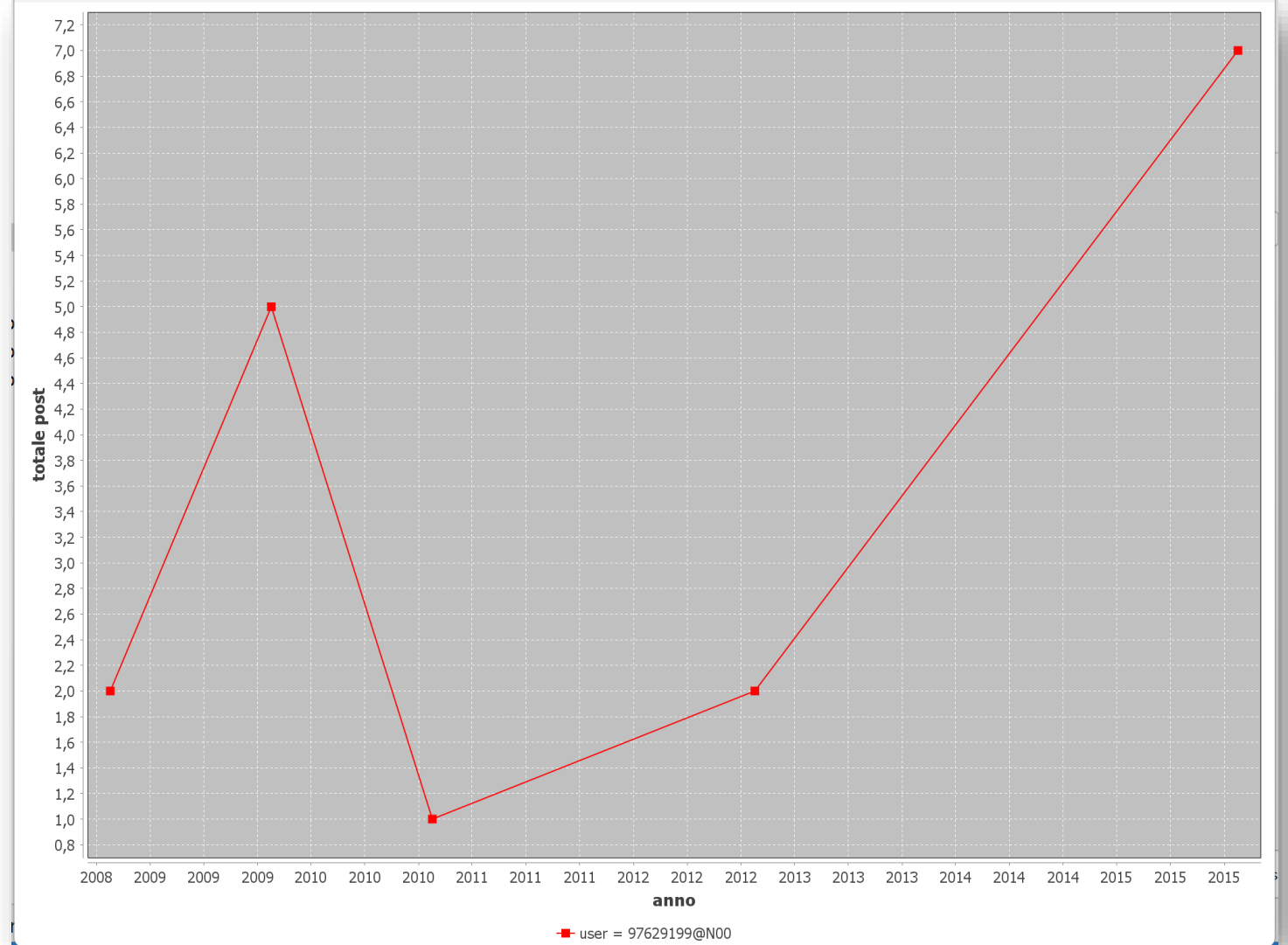
Andamento temporale delle pubblicazioni di un utente

Due parametri:

- Visualizzazioni
- Numero di post

Serie temporali realizzate per:

- Anno
- Mese
- Giorno

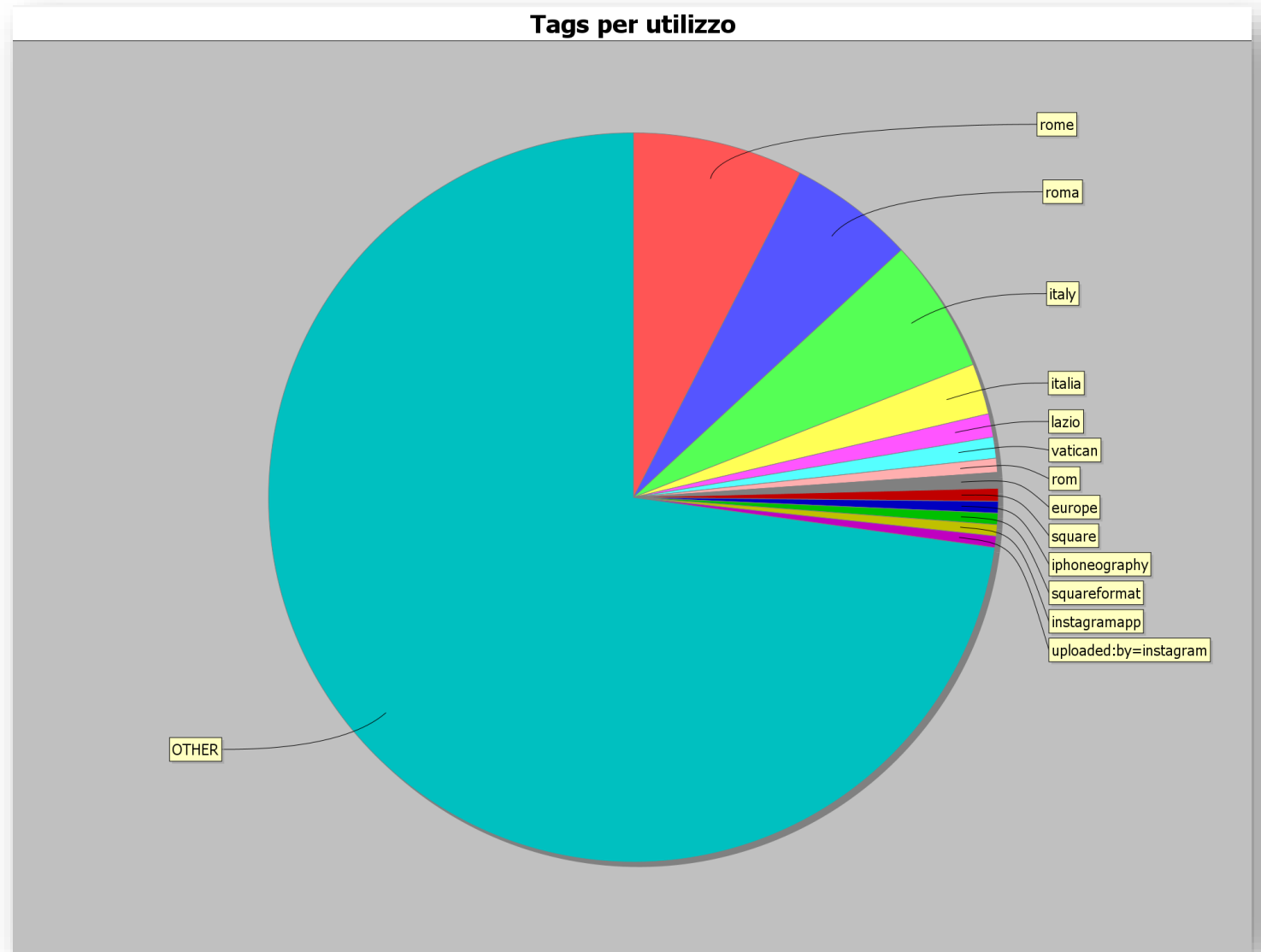


Analisi riguardanti l'utilizzo dei tag



Tag maggiormente utilizzati

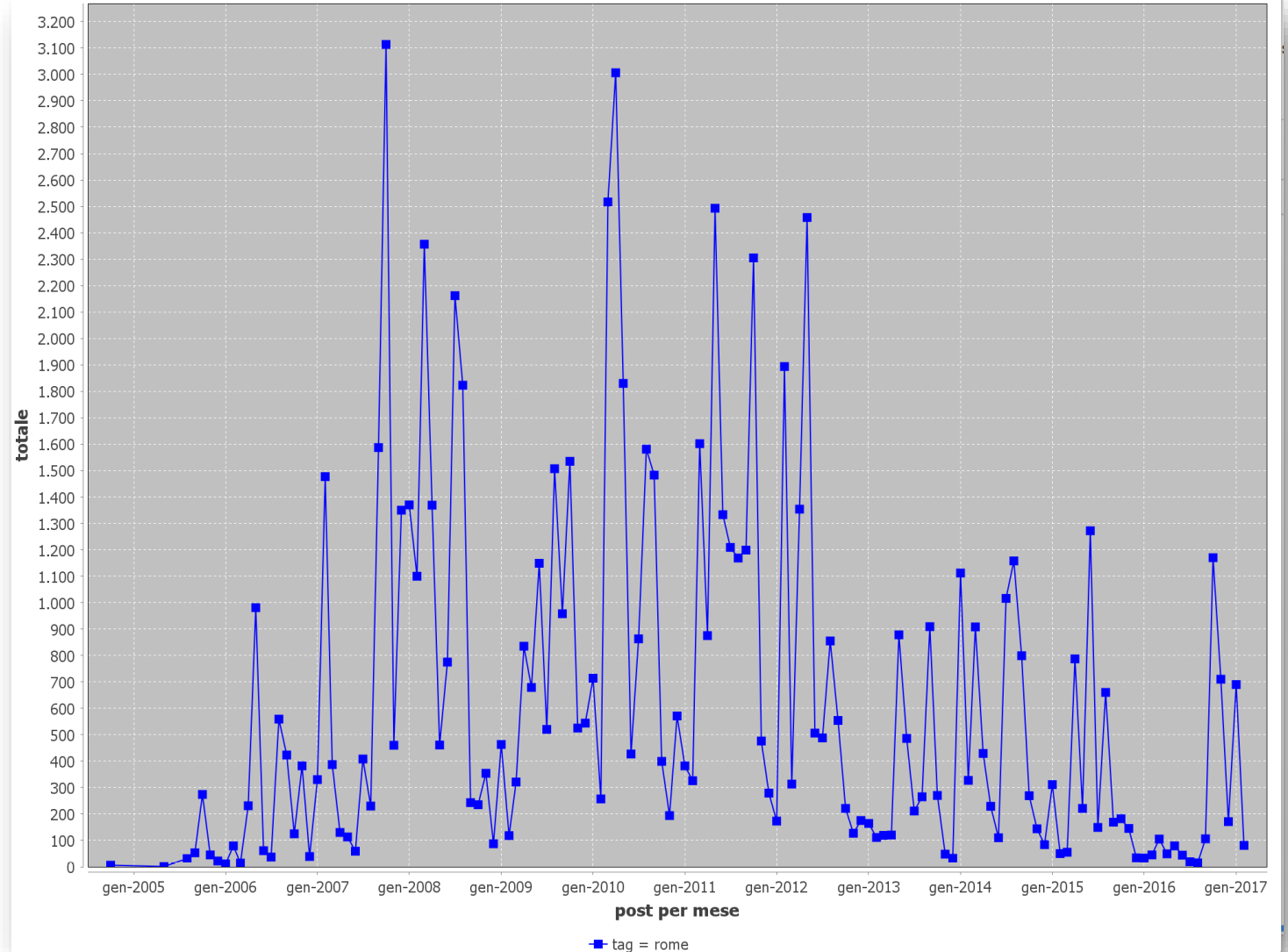
- Si è studiata la distribuzione dei tag sulla base del numero dei post ad essi relativi ed il numero di visualizzazioni totalizzate.

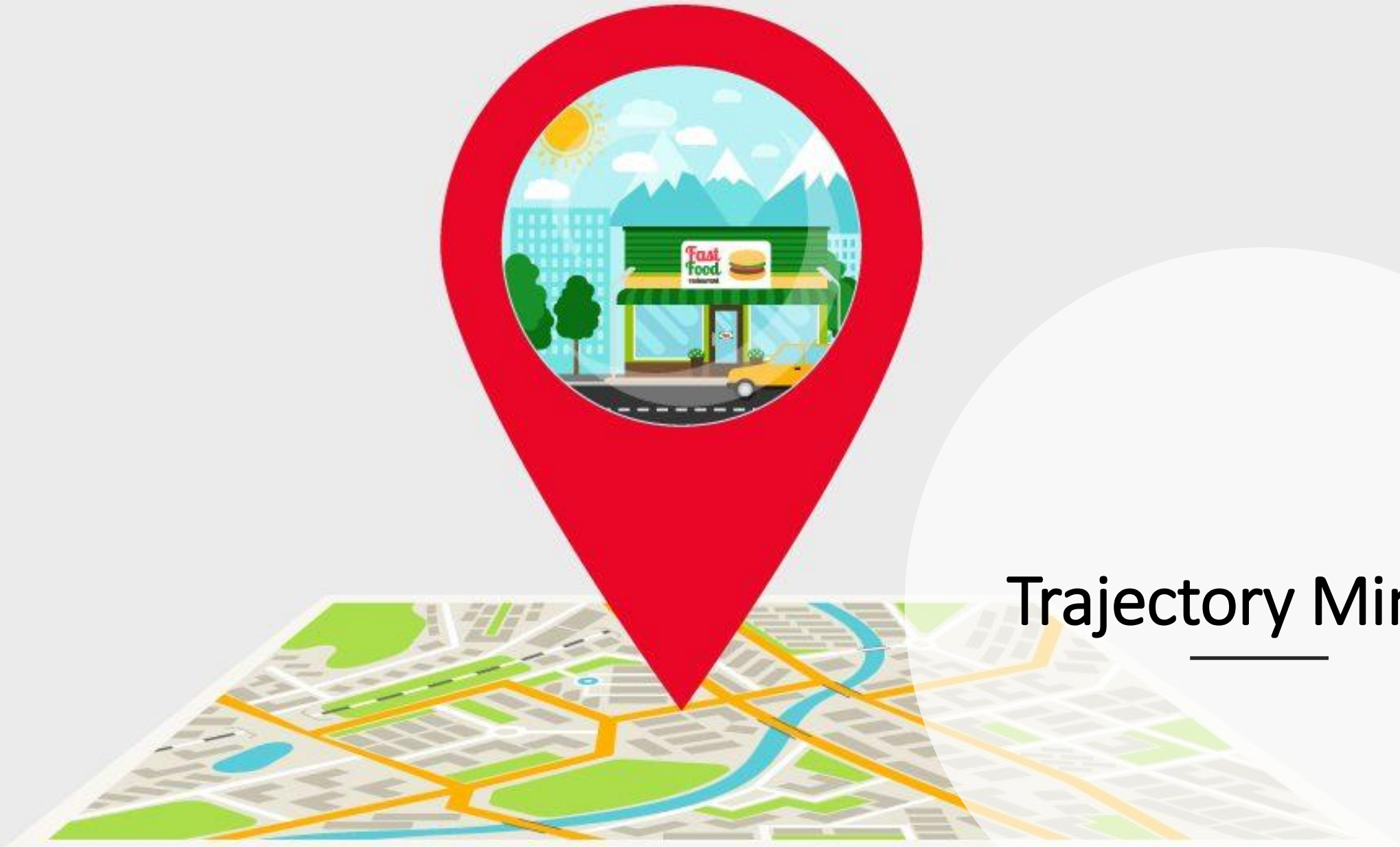


Andamento temporale dell'utilizzo di un tag

Serie temporali realizzate
per:

- Anno
- Mese
- Giorno





Trajectory Mining

Approccio adottato



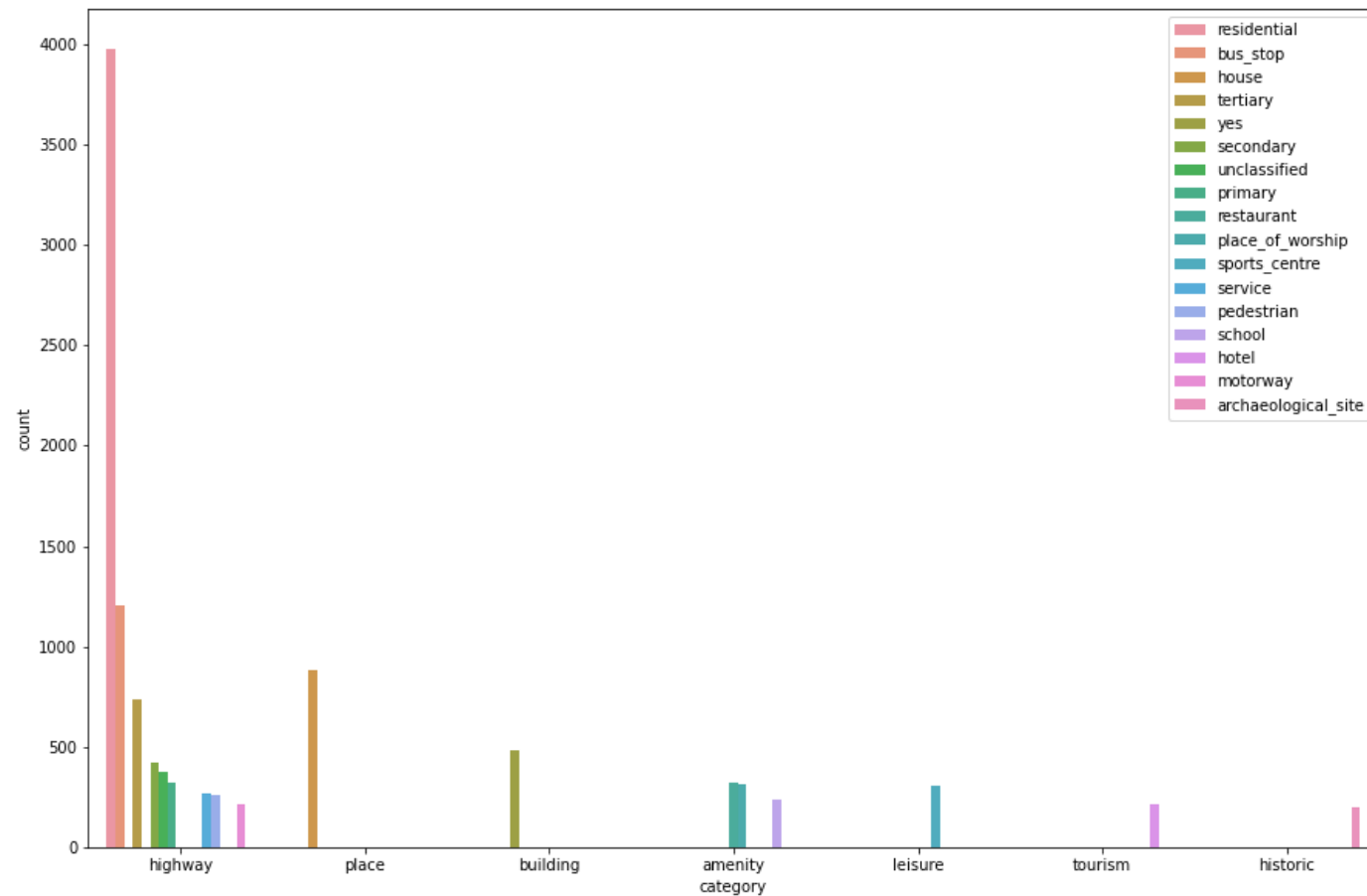
Task 1: post credibili

Si eliminano dal dataset tutti i post per i quali

- la data di acquisizione della foto è mancante
- la data di acquisizione della foto è antecedente il 1/1/2004
- la data di acquisizione della foto è conseguente al 1/1/2020.

Task 2: attribuzione di un luogo ad un post

1. si individuano le coppie (*latitudine* , *longitudine*) presenti nel dataset;
2. si effettua un rounding a 3 cifre delle coppie;
3. si filtrano le sole coppie distinte;
4. si effettua una richiesta alle API di **OpenStreet-Map** per tutte le coppie al fine di ottenere info sul luogo;
5. gli oggetti **GeoDFItem** ottenuti al passo (4) saranno raccolti in un dataset, filtrato sulla base di analisi delle distribuzioni degli attributi delle sue tuple;
6. ad ogni post presente nel dataset, verranno assegnati tutti i luoghi così trovati (mediante join) filtrando solo quelli che si trovano nel raggio di 300 metri dalla posizione dello scatto fotografico.



Task 3: Dataset per Prefix Span

Examples

Scala

Java

Python

R

Refer to the [Scala API docs](#) for more details.

```
import org.apache.spark.ml.fpm.PrefixSpan

val smallTestData = Seq(
  Seq(Seq(1, 2), Seq(3)),
  Seq(Seq(1), Seq(3, 2), Seq(1, 2)),
  Seq(Seq(1, 2), Seq(5)),
  Seq(Seq(6)))

val df = smallTestData.toDF("sequence")
val result = new PrefixSpan()
  .setMinSupport(0.5)
  .setMaxPatternLength(5)
  .setMaxLocalProjDBSize(32000000)
  .findFrequentSequentialPatterns(df)
  .show()
```

<https://spark.apache.org/docs/latest/ml-frequent-pattern-mining.html#prefixspan>

Task 3:

Dataset per Prefix Span

```
private def user_loc_seq(dataset: Dataset[FlickrPost], datasetGeo: Dataset[GeoDFItem]): DataFrame = {  
    val res = best_loc_guess_all_data(dataset, datasetGeo).repartition(200)  
    {...}  
  
    .toDF("DATE", "ID_OWNER", "SEQUENCE")  
  
    all_sequences.write.json("sequences_json")  
  
    all_sequences  
  
def seq_for_prefix_span(sequences: DataFrame): DataFrame = {  
    val all_sequences_df =  
        sequences  
        .map(x => x.get(2).asInstanceOf[Seq[String]].map(y => Seq(y)))  
        .toDF("sequence")  
  
def frequent_seq_pat(dataframe: DataFrame): DataFrame = {  
    val result = new PrefixSpan()  
        .setMinSupport(0.02)  
        .setMaxPatternLength(10)  
        .setMaxLocalProjDBSize(32000000)  
        .findFrequentSequentialPatterns(dataframe)  
  
    {...}  
}
```


Task 4:

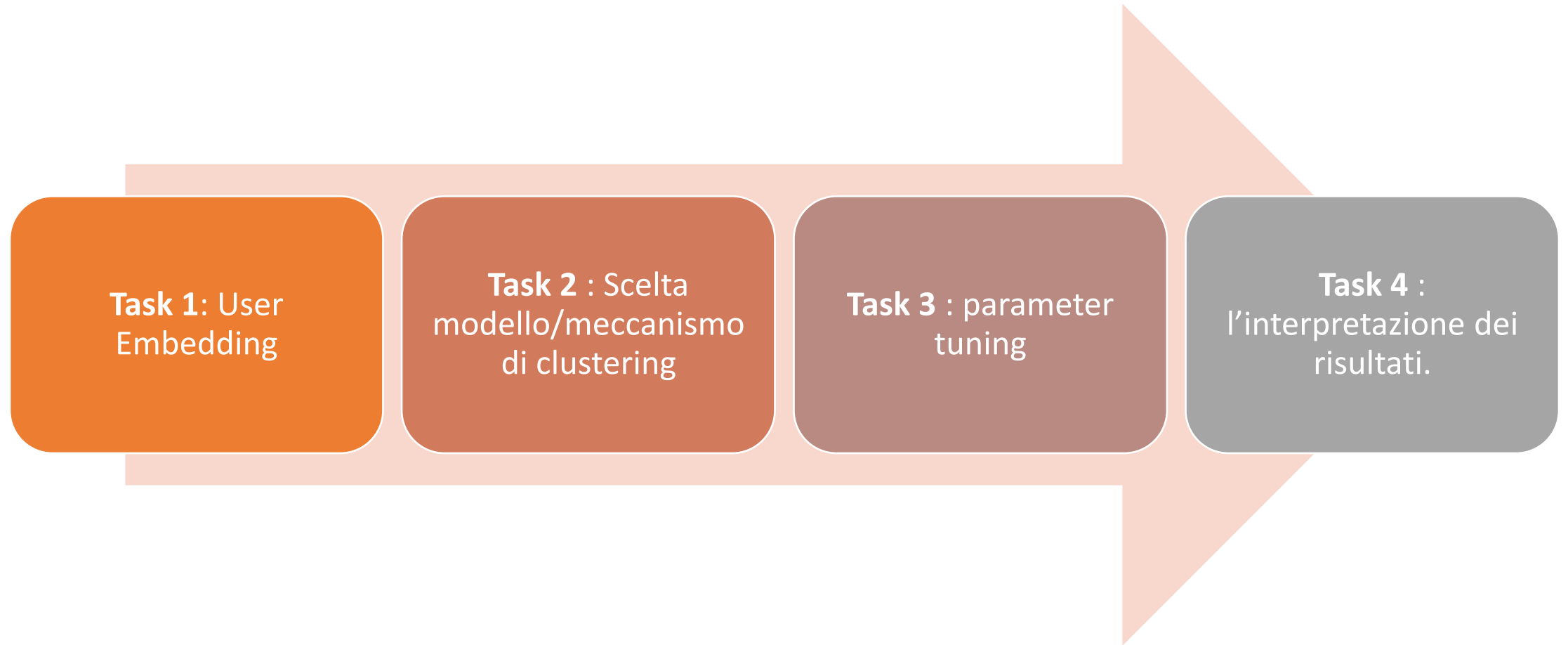
Interpretazione dei risultati

Pattern	Supporto
Piazza del Campidoglio → Foro di Traiano	11.60%
Colosseo → Domus Aurea	10.22%
Colosseo → Ludus Magnus	9.59%
Basilica Sancti Petri → Braccio di Carlo Magno	10.37%
Forum Romanum → Tempio del Divo Giulio	7.52%
Musei Vaticani → San Nilammone Forum Sancti Petri → Braccio di Carlo Magno	4.60%
Colosseo → Santi Luca e Martina al Foro Romano → Tempio della Concordia	4.02%
Basilica Sancti Petri → Palazzo del Sant'Uffizio → Braccio di Carlo Magno	10.04%
Forum Romanum → Tempio del Divo Giulio → Foro di Nerva	7.31%
Casa delle Vestali → Forum Romanum → Foro di Nerva	6.51%
Colosseo → Ludus Magnus → Antiquarium del Celio → Domus Aurea	6.45%
Casa delle Vestali → Forum Romanum → Tempio del Divo Giulio → Necropoli arcaica	6.98%
Piazza Venezia → Santa Maria in Aracoeli → Insula dell'Ara Coeli → Foro di Traiano	7.03%
Obeliscus Vaticanus → Basilica Sancti Petri → Palazzo del Sant'Uffizio → Braccio di Carlo Magno	8.33%
Tipografia Vaticana → Basilica Sancti Petri → San Nilammone → Braccio di Carlo Magno	7.69%



User Clustering

Approccio adottato



Task 1: User Embedding

- Scelta feature dai post (*title + description*)
- Poi, diversi approcci al problema:
 - Modelli preallenati (su cosa? Quali lingue?)
 - Es: Doc2Vec
 - Modelli da addestrare (come? Parameter tuning)
 - Es: LDA

Si è optato per *Multilingual Universal Sentence Encoder*

Task 1: User Embedding

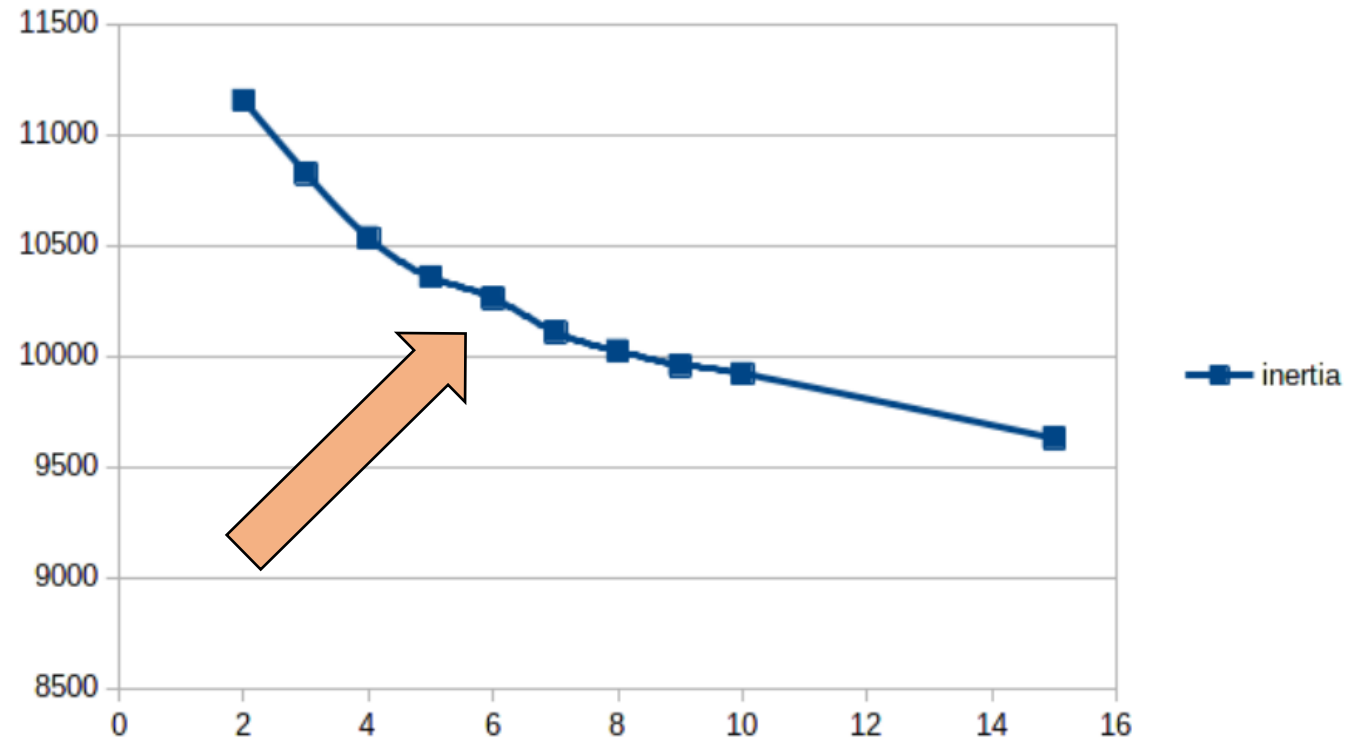
L'embedding di un singolo utente sarà dato da:

$$embedding(user) = \frac{\sum_{x \in posts(user)} embedding(x)}{|posts(user)|}$$

Lo spazio latente (\mathbb{R}^{512}) è condiviso tra post e utenti

Task 2, 3:
Scelta
meccanismo di
clustering e
parameter
tuning

- Serve un modello con ragionevole *explainability*
- Si è optato per KMeans, $k = 6$



Task 4:

Interpretazione dei risultati

- Cluster maggiormente
numeroso

Utente	Parole più usate
98274023@N00	vatican, st, rome, museum, i, pantheon, hotel, peter, paul, new
32076237@N00	rome, the, san, basilica, Pietro, saint, peter, vatican, piazza, colosseum
44192643@N02	peter, st, vatican, square, santa, rome, maria, dome, inside, colosseum
83031170@N00	the, rome, st, peter, inside, vatican, forum, basilica, fountain, coliseum
22094057@N05	the, peter, st, vatican, coliseum, basilica, fountain, trevi, museum, forum

Cluster 0

Task 4: Interpretazione dei risultati

- Turisti abbastanza descrittivi all'interno dei loro post

Utente	Parole più usate
55391611@N00	rome, the, basilica, vatican, city, peter, saint, santa, maria, church
22158962@N07	roma, san, org, santa, wikipedia, http, href, wiki, www, piazza
28353725@N00	rome, i, the, piazza, street, quot, church, it, one, nuns
77547214@N00	the, rome, fountain, roman, forum, temple, colosseum, quot, st, peter
25718393@N04	quot, the, rome, roma, wikipedia, href, com, see, http, flickr

Cluster 1

Task 4: Interpretazione dei risultati

- Utenti che pubblicano post con pensieri ed aforismi personali

Utente	Parole più usate
34857532@N00	en, wiki, org, wikipedia, rome, http, href, i, villa
24793644@N08	crunch, i, roma, dsc, jpg, foto, storico, 5, b,
63327992@N07	i, a, the, rome, old, nice, our, fountain, water, maxentius
19446102@N00	rome, quot, href, http, i, roma, rel, nofollow, via, piazza
33399095@N00	roma, www, com, href, http, flickr, quot, photos, e, mm

Cluster 2

Task 4: Interpretazione dei risultati

- Utenti che inseriscono URL verso altri siti, che ottengono molte visualizzazioni. Sfruttano Flickr per pubblicizzarsi.

Utente	Parole più usate
8099187@N06	com, omogirando, href, rel, nofollow, www, http, jimdo, b, facebook
69912818@N00	href, http, rel, nofollow, www, com, b, roma, sound36
11432907@N00	href, http, com, rel, nofollow, large, amp, bighugelabs, onblack, php
11102419@N00	href, http, org, wikipedia, wiki, en, rome, com, flickr, the
21336230@N08	href, http, rel, nofollow, com, roma, www, amp, view, large

Cluster 3

Task 4: Interpretazione dei risultati

- Utenti che utilizzano titolo e descrizione del post in maniera poco significativa

Utente	Parole più usate
29223649@N04	de, roma, piazza, san, via, en, n, 2, y, palazzo
96291012@N00	img, com, www, bertolinidennis, myspace, 20110602, http, href, 06, edited
47211255@N05	img, a, s, day, piazza, navona, roma, lina, ivo, not
58826214@N00	column, trajan, roman, rome, aurelius, marcus, it, built, the, spiral
25538307@N00	dsc, img, 8, marzo, eucalipti, minirugby, ios, u16, s, photos

Cluster 4

Task 4: Interpretazione dei risultati

- Utenti appassionati di fotografia che inseriscono descrizioni sulle macchine fotografiche utilizzate, obiettivi, ecc.

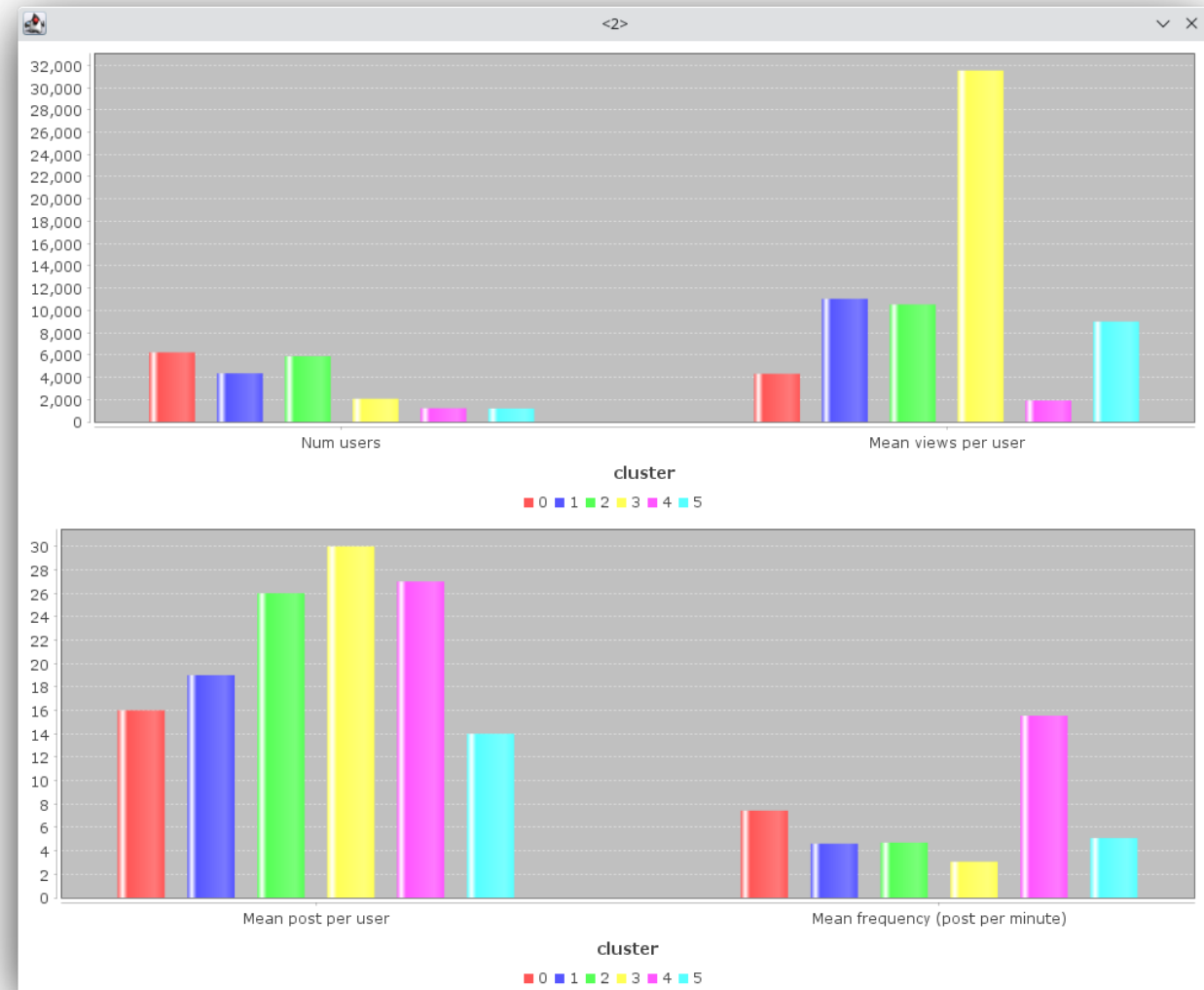
Utente	Parole più usate
63558118@N07	voigtlander, 5, f, roma, 20mm, la, d200, valeria, rome, steps
136373368@N02	rome, roma, gh4, panasonic, picture, villa, autumn, borghese, shooted
13958243@N08	ilford, rome, nikon, epson, v750, kodak, nikkor, quot, tmax, d76
33920763@N08	de, ce, societ, 1, nikon, n, embe, famo, vino, statue
27818145@N00	http, href, kodak, com, 2, f, planar, 80mm, hasselblad, 501cm

Cluster 5

Task

4: Interpretazione dei risultati

In figura a fianco, statistiche relative ai vari cluster

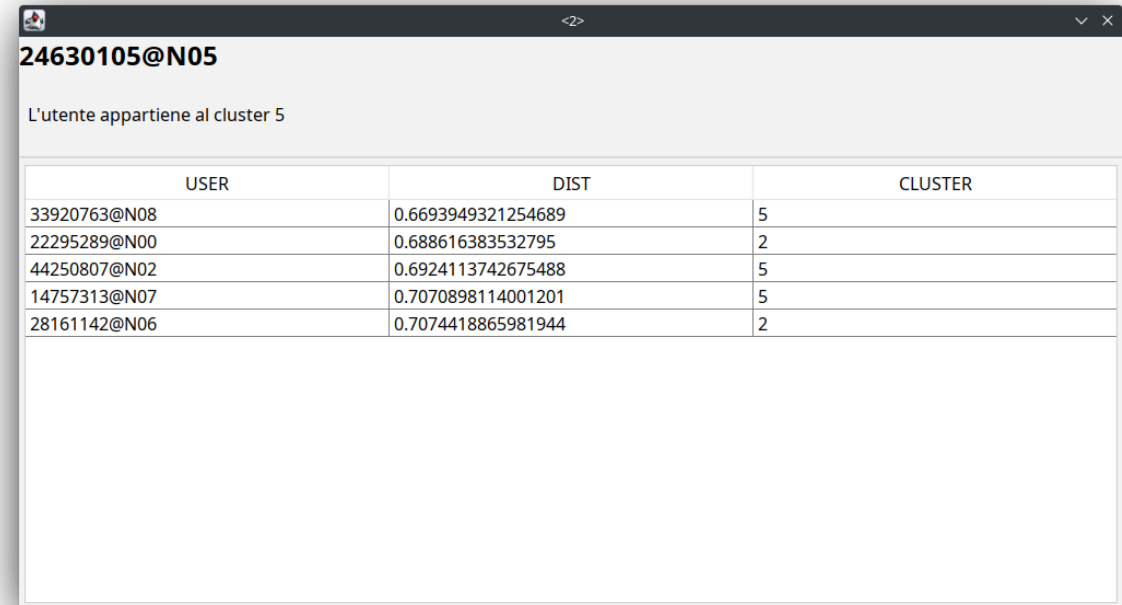


Indicazioni di longevità dell'utente

Task 4: Interpretazione dei risultati

In figura, utenti simili ad un dato utente.

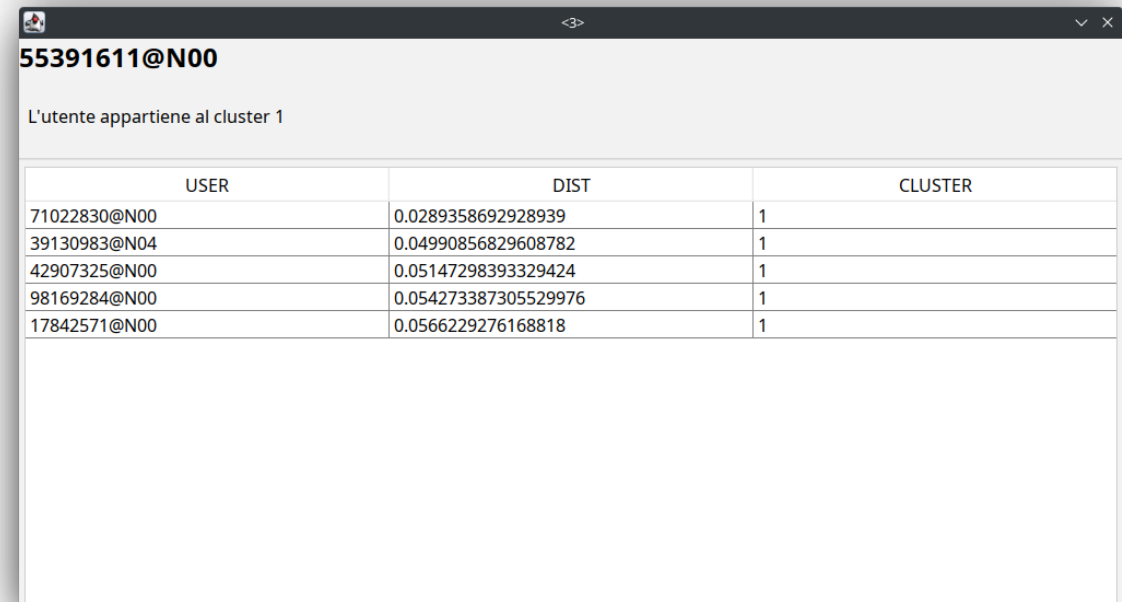
La dimensionalità rende molto meno efficace la distanza euclidea.



24630105@N05

L'utente appartiene al cluster 5

USER	DIST	CLUSTER
33920763@N08	0.6693949321254689	5
22295289@N00	0.688616383532795	2
44250807@N02	0.6924113742675488	5
14757313@N07	0.7070898114001201	5
28161142@N06	0.7074418865981944	2



55391611@N00

L'utente appartiene al cluster 1

USER	DIST	CLUSTER
71022830@N00	0.0289358692928939	1
39130983@N04	0.04990856829608782	1
42907325@N00	0.05147298393329424	1
98169284@N00	0.054273387305529976	1
17842571@N00	0.0566229276168818	1

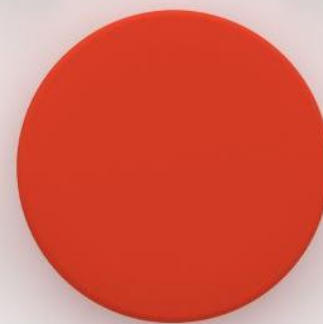
Query di Machine Learning



Obiettivi

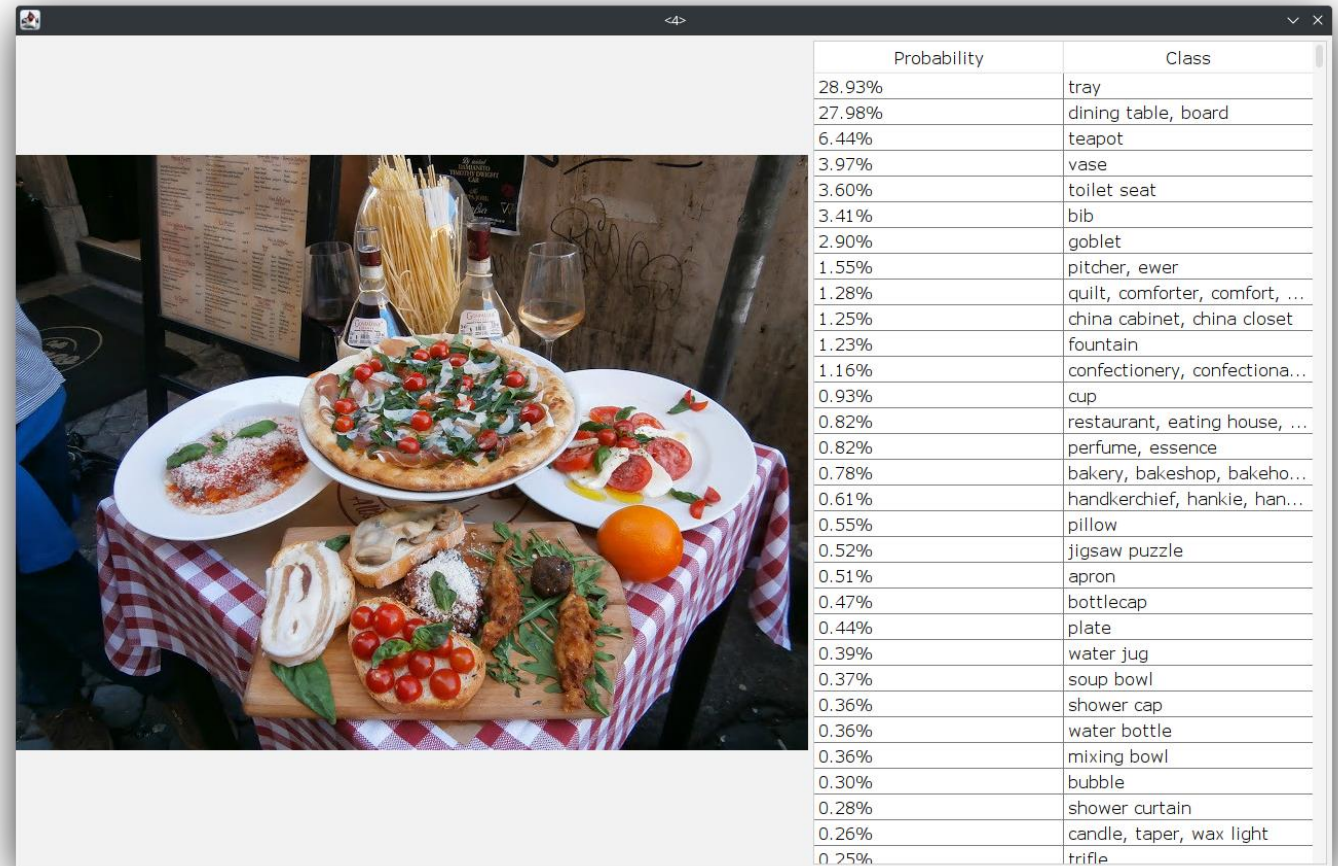
Sono state implementate query per:

- Ottenere feature da immagini
- Ricavare la lingua di un post



Query per immagini

- In figura, è possibile osservare l'output della rete GoogleNet.
- La rete, preallenata, è stata utilizzata tramite la libreria SynapseML di Microsoft, la quale supporta il formato ONNX.

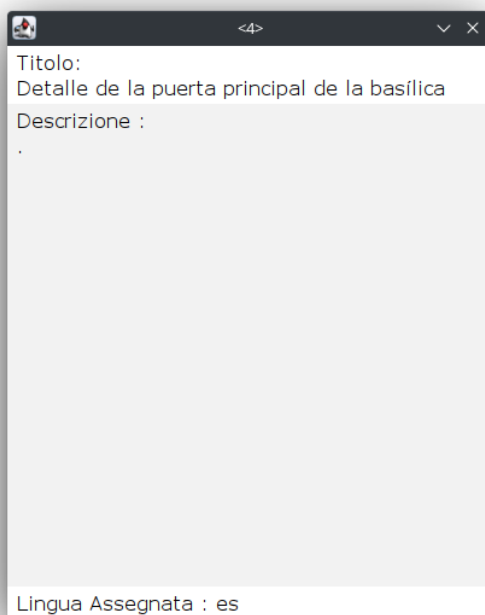


The image shows a software window displaying a photograph of a table set with food, including a pizza, bread, and a drink. To the right of the image is a table listing the detected objects and their probabilities.

Probability	Class
28.93%	tray
27.98%	dining table, board
6.44%	teapot
3.97%	vase
3.60%	toilet seat
3.41%	bib
2.90%	goblet
1.55%	pitcher, ewer
1.28%	quilt, comforter, comfort, ...
1.25%	china cabinet, china closet
1.23%	fountain
1.16%	confectionery, confectiona...
0.93%	cup
0.82%	restaurant, eating house, ...
0.82%	perfume, essence
0.78%	bakery, bakeshop, bakeho...
0.61%	handkerchief, hankie, han...
0.55%	pillow
0.52%	jigsaw puzzle
0.51%	apron
0.47%	bottlecap
0.44%	plate
0.39%	water jug
0.37%	soup bowl
0.36%	shower cap
0.36%	water bottle
0.36%	mixing bowl
0.30%	bubble
0.28%	shower curtain
0.26%	candle, taper, wax light
0.25%	trifle

Query per lingua di un post

In figura, è possibile osservare l'output del *LanguageDetectorDL*, fornito da Spark NLP, per post di lingua differente.



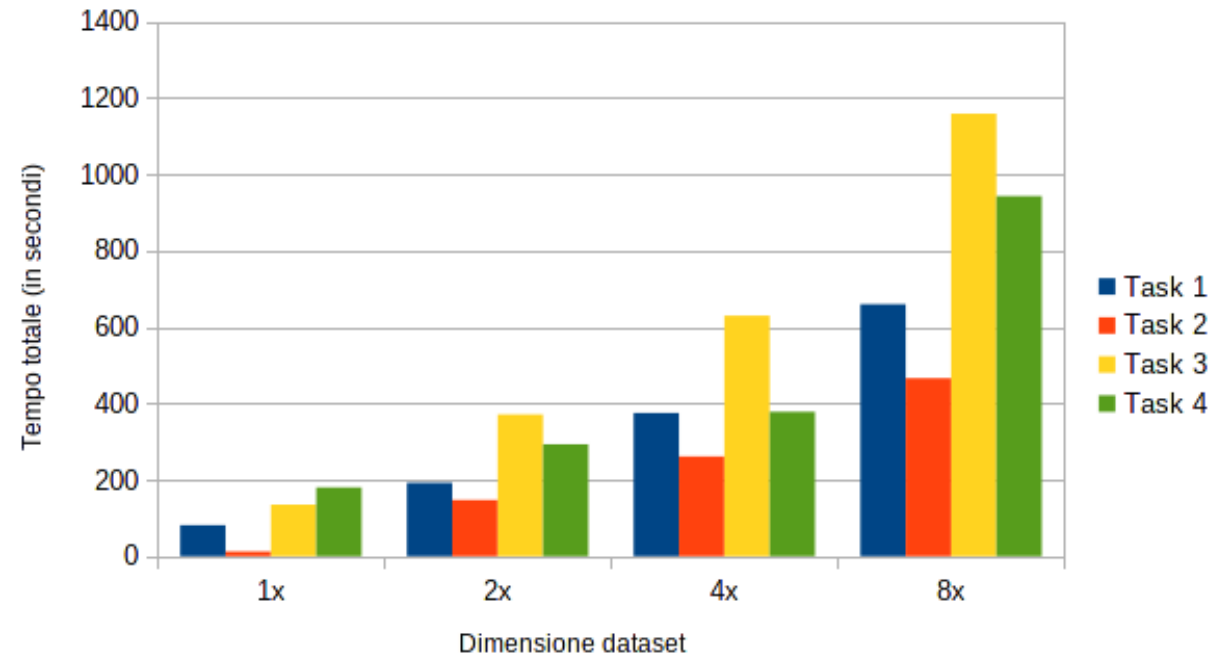
Analisi delle prestazioni



Analisi su dati sintetici

I task analizzati sono i seguenti:

- task 1: andamento dei post di un utente per anno;
- task 2: utenti più influenti sulla base di visualizzazioni e post pubblicati (score);
- task 3: andamento dei tag negli anni;
- task 4: distribuzione dei tag nel dataset;



Grazie per l'attenzione!

A thick, orange, wavy horizontal line that spans the width of the text above it, serving as a decorative underline.