HOW PEOPLE TALK ABOUT HEALTH?

Nicola Procopio 25-01-2022

Talent Garden Cosenza

AI DEMOCRATIZE

iaml
ITALIAN ASSOCIATION FOR MACHINE LEARNING

About me

Nicola Procopio

Senior Data Scientist

Contacts

nicola.procopio@healthwareinternational.com

https://it.linkedin.com/in/nicolaprocopio

https://github.com/nickprock

https://www.slideshare.net/NicolaProcopio

Actually @

healthware international

Background

integris

ICAR CNR

Affare

KDD Lab

REGIONE CALABRIA

UNIVERSITÀ DELLA CALABRIA

Community

AI DEMOCRATIZE

tag Talent Garden Cosenza

iaml ITALIAN ASSOCIATION FOR MACHINE LEARNING

GDG

SC STARTUP CALABRIA

Mission

healthware international

The full-service healthcare agency of Healthware Group

We play at the intersection of science, creativity, boundless curiosity, and our understanding of human needs. That's how we design transformational healthcare experiences that engage, simplify and empower people's lives.

We are digital natives and multi-talented coders, connected and passionate to learn and innovate.

Our mission is to design and develop successful solutions and digital products.

Summary

1. Google Flu Trends (2009)
2. Online Clustering Algorithm on Twitter Data (2016)
3. Observational Study on an Online Decision Support System Data (2021)

Google Flu Trends

- Started in 2008
- It provided estimates of influenza activity for more than 25 countries.

“Google web search queries can be used to accurately estimate influenza-like illness percentages in each of the nine public health regions of the United States” [1]

Google Flu Trends

- Google Flu Trends stopped publishing current estimates on 9 August 2015.

SCIENZA | MARTEDÌ 18 MARZO 2014

Google non ci ha preso con l'influenza

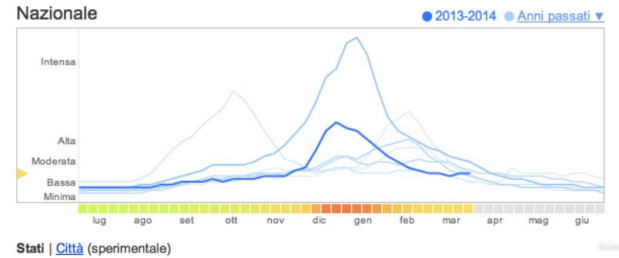
Una ricerca su "Science" critica il sistema di analisi per prevedere l'andamento dell'influenza stagionale nel mondo, sulla base delle cose cercate dagli utenti

nòva

Scienza | Tecnologia | Creatività | Social Innovation | Dossier | Blog

Scopri i trend influenzali - Stati Uniti

Abbiamo scoperto che determinati termini di ricerca sono validi indicatori dell'attività influenzale. Google Trend influenzali utilizza dati di ricerca aggregati di Google per stimare l'attività influenzale.
Ulteriori informazioni »



Stati | Città (sperimentale)

Google Flu Trends: big data senza big theory

6 aprile 2014 Cristina Cenci BIG DATA

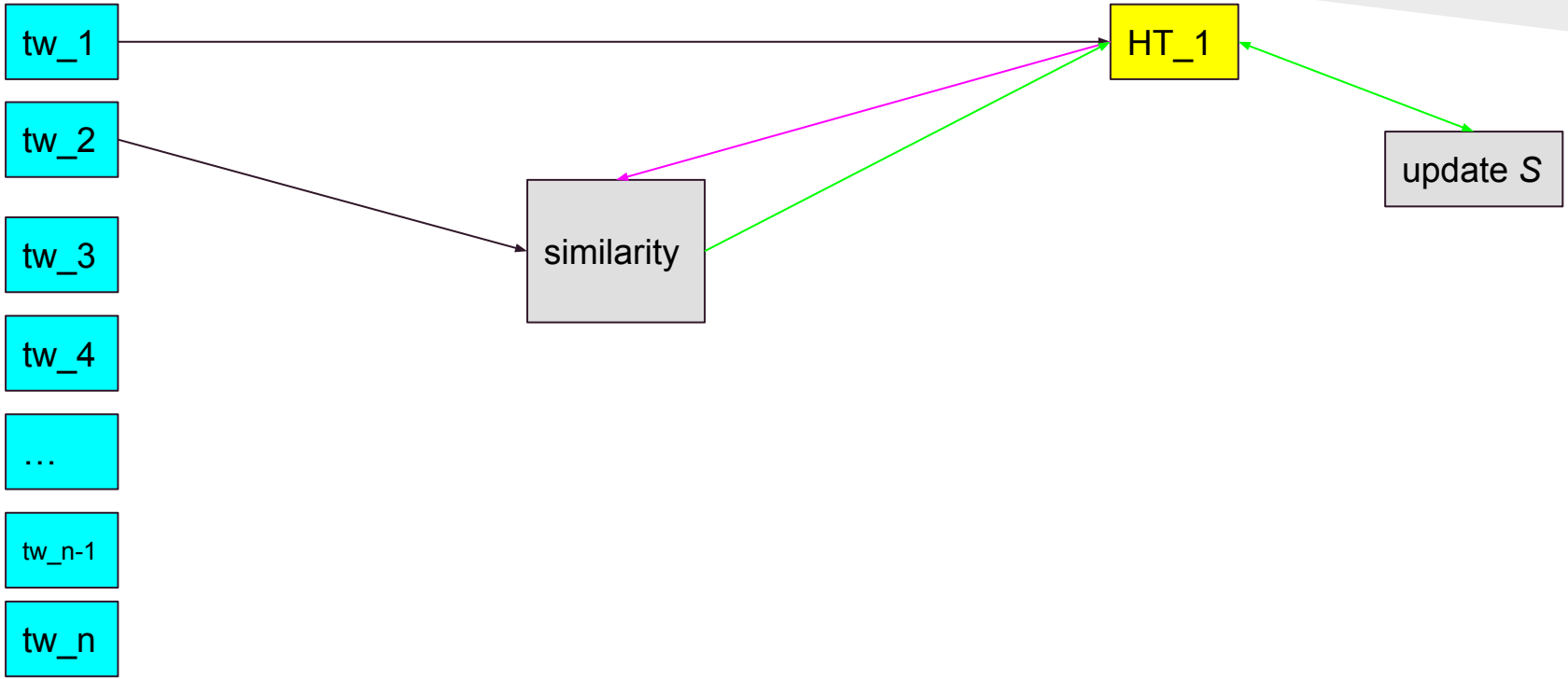
HealthS-Tweet

- *"Health Surveillance through Twitter"* is a specialized version for the healthcare domain of a previous work about Online Clustering for Topic Detection[3]
- Incremental clustering
- Able to detect low-frequency topics
- not based on word frequencies, but on similarity of the words and hashtag used
- Tweets posted in USA from September 2015 to April 2016

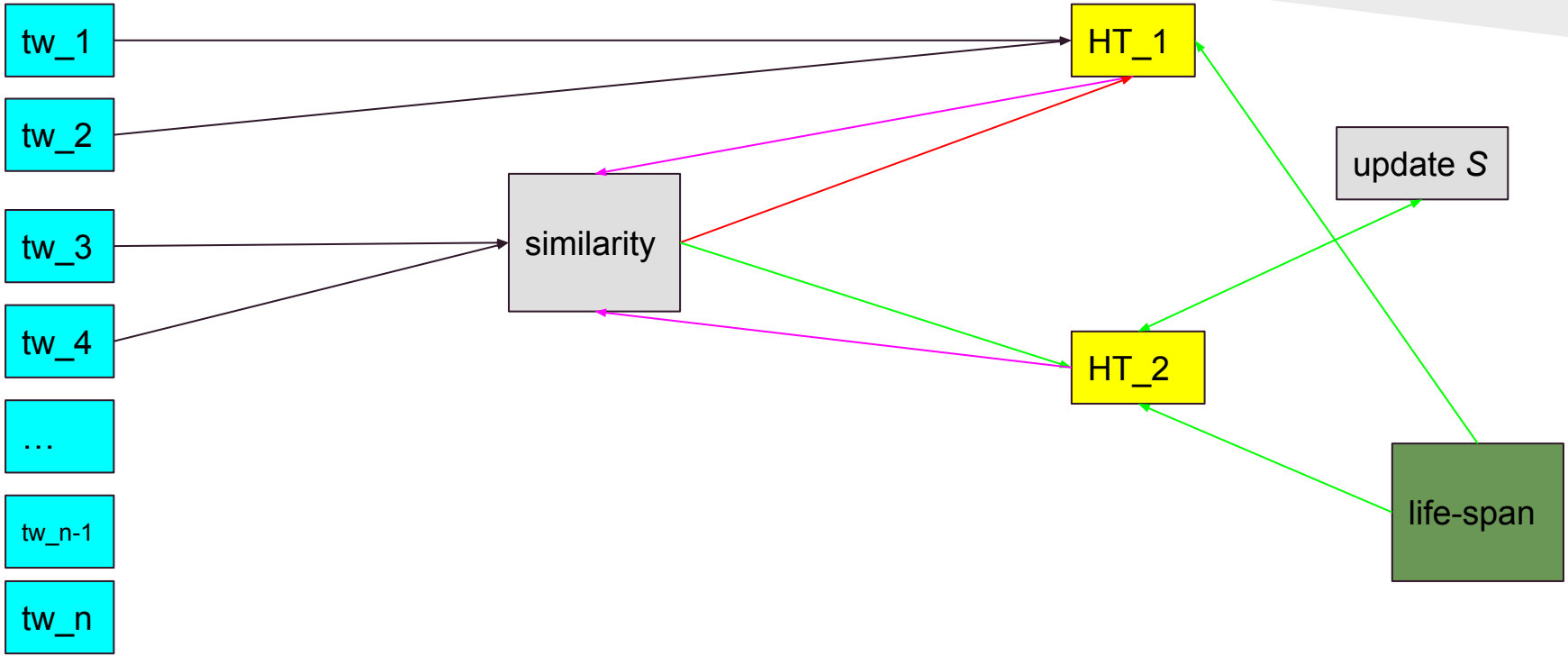
HealthS-Tweet: Data

- ***tweet*** → *tw: (id, u, l, tf, fv)*
  - *id: tweet id*
  - *u: user id*
  - *l: location [Lat.; Lon.]*
  - *tf: creation time*
  - *fv: (wu, wb, hu, hb)*
    - *wu, wb: word unigram, word bigram*
    - *hu, hb: hashtag unigram, hashtag bigram*
- ***Topic Summary*** → *S like a tweet extended with (ht, t0, tc, fvt)*
  - *ht: health topic, the label*
  - *t0, tc: creation time, last update time*
  - *fvt: (fv, ff)*
    - *fv: analogous to the tweet fv*
    - *ff (fwu, fwb, fhu, fhb): the frequencies associate to unigram and bigram*

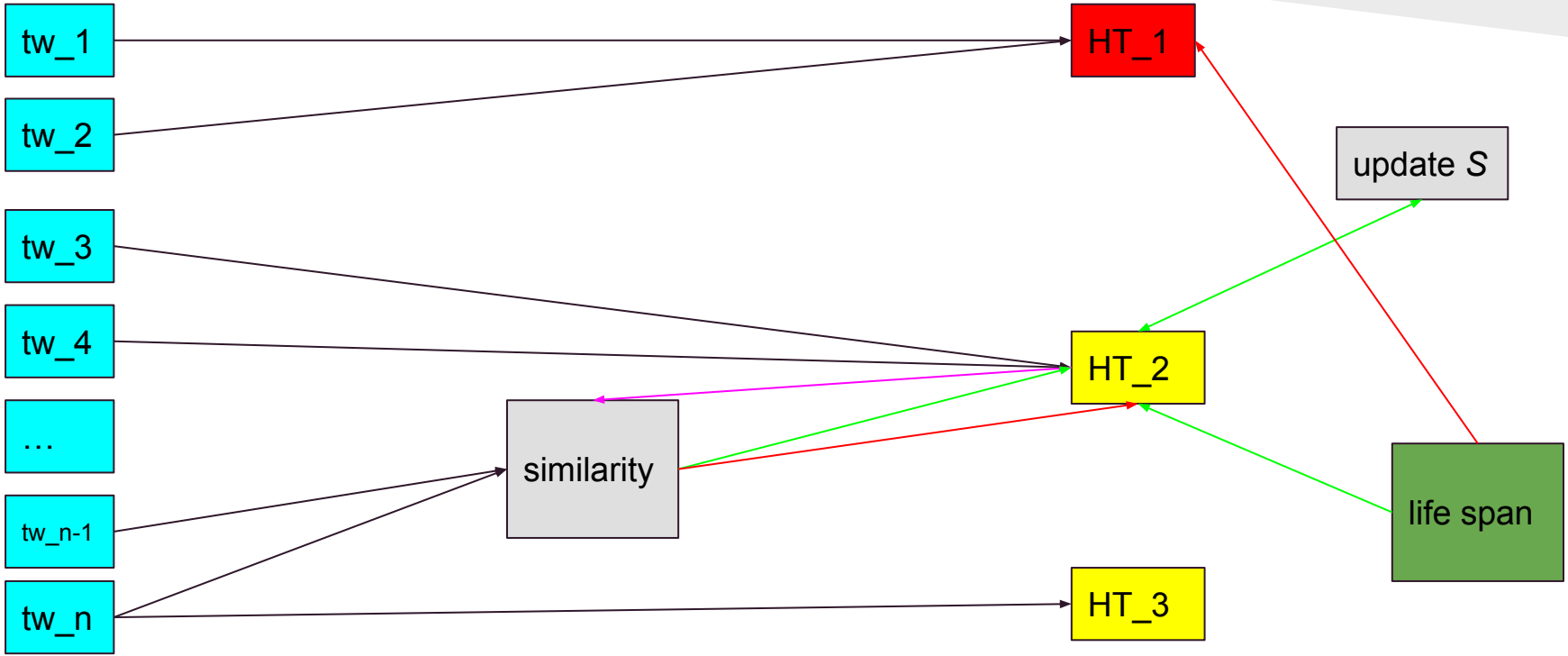
HealthS-Tweet: Algo



HealthS-Tweet: Algo



HealthS-Tweet: Algo



HealthS-Tweet: Results

- outperform traditional topic modeling
- the topic are very different (spanning from seasonal disease to common disease)
- future developments*:
  - HT sentiments
  - improve the labeling about HT
  - automatic hyperparameters tuning

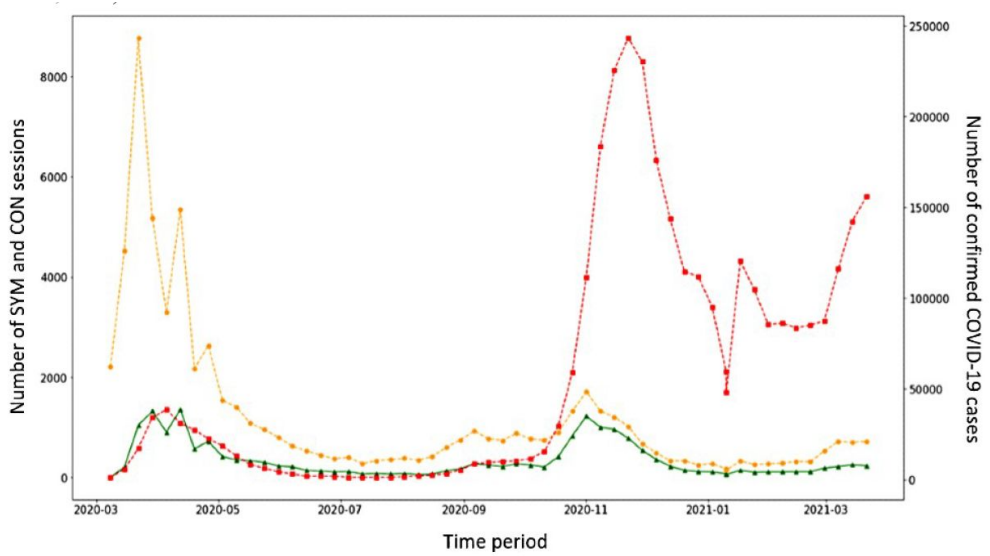
Online DSS

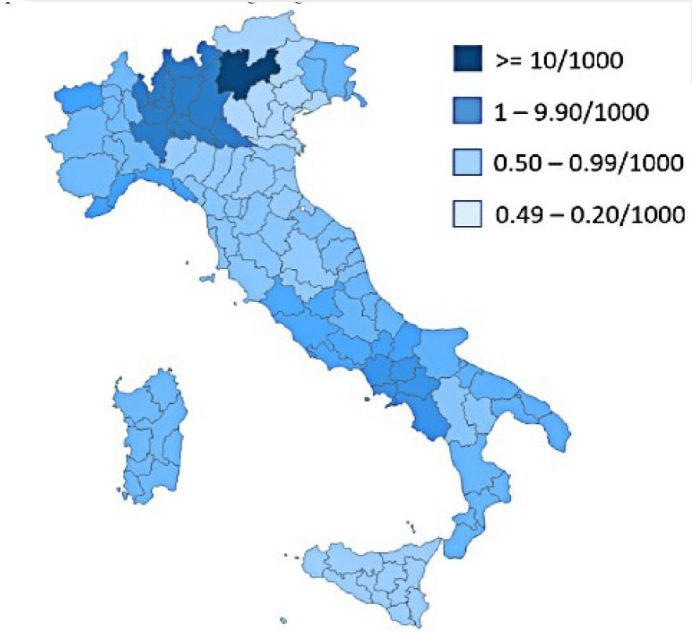
A digital Italian health tech startup, ***Paginemediche***, developed a noncommercial, online DSS with a chat user interface early 2020: **VISITAMI**

This study aimed to compare the trend in online DSS sessions with that of COVID-19 cases reported by the national health surveillance system in Italy, from February 2020 to March 2021.[4]

Online DSS: dataset

- 75557 sessions
  - 65207 were sessions by symptomatic users
  - 19062 were by contacts of individuals with COVID-19





Online DSS: Analysis

- To perform the comparison, we first used moving averages to display the curves.
  - K was set at a value of 7, as data from the national surveillance system were published weekly.

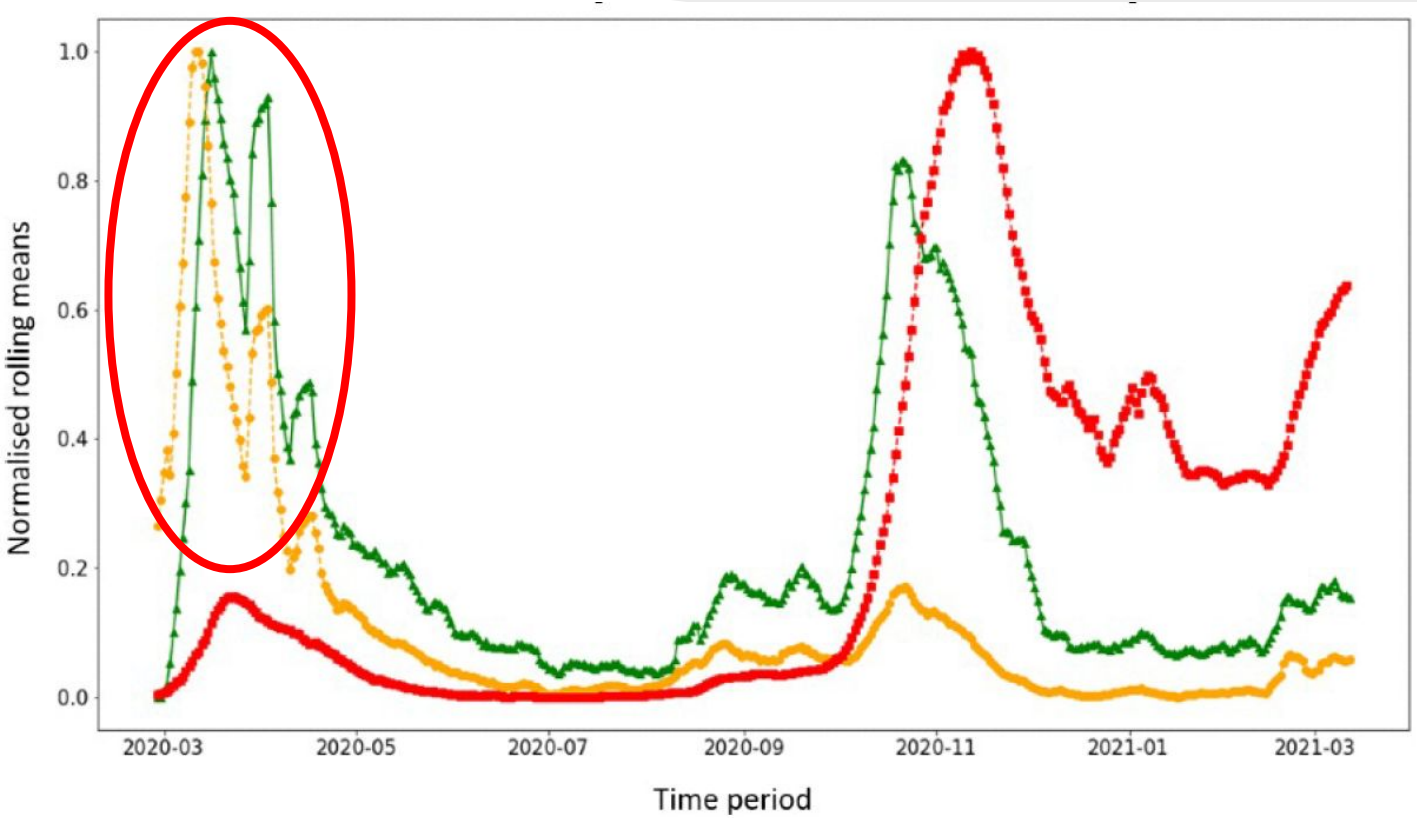
$$MA_k = \frac{1}{k} \sum_{i=n-k+1}^{n} p_i$$

- We then scaled each time series in a range between 0 and 1

On the scaled Time Series:

- We first applied Symbolic Aggregate approXimation (SAX) encoding
- Then, we calculated the Hamming distance between SAX Strings
- To verify that the online DSS anticipated the trends observed in notified cases, we shifted its time series 1 week ahead.

Online DSS: Analysis



SAX Encoding

Symbolic Aggregate approXimation Encoding:

- developed by Keogh and Lin in 2002
- transforms time series in sequence of symbols
- robust for missing values
- unsupervised

Adolphe Sax saxophone creator.
Correlation(SAX, Sax) = 0.00

Data Preparation

Il SAX Encoding needs data organized as follows:

- for each row a time series
- for each column a timestep
- standardized data

Regione	1	2	...	57
Abruzzo	0	0	...	2067
Basilicata	0	...	...	245
...	...	...	...	...
Veneto	32	...	...	10077

Regione	1	2	...	57
Abruzzo	-1.067762	-1.067762	...	1.699956
Basilicata	-1.063952	...	...	1.041431
...	...	...	...	...
Veneto	-1.244880	...	...	10766

protezione civile's data about Covid-19

Piecewise Aggregate Approximation

Idea

"SAX encoding is a method used to simplify time series by symbolically representing periods, the data becomes much smaller and easier to deal with, while still capturing its important aspects."

A Time Series $Y = [Y_1, Y_2, \ldots, Y_n]$ can be reduced in a sequence $X = [X_1, X_2, \ldots, X_m]$ with $m \leq n$ using:

$$\bar{X}_i = \frac{m}{n} \cdot \sum_{j=n/N(i-1)+1}^{(n/M)\cdot i} x_j$$

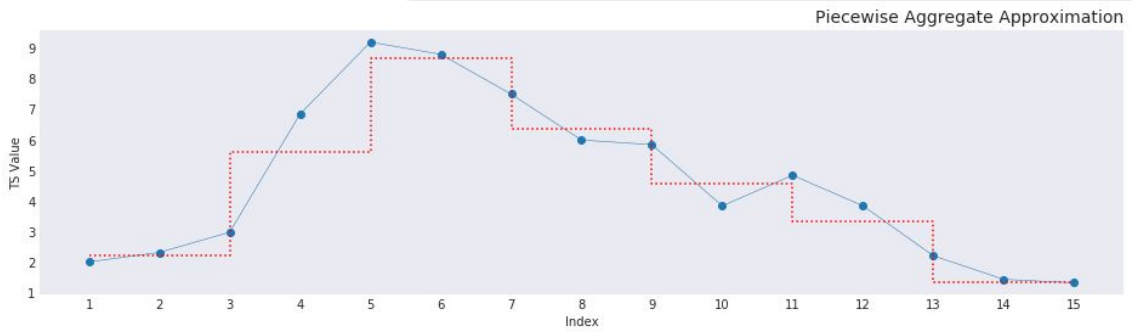
special cases:

- m=n return the original series
- m=1 return only one value, the average of the original series

Piecewise Aggregate Approximation

Important, the setting of hyperparameter ***w,* the time window.**

It controls the number of segments, if ***len(TS)/w*** isn't integer round up so as not to lose information.



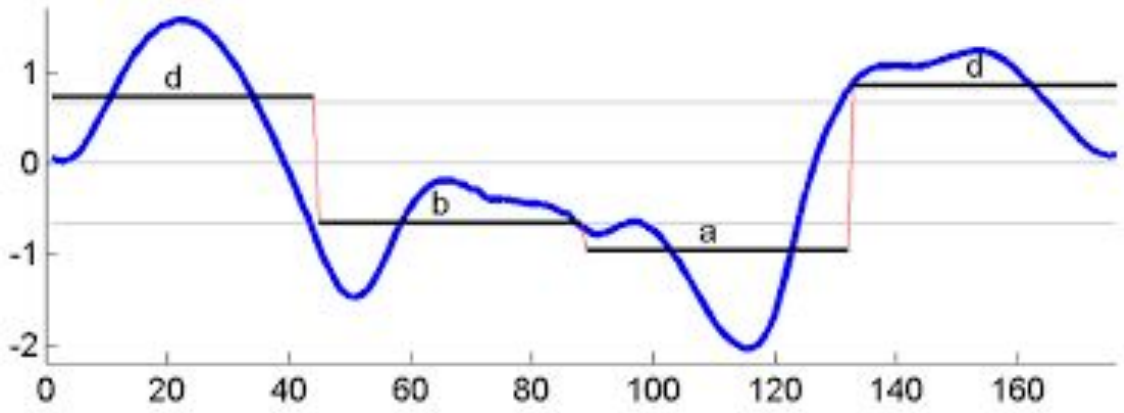
denominazione_regione	0	1	2	3	4
Abruzzo	-1.060876	-0.785615	0.352059	1.252060	1.696609
Basilicata	-1.055972	-0.846662	0.516006	1.239693	1.028541
Calabria	-1.088706	-0.797959	0.449874	1.246324	1.333274

SAX String

Creation of the SAX Strings:

- choose how many levels
- choose the bounds
- labeling periods returned by PAA

denominazione_regione	SAX_string
Abbruzzo	CAABC
Basilicata	CABBA
Calabria	BABBB
...	...

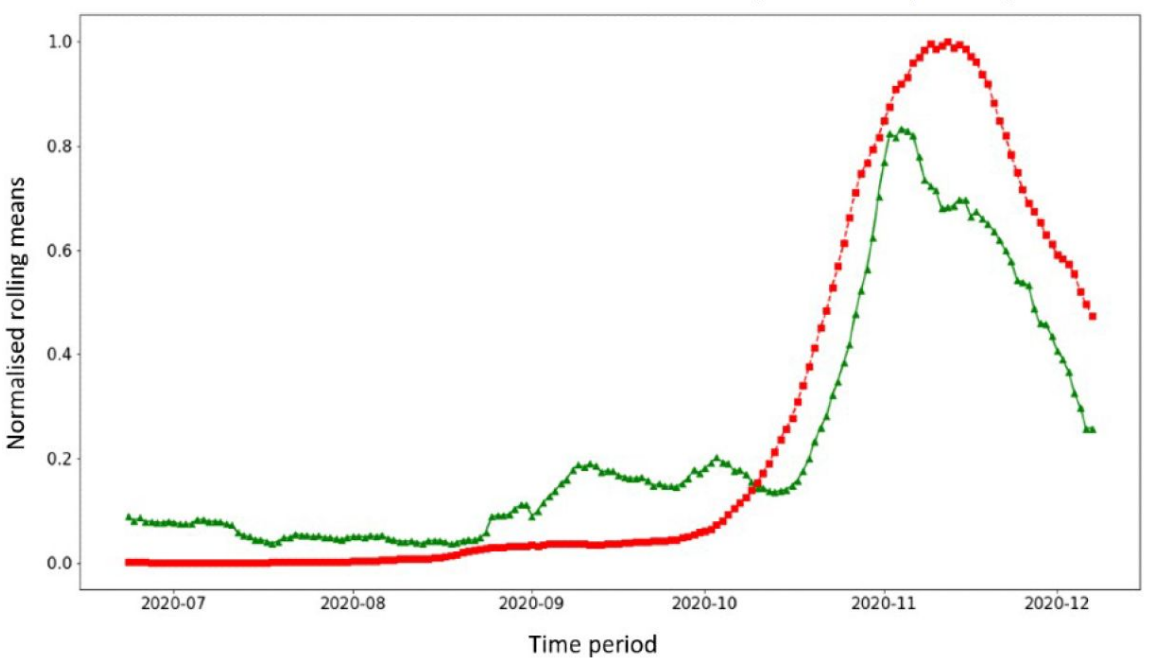


Online DSS: Results

- The series "user with contacts" was more consistent with the trend in confirmed cases than "user with symptoms"

- Throughout the period, after applying the 1-week shift, the Hamming distance improved from 0.49 to 0.46

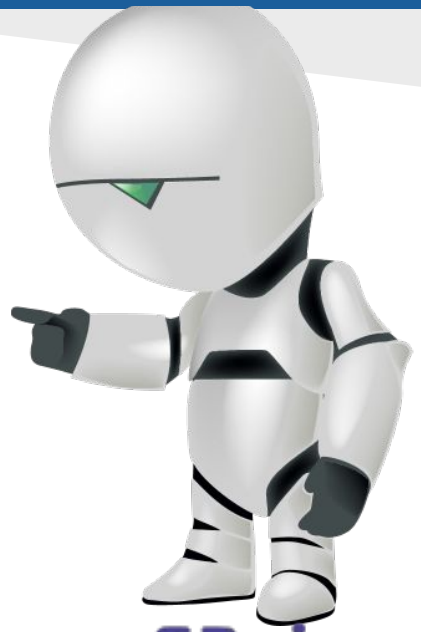
- July - December 2020, after applying the 1-week shift, the Hamming distance improved from 0.16 to 0.08



I haven’t a “Conclusion”

Let’s Chat!

Thank You!



tag Talent Garden Cosenza

AI DEMOCRATIZE

iaml
ITALIAN ASSOCIATION FOR MACHINE LEARNING

References

1. Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* 457.7232 (2009): 1012-1014.

2. Comito, Carmela, Clara Pizzuti, and Nicola Procopio. "How people talk about health? Detecting health topics from Twitter streams." *Proceedings of the International Conference on Big Data and Internet of Thing*. 2017.

3. C. Comito, C. Pizzuti and N. Procopio, "Online Clustering for Topic Detection in Social Data Streams," *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2016, pp. 362-369, doi: 10.1109/ICTAI.2016.0062.

4. Tozzi A, Gesualdo F, Urbani E, Sbenaglia A, Ascione R, Procopio N, Croci I, Rizzo C. "Digital Surveillance Through an Online Decision Support Tool for COVID-19 Over One Year of the Pandemic in Italy: Observational Study", J Med Internet Res 2021;23(8):e29556.

5. "Individuare Pattern con il SAX Encoding". IAML Blog