Big Data Management

Technische Universiteit Eindhoven University of Technology

Milestone 1

Group 6
Tingyuan Zhang, Ruichen Hu, Simin Sun,
Mingzhe Shi, Siyue Chen, Jing Wu

Video link: https://youtu.be/S7mFgBKu6Ds

Dataset description

Wikipedia page visit frequency Obtained from Massviews Analysis [1]

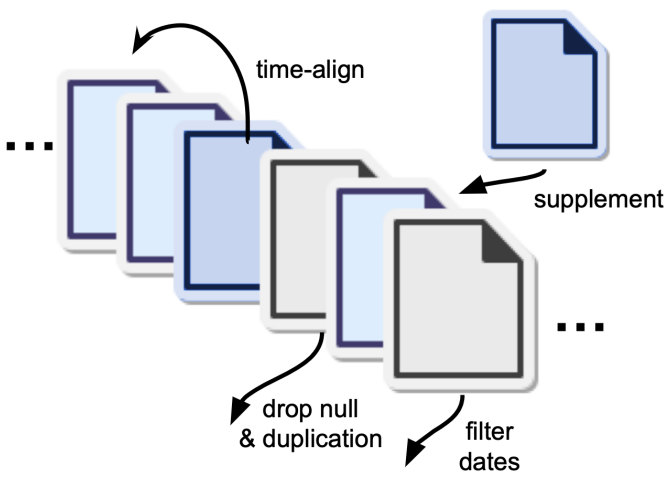
Content	Wiki Topic, Date, Visit frequency
Size	32 MB
Number of time series	1001
Length of each time series	1305
Total rows	1326326
Date range	01/01/2016 - 31/12/2020

Stock Obtained from the lecture.

Content	Stock name, Date, Price, Volume
Size	3.3 GB
Number of time series	29821
Total rows	38374198
Date range	01/01/2016 – 31/12/2020

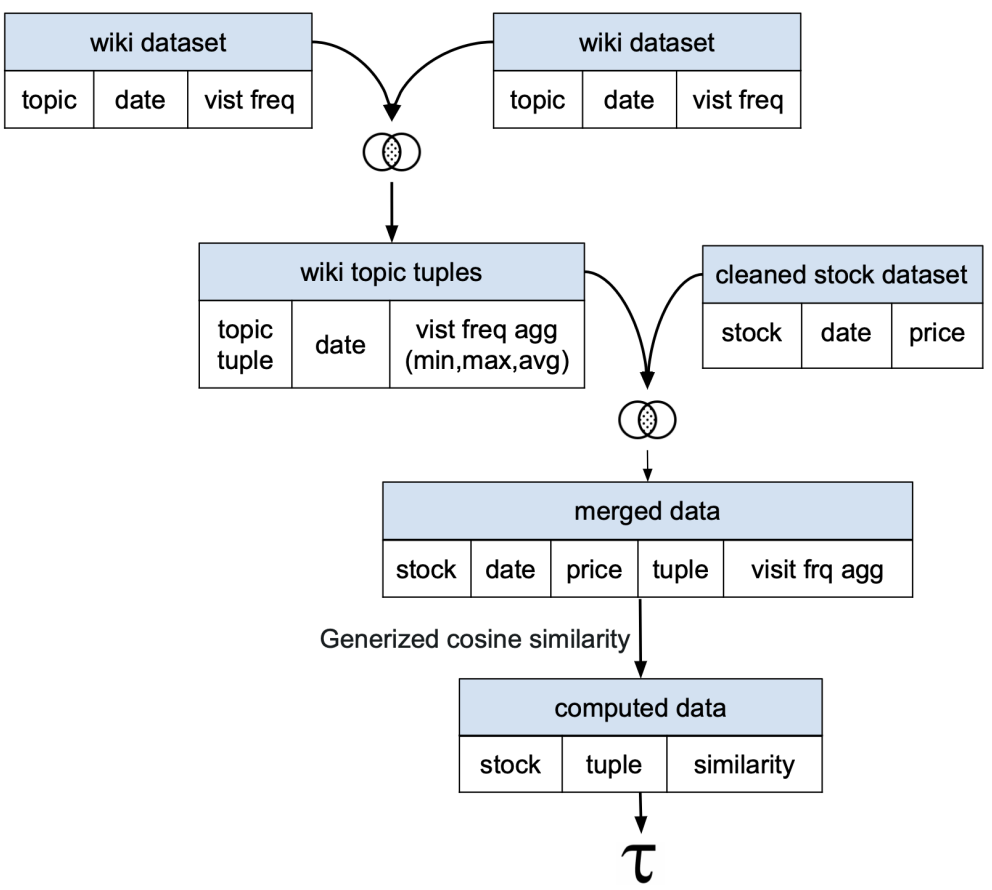
Data preprocessing

In part 1, we preprocess and clean the data from the stock dataset with the following steps:



Query architecture

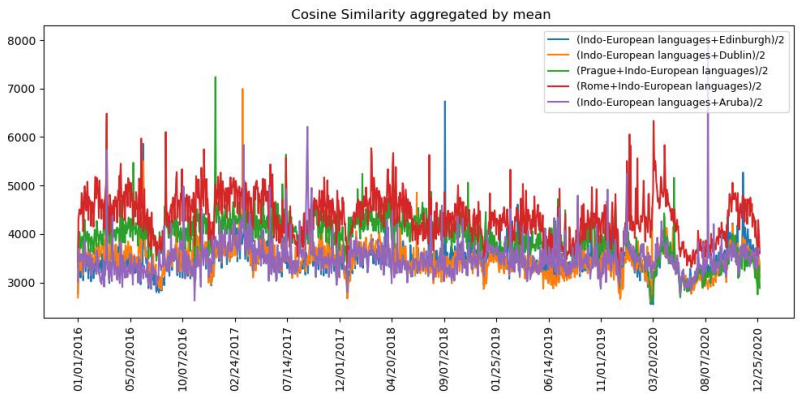
We try to find the relationship between two datasets. So we implement a series of query operations on them as shown below.



Ethical aspects

- There are certain relationships between some stocks price and wikipedia topics views.
- It is not always ethical to merge different datasets, especially to the obviously uncorrelated ones
- Correlation does not imply causation, thus the insight might be bias. We can mine some not easily perceived information, but it may in turn mislead us

Result



	τ **value**	#result	runtime part2	runtime part3
Avg	0.994122379	20	331.4s	1217.8s
Min	0.995819854	20	336.3s	1.5 s
Max	0.991408893	20	371.8s	1.4s

[1] https://pageviews.wmcloud.org/massviews/

Stock_code	Average		Min		Max	
26933*	Indo-European languages	Edinburgh	Belgium	Romania	Indo-European languages	Dublin
26933	Indo-European languages	Dublin	Czech Republic	Romania	Prague	Ancient Greece
26933	Indo-European languages	Prague	Romania	Kazakhstan	Prague	Indo-European languages
26933	Indo-European languages	Rome	Caribbean	Oceania	Indo-European languages	Edinburgh
26933	Indo-European languages	Aruba	Saint Petersburg	French language	Indo-European languages	"Latin America"

* Futures--Indices--ETF_ETF_iShares-eb.rexx-R-Gov.-Germany