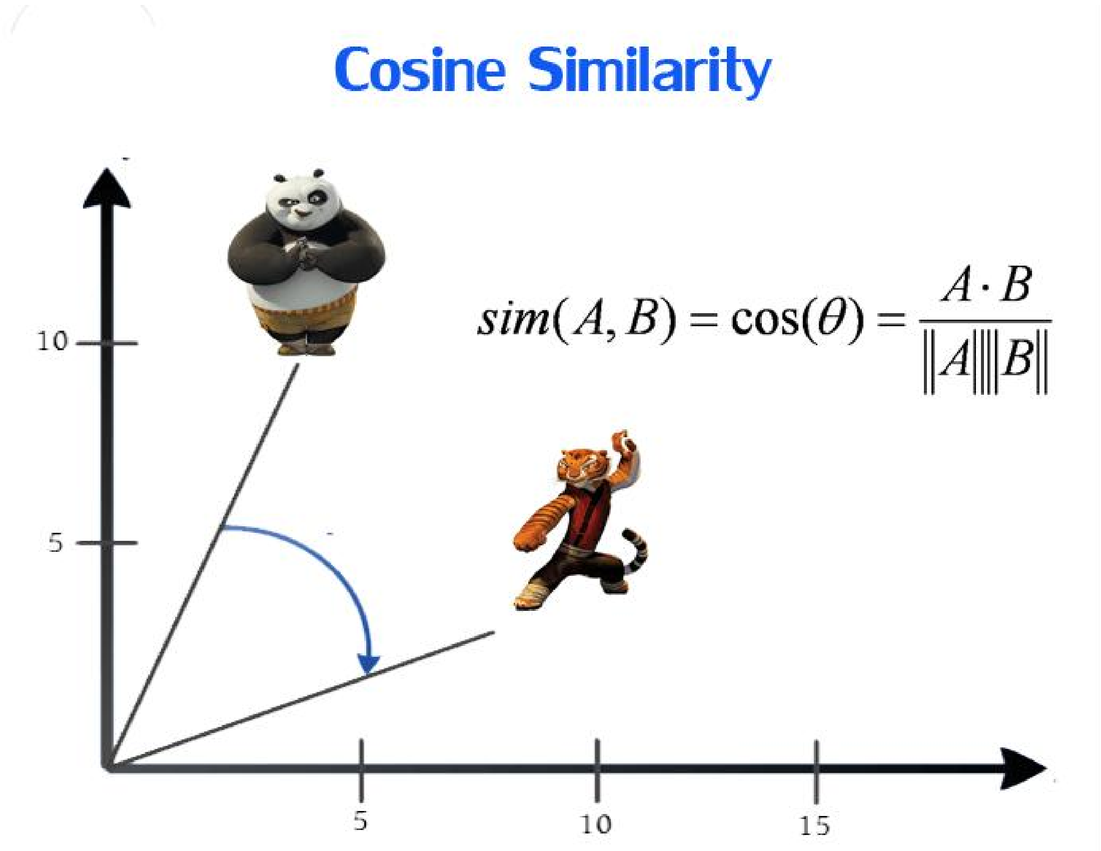
Dedup - Deduplication of Records

An analytics tool for deduplicating your records in a file

Technique Used: Cosine Similarity



- Cosine Similarity is a measure of the cohesiveness of items within a cluster, i.e. we can use it to check how close is a word to the other word !

- It is based on the **orientation** of the word rather than its magnitude, i.e. cos(theta) is calculated based on the vectors derived from the text itself. Ex:
    text1 = "Vladimir Antonio Frometa Garo"
    text2 = "Vladimir Antonio F Garo"

- Now the similarity between text1 and text2 = 0.75

-Calculation:

```
##########################
#   cos(theta) = v1.v2 / (||v1||.||v2||)    #
##########################
```

where, v1 and v2 are vectors derived from text1 and text2

And

||x|| means norm of vector x.

So, here v1 = [1 1 1 1] and v2 = [1 1 1 0]
thus,

Similarity = cos(theta) = 0.75

and a threshold is set by analysing the facts to capture similar words.

Workflow:

- Reading .csv file
- Making a dictionary on the basis of ‘DOB and Gender’ because it will initially segregate the data into **mega - clusters**, i.e. Ex. Key = ‘24/11/34M’ will give the names of all **Males** having **DOB** as **24/11/34**
- After making mega - clusters, I have applied cosine similarity on each mega- cluster to make **mini clusters** among them so that all the people that looks similar aggregates into a cluster and other people in other cluster and so on,
- In this way, all the duplicated records have been removed from the given .csv file of records.