NLP-Text Clustering

Sarang Mete · Following

2 min read · Nov 12, 2022

Photo by Pawel Czerwinski on Unsplash

Grouping similar sentences/text together.

Methods to check closeness/similarity:

- Euclidean distance: ||a-b||2 = √(Σ(ai-bi))
- Squared Euclidean distance: ||a-b||22 = Σ((ai-bi)2)
- Manhattan distance: ||a-b||1 = Σ|ai-bi|
- Maximum distance:||a-b||INFINITY = maxi|ai-bi|
- Mahalanobis distance: √((a-b)T S-1 (-b)) {where, s : covariance matrix}

There are hundreds of clustering algorithms but commonly used are Centroid models(like K-means) and Hierarchical clustering(like DBSCAN).

Hierarchical (bottom-up) :

1.Assign each data point as a cluster

2.Merge 2 nearest clusters until we've only 1 big cluster

3. Then decide, number of clusters to keep.

We can't predict new data point cluster with DBSCAN because it works on whole text unlike k-means. So if prediction is needed then go with k means else go with DBSCAN.

Great article on DBSCAN.

KNN vs K means

KNN:

Check k nearest neighbors of new data point and assign label on majority label of neighbors.

K-means:

1. Randomly initialize centroids from data points.
2. Assign each data point to cluster using min distance
3. Move centroids to mean of newly assigned data points
4. Repeat steps-2 and 3 until convergence or given number of iterations are over.

K means++: Instead of randomly initializing centroids,use distribution

How do you decide the number of clusters?

1. **The Elbow Method**

Find error/distance of a data point from centroid for each value of k,

choose k where error stops reducing.

1. **The Silhouette Method**

measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation)

I've created a complete end to end project . You can refer it here.

If you liked the article or have any suggestions/comments, please share them below!

Let's connect and discuss on LinkedIn

Resources:

How to Utilize Machine Learning to Automatically Detect Patterns in Text

In the last post, we talked about Topic Modeling, or a way to identify several topics from a corpus of documents. The...

www.dataknowsall.com

https://ruffinisilvia.medium.com/textual-clustering-summarization-and-visualization-e0dcc5c2d3b

https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Breaking it Down: K-Means Clustering

Exploring and visualizing the fundamentals of K-means clustering with NumPy and scikit-learn.

towardsdatascience.com

https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/

Text Clustering NLP K Means Clustering Nltk Sklearn

More from the list: "NLP"

Curated by Himanshu Birla

Jon Gi... in Towards Data ...

Characteristics of Word Embeddings

· 11 min read · Sep 4, 2021

Jon Gi... in Towards Data ...

The Word2vec Hyperparameters

· 6 min read · Sep 3, 2021

Jon Gi... in

The Word2ve

· 15 min rea

View list

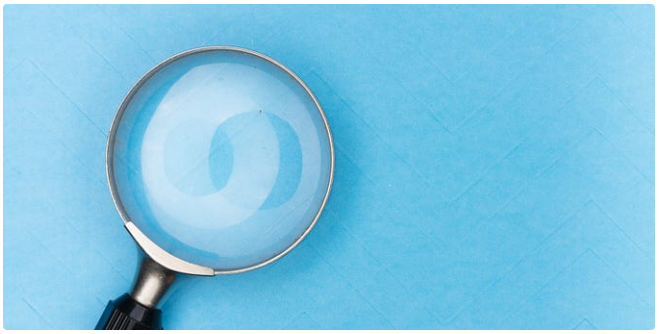
Written by Sarang Mete

162 Followers

Following

Senior NLP Engineer

More from Sarang Mete



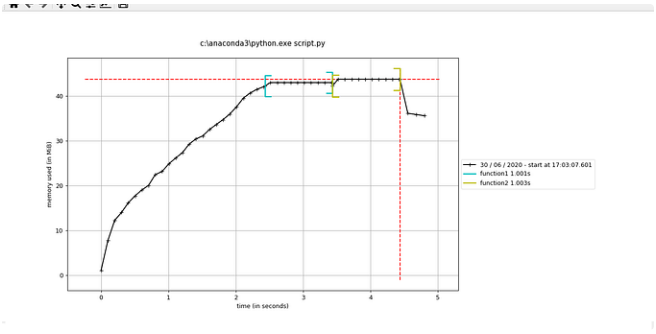
Sarang Mete

NLP-Semantic search using elasticsearch and embeddings

Semantic Search

3 min read · Nov 14, 2022

18 1



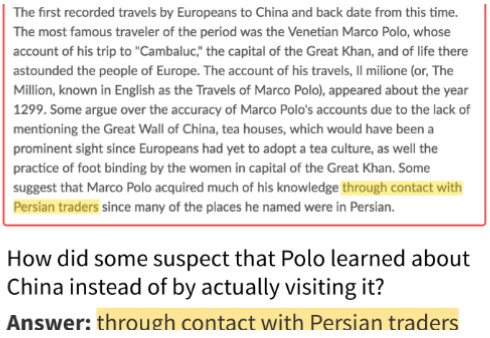
Sarang Mete in Analytics Vidhya

MLOps-Calculate Memory Consumption of Python Code

Python developers should be aware of the memory consumption of the code they are...

3 min read · Dec 16, 2021

10



Sarang Mete in Analytics Vidhya

Question Answer System

Create Question-Answer system in python in few steps.

2 min read · Apr 22, 2020

Sarang Mete in Analytics Vidhya

OCR Corrector in RegEx Extraction

There are different methods to extract entities from textual images.Some of the methods are:

3 min read · May 30, 2020

See all from Sarang Mete

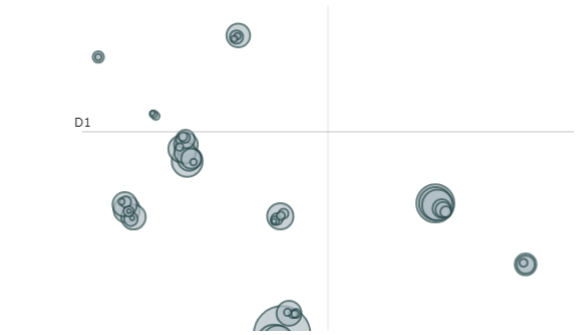
Recommended from Medium

Sirsh Amarteifio

Cluster chatter: HDBSCAN + LLM

HDBSCAN is a density based (hierarchical) clustering algorithm. Clustering algorithms...

5 min read · Jun 13



Jawwad Shadman Siddique

Topic Modeling Using BERTopic on Newsgroup Dataset: Python...

We go step by step from creating a google collab workspace to visualizing the cluster o...

7 min read · Jul 5

Lists

Natural Language Processing

669 stories · 283 saves

The New Chatbots: ChatGPT, Bard, and Beyond

13 stories · 133 saves

New_Reading_List

174 stories · 133 saves

Staff Picks

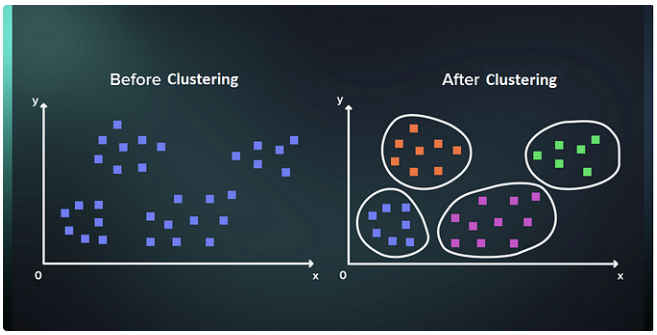
465 stories · 317 saves

Guillaume Michel in Kensho Blog

Kensho Classify: The Solution to Common Challenges of Text…

Text classification is widely used in many industries and often serves as a pillar for mo…

4 min read · Sep 7



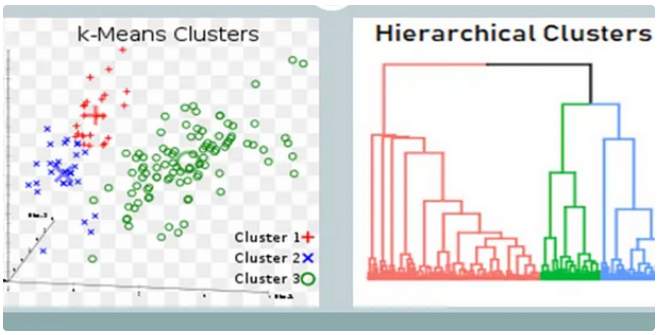
Deniz Gunay

Clustering

Clustering

20 min read · Sep 18

26



İlyurek Kılıç

Unsupervised Learning: Clustering with K-Means and Hierarchical…

Bryan

From Corpus to Multi-Label Classification

2 min read · Sep 25

14

A Practical Guide

2 min read · Sep 9

See more recommendations