Data to Discovery: Unveiling Clustering in BERTopic Topic Modeling

Abhiram Ravikumar
ML Engineer at Collinson

Jaspal Singh
Lead Data Scientist at Collinson

Conf42 | 18th May 2023

Who are we?

Abhiram Ravikumar

- Cloud Machine Learning Engineer, Collinson
- MSc in Data Science, King’s College London
- Data Science Research Fellow, SAP Labs (ex)
- LinkedIn Learning Instructor
- Volunteer at DataKind Bengaluru
- Loves to play badminton and listen to 80’s rock music
- Twitter: @abhi12ravi

Jaspal Singh

- Lead Data Scientist, Collinson
- Expertise: AI, Python, AWS and Data Products
- CBA, Advanced Business Analytics, Indian School of Business
- Loves to play football and favourite football club is Arsenal
- Twitter: @jaspalsingh26

Agenda

- Topic Modeling Use Case
- Why BERTopic?
- BERTopic end-to-end flow
- Clustering (HDB Scan)
- Dataset Description (Amazon Alexa reviews)
- Run through of topic modelling on Google Collab
- Conclusion and Future Scope

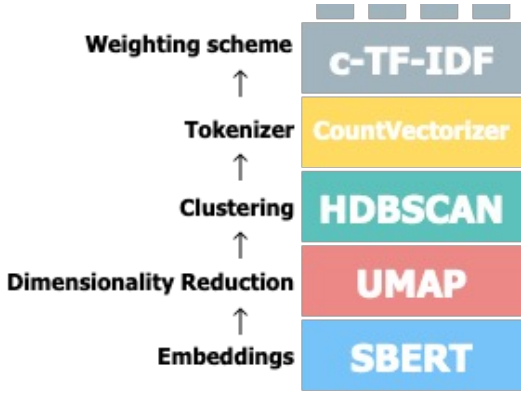
Topic Modeling Use Case

Topic Modeling

- Topic 1
- Topic 2
- .
- .
- .
- Topic n

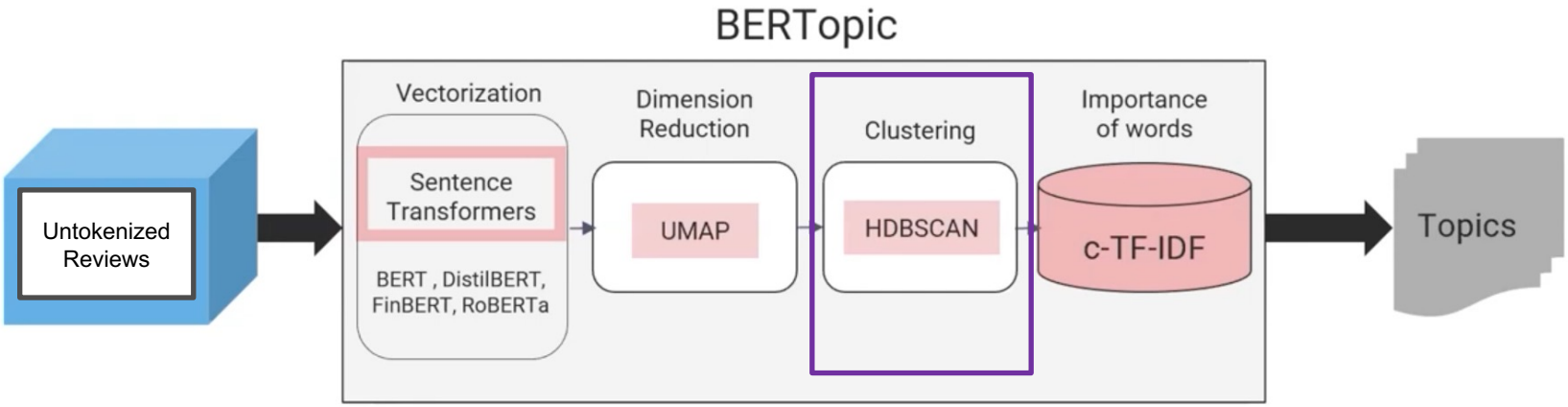
Why BERTopic?

- Works on unstructured data
- Takes advantage of transformer models
- Offers modularity
- Contextual embeddings
- Flexible structure
- New advancements in clustering can be adapted easily
- c-TF-IDF extraction of topic representations



Source: Clustering in BERTOPIC

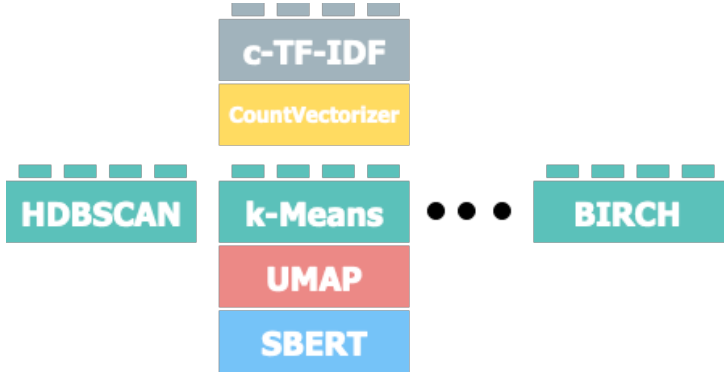
BERTopic End-to-End Flow



Source: [DataHour by Bharath Kumar Bolla]

Clustering

- HDBSCAN
- K-Means
- cuML HDBSCAN
- sklearn algorithms
  - Agglomerative clustering



Credits: Clustering in BERTOPIC

Dataset Description

rating	date	variation	verified_reviews	feedback
5	31-Jul-18	Charcoal Fabric	Love my Echo!	1
5	31-Jul-18	Charcoal Fabric	Loved it!	1
4	31-Jul-18	Walnut Finish	Sometimes while playing a game, you can answer a question correctly but Alexa says you got it wrong and answers the same as you. I like being able to turn lights on and off while away from home.	1
5	31-Jul-18	Charcoal Fabric	I have had a lot of fun with this thing. My 4 yr old learns about dinosaurs, i control the lights and play games like categories. Has nice sound when playing music as well.	1
5	31-Jul-18	Charcoal Fabric	Music	1
5	31-Jul-18	Heather Gray Fabric	I received the echo as a gift. I needed another Bluetooth or something to play music easily accessible, and found this smart speaker. Can't wait to see what else it can do.	1
3	31-Jul-18	Sandstone Fabric	Without having a cellphone, I cannot use many of her features. I have an iPad but do not see that of any use. It IS a great alarm. If u r almost deaf, you can hear her alarm in the bedroom from out in the living room, so that is reason enough to keep her.It is fun to ask random questions to hear her response. She does not seem to be very smartbon politics yet.	1

Amazon Alexa Reviews dataset - Kaggle

Hands-on: BERTopic

Diving Deep into HDBSCAN

Abhiram Ravikumar and Jaspal Singh | Conf42 | 18th May, 2023

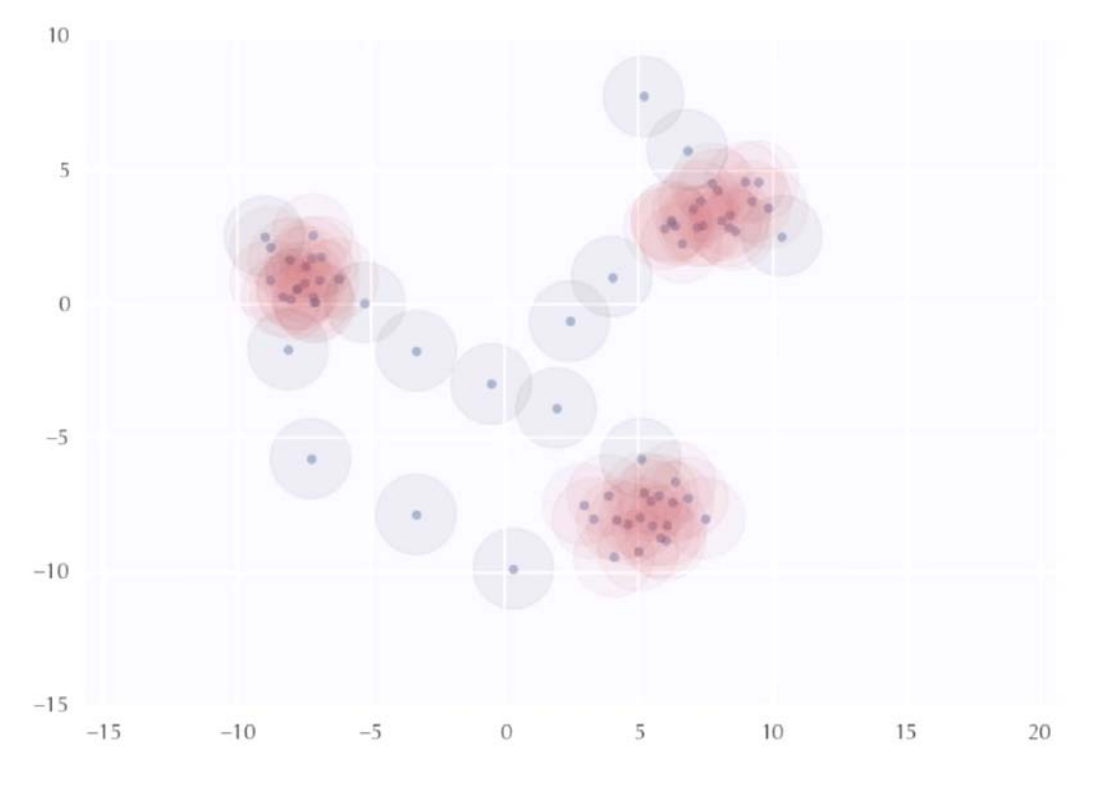
What is HDBSCAN?

HDBSCAN is a long acronym AND a clustering algorithm!

Hierarchical **D**ensity **B**ased **S**patial **C**lustering of **A**pplications with **N**oise

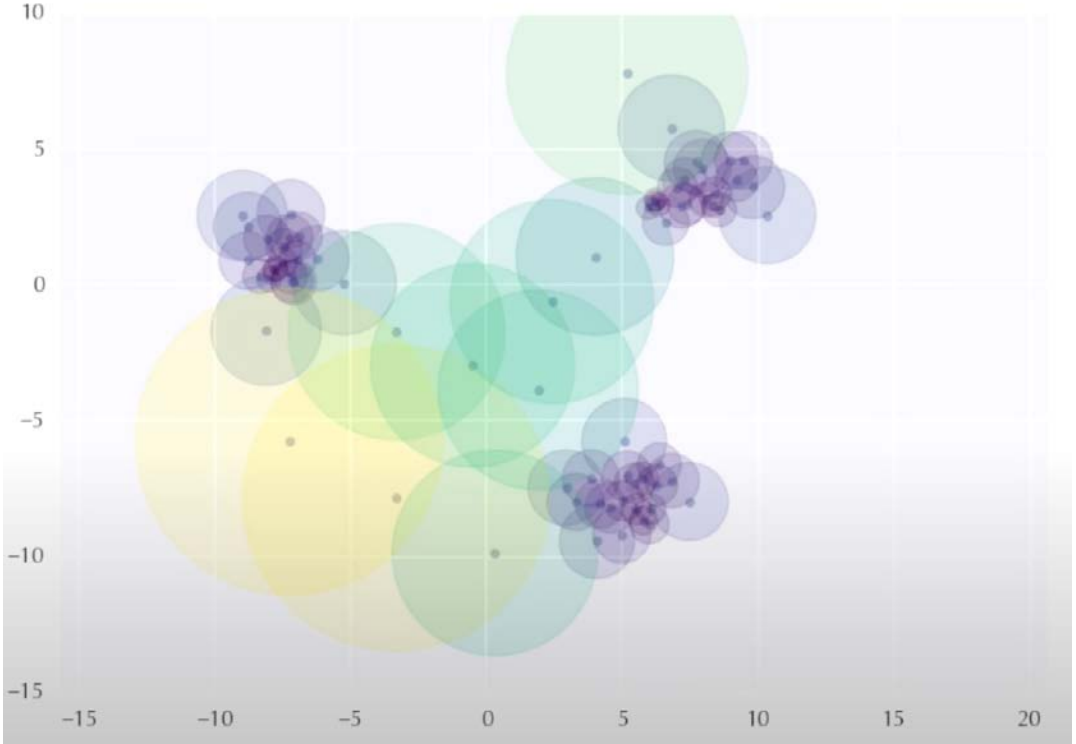
To understand HDBSCAN we need to know DBSCAN

- Clusters based on density.
- Circles/hyperspheres around data points of fixed radius “epsilon”
- Robust, flexible, and outlier-resistant.
- No predefined number of clusters
- Need to define radius of the circle
- SLOW!



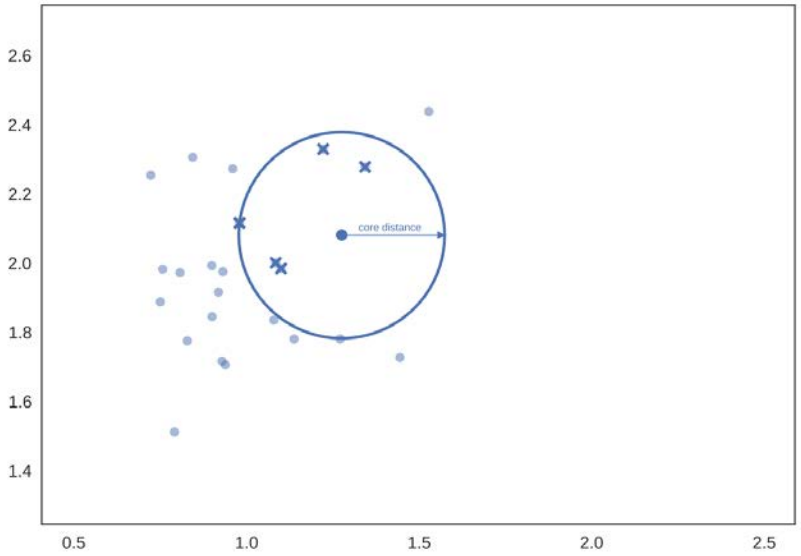
What if there was no fixed radius?

- No fixed radius
- Helps identify dense region
- FAST!

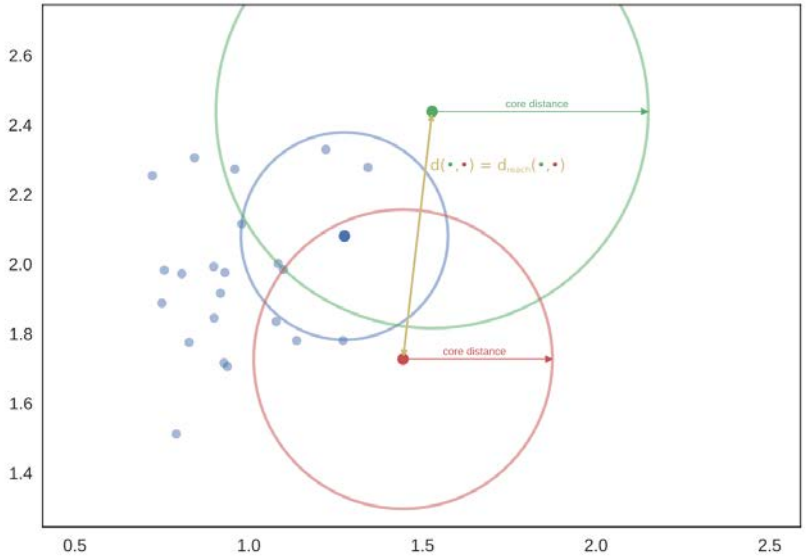


K-NN algorithm to define radius

K=5



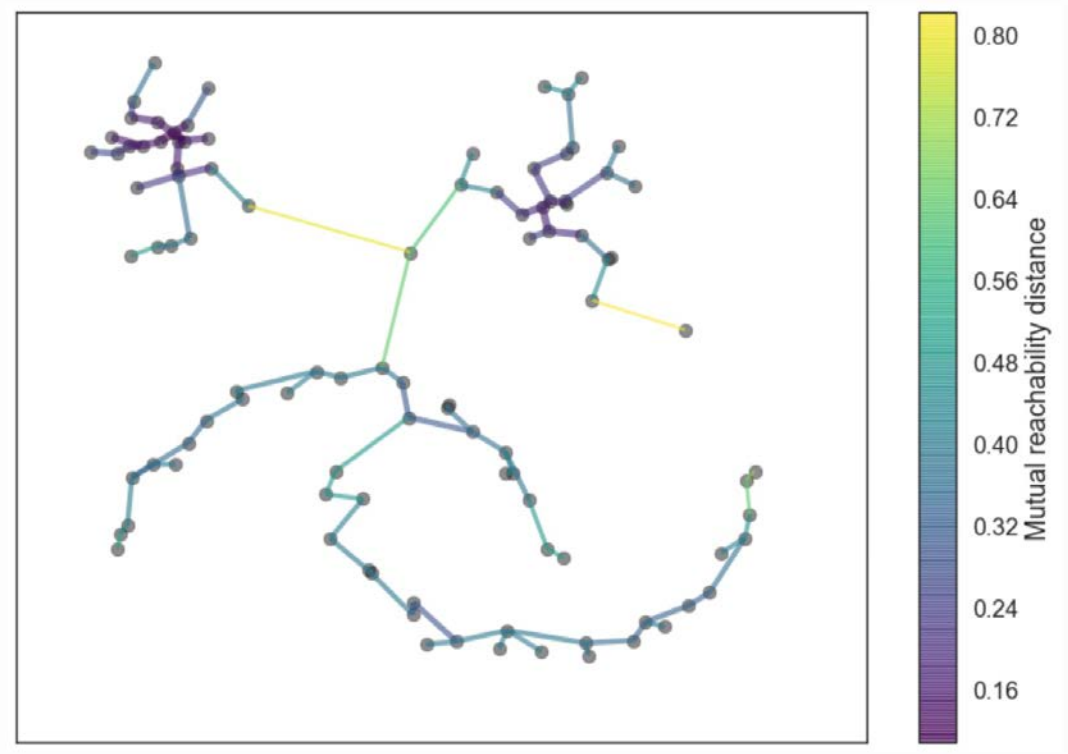
Core Distance



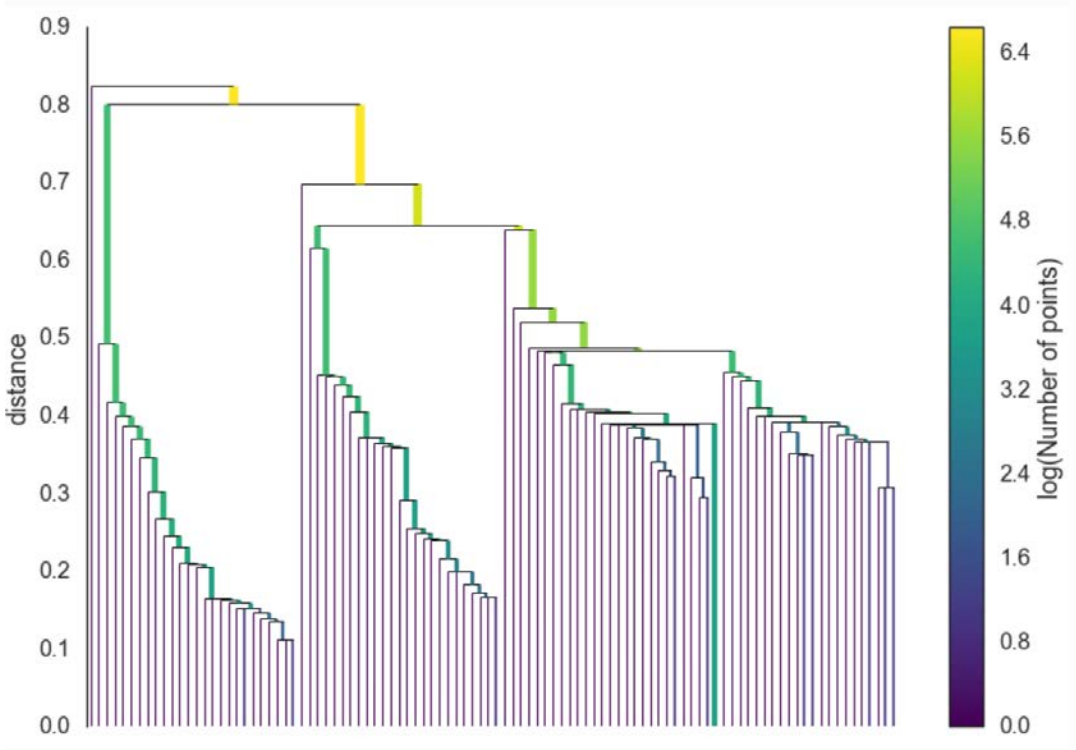
Mutual Reachability Distance

Minimum spanning tree finds density and hierarchy

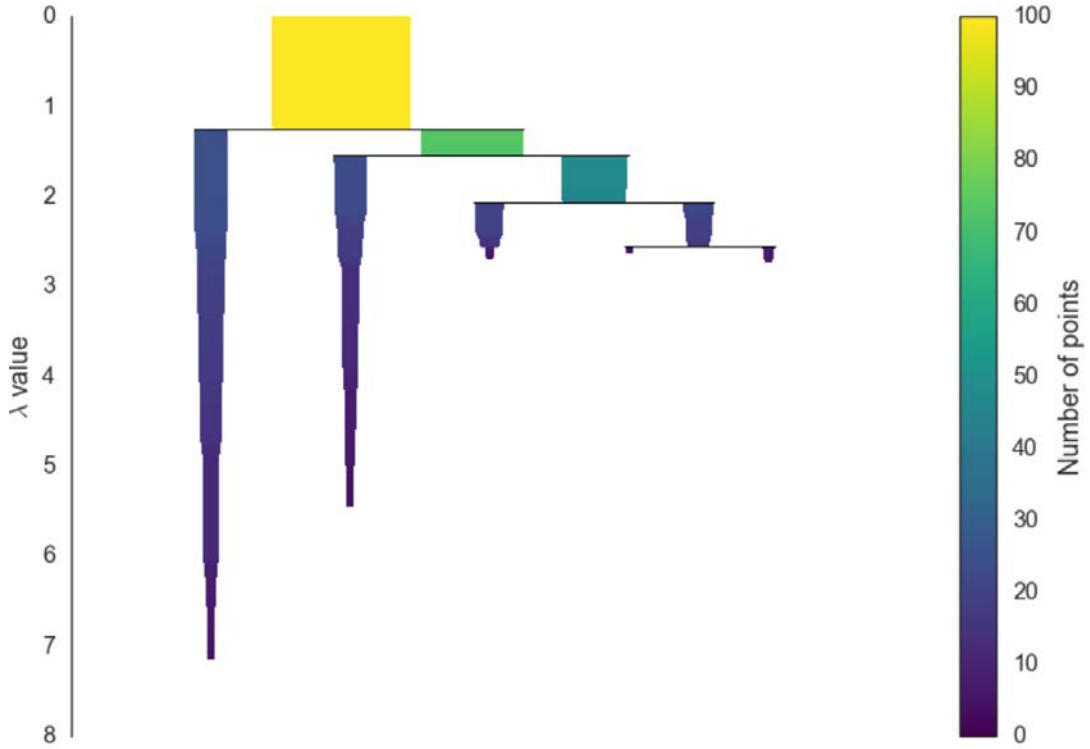
- Identifying connected components
- Hierarchical structure
- Clustering interpretation
- Efficient computation



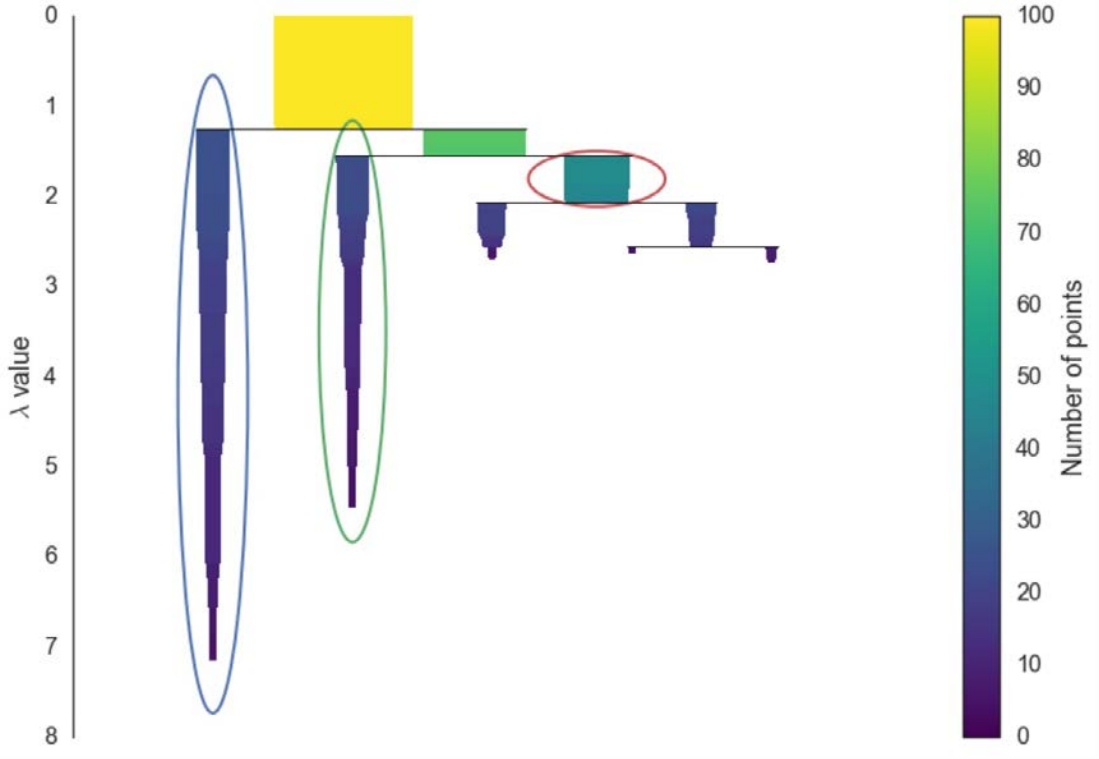
Density Based Spatial Clustering



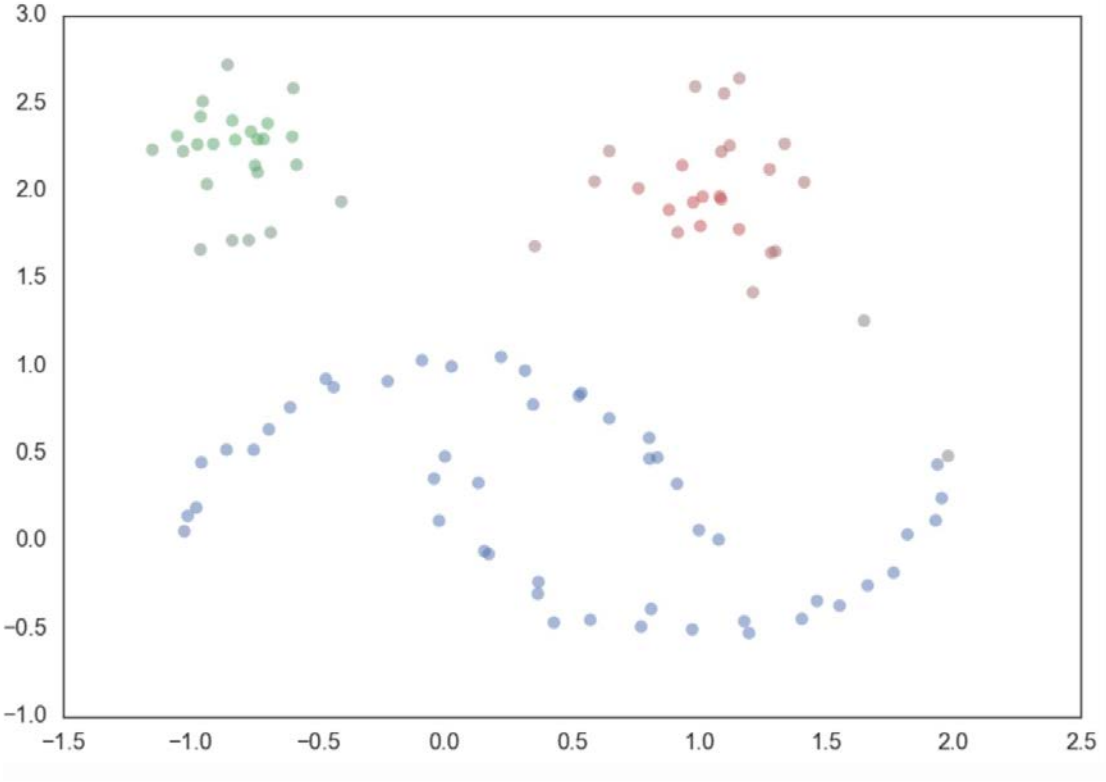
Stability score "λ"



Stability score "λ"

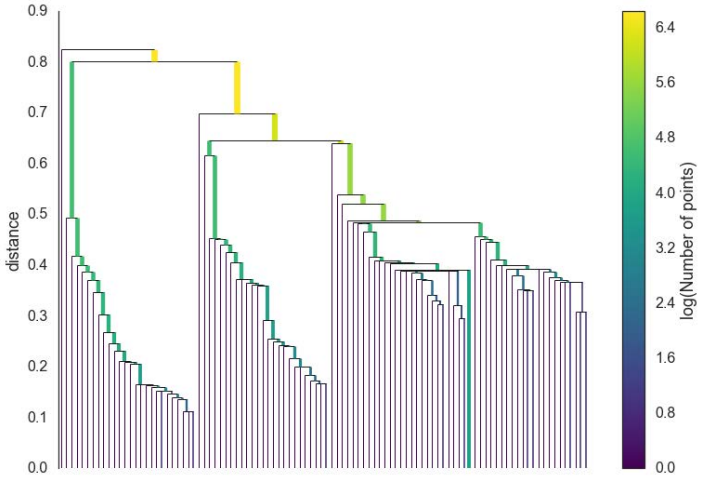


Final Clusters



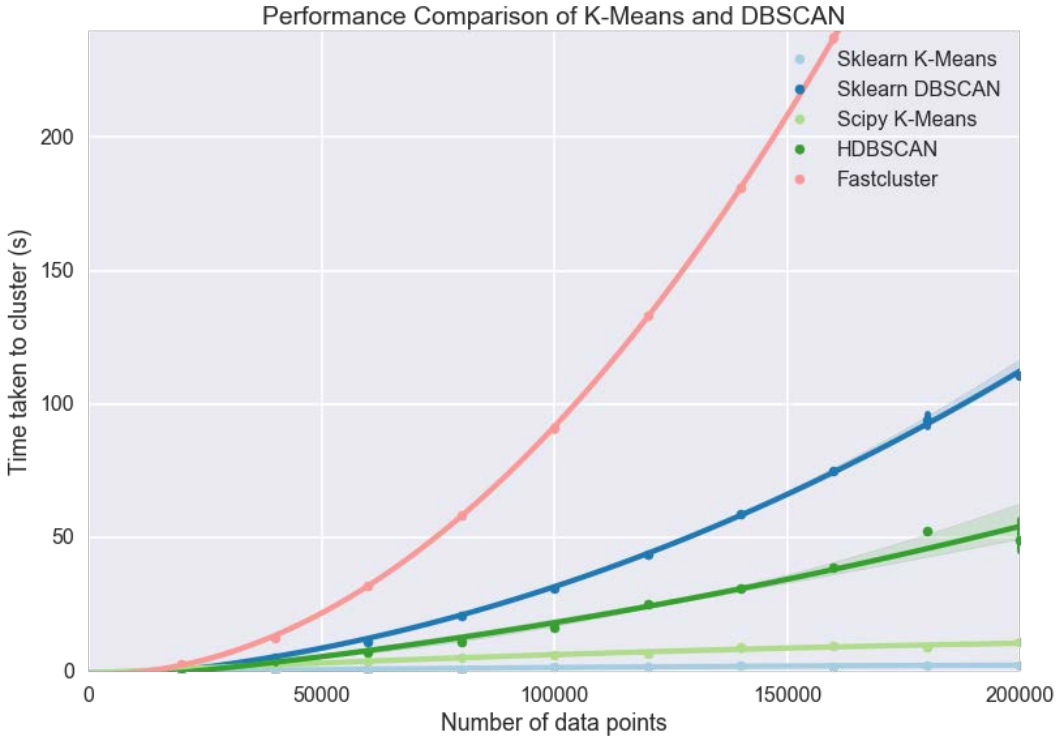
HDBSCAN steps

1. Transform the space as per density
2. Build minimum spanning tree
3. Construct cluster hierarchy
4. Condense the hierarchy based on min cluster size
5. Extract stable clusters from the condensed tree



Source: How HDBSCAN works

HDBSCAN – performance comparison



Source: http://hdbscan.readthedocs.io

HDBSCAN – strengths and weaknesses

- HDBSCAN focuses on high-density clustering -> reduces noise clustering problem
- Min-cluster-size parameter can be set, relatively fast
- Difficulty in handling large amounts of data
    - cuML HDBSCAN speeds up HDBSCAN using GPU acceleration

- Read: Comparing Python Clustering Algorithms
  https://hdbscan.readthedocs.io/en/latest/comparing_clustering_algorithms.html

Conclusion and Future Scope

- BERTopic – modular, scalable, flexible
- Clustering is modular
- BERTopic assumption: Every document contains only one topic
- Large Language Models (ChatGPT, etc.) could impact topic modeling
  - Be cautious of inherent biases, and ethical issues with LLMs
- Online topic modeling
- Cloud vendors
- Operationalization

References

1. The BERTopic Algorithm
2. Kaggle Dataset
3. BERTopic paper
4. HDBSCAN package
5. Pinecone HDBSCAN notebook

Session Resources

1. Slides on Speaker Deck: https://speakerdeck.com/abhi12ravi/
2. Notebooks on GitHub: https://github.com/abhi12ravi/BERTopic_Conf42/

Thank you!