Unsupervised Machine Learning

Part Two: Clustering

DistrictDataLabs

Agenda

- Clustering
    - Algorithms Review
    - Evaluation
    - Visualization

Clustering Algorithms Review

Clustering Algorithms

- K-Means
- Hierarchical
- Parallel canopy

K-Means

Straight Forward, Mature Algorithm:

- Select predefined K
- Pick K random points in the data set to be centroids
- For each point, assign it to closest centroid
- Compute middle of cluster, move centroid
- Repeat previous 2 steps until centers don’t move.

Considerations:

- Distance metric
- How do you choose K?

Hierarchical Clustering

We want strong membership as a hierarchy.

- Start with all data points as their own cluster
- Repeat until only a single cluster is left:
    - Find 2 closest points x_i and x_j
    - Merge points into a single cluster
    - Remove previous singleton clusters

This method creates a dendrogram of clusters- a hierarchical tree representing the cluster structure!

Canopy Clustering

An unsupervised *pre-clustering* algorithm that is often used as a preprocessing step for K-Means or Hierarchical clustering.

This algorithm is intended to speed up other clustering algorithms, particularly in large data sets that make these algorithms impractical.

Basically canopies are a form of “blocking” - reducing the computational space and the number of required pairwise distance comparisons.

Clustering Evaluation

There is no gold standard for evaluation so …

Internal Evaluation

Inspect the data that was clustered for quality:

1. Ratio of intra-cluster vs. inter-cluster distances.
2. Density of Clusters
3. Average distance to points in the cluster as opposed to outside (Silhoutte)

Usually highly dependent on algorithm choice.

External Evaluation

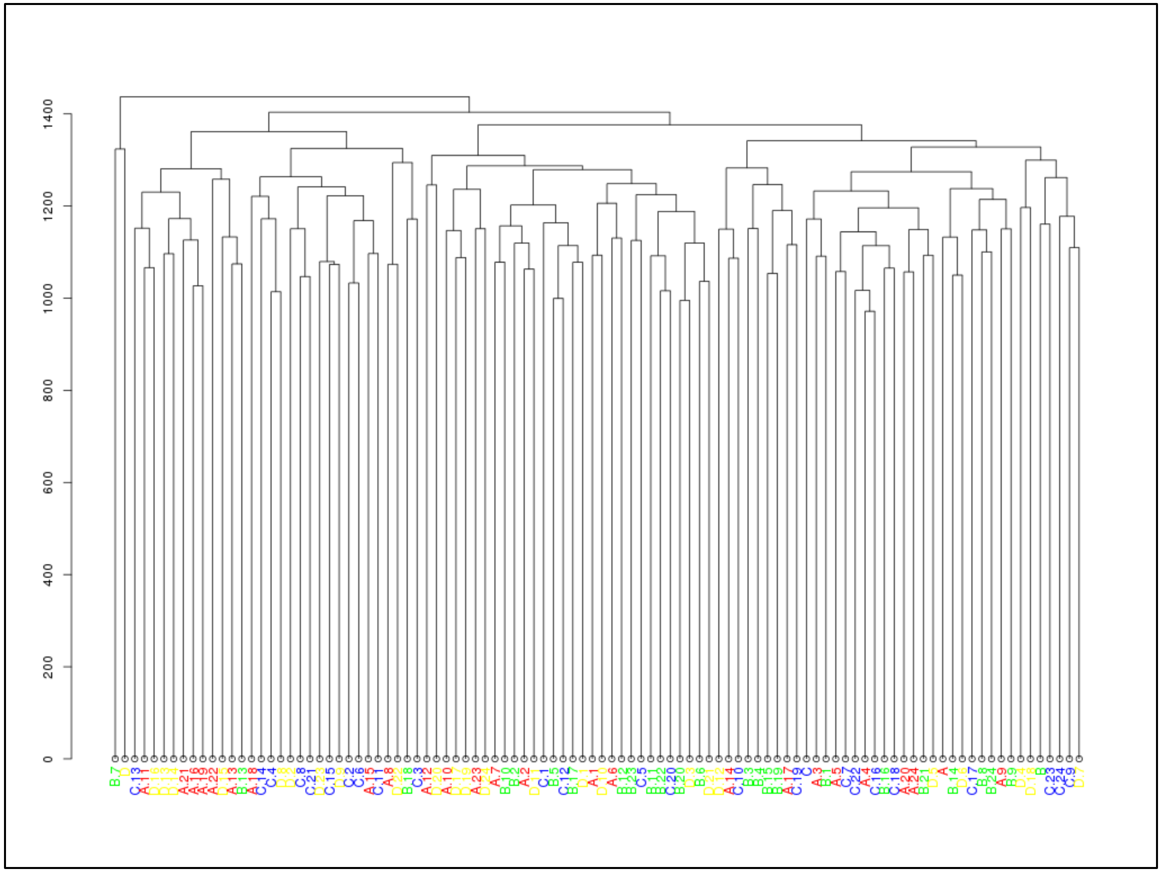
Evaluate based on known data that was not clustered.

1. Benchmarking
2. Pre-Classification

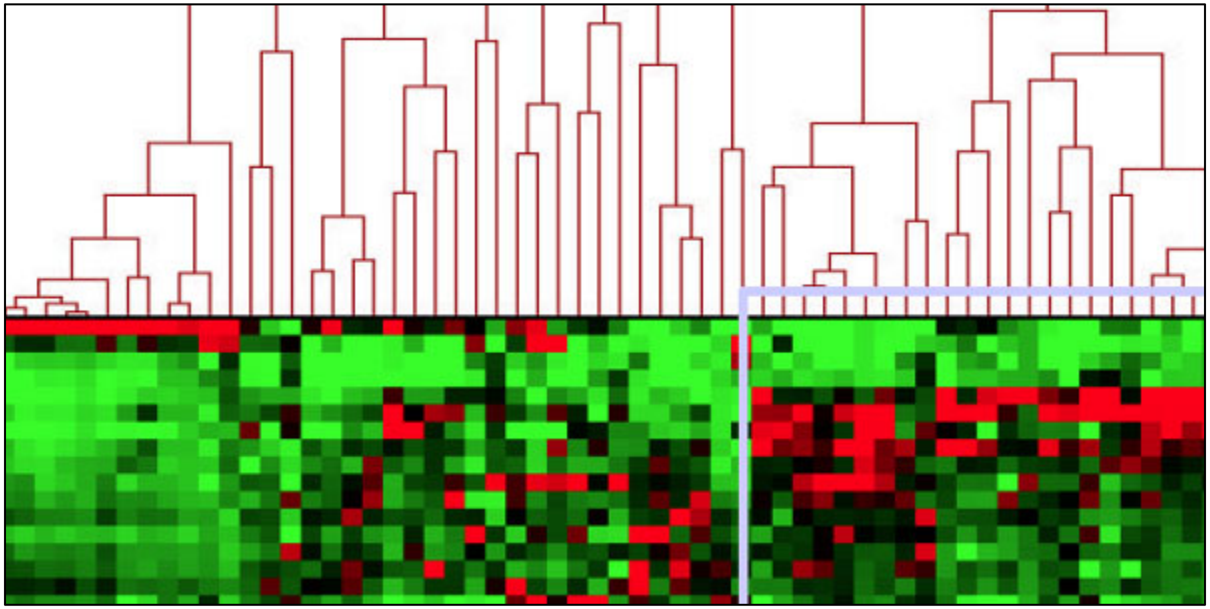
Similar techniques to classification.

Cluster Visualization

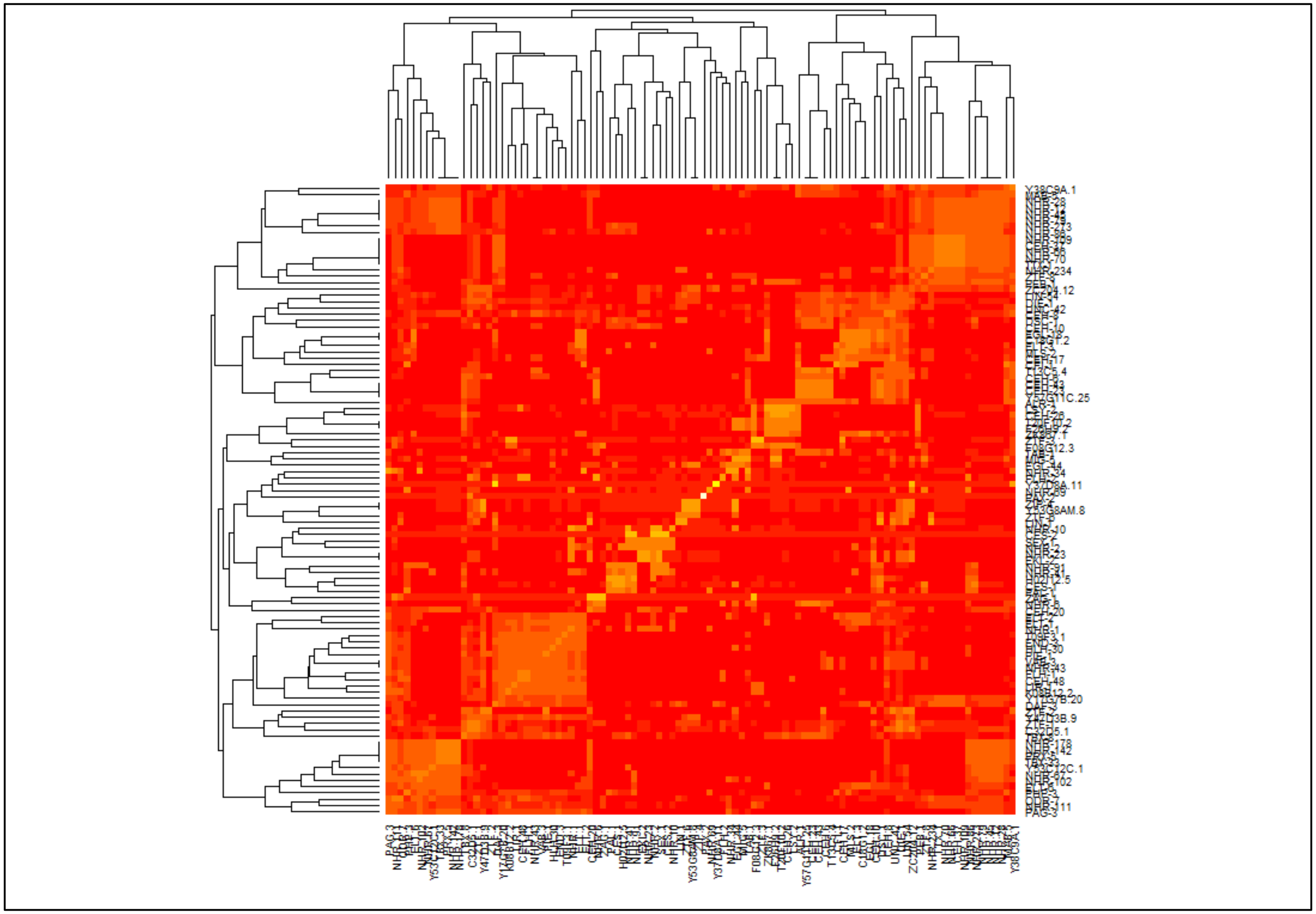
Dendogram



Hierarchical Clustering Explorer

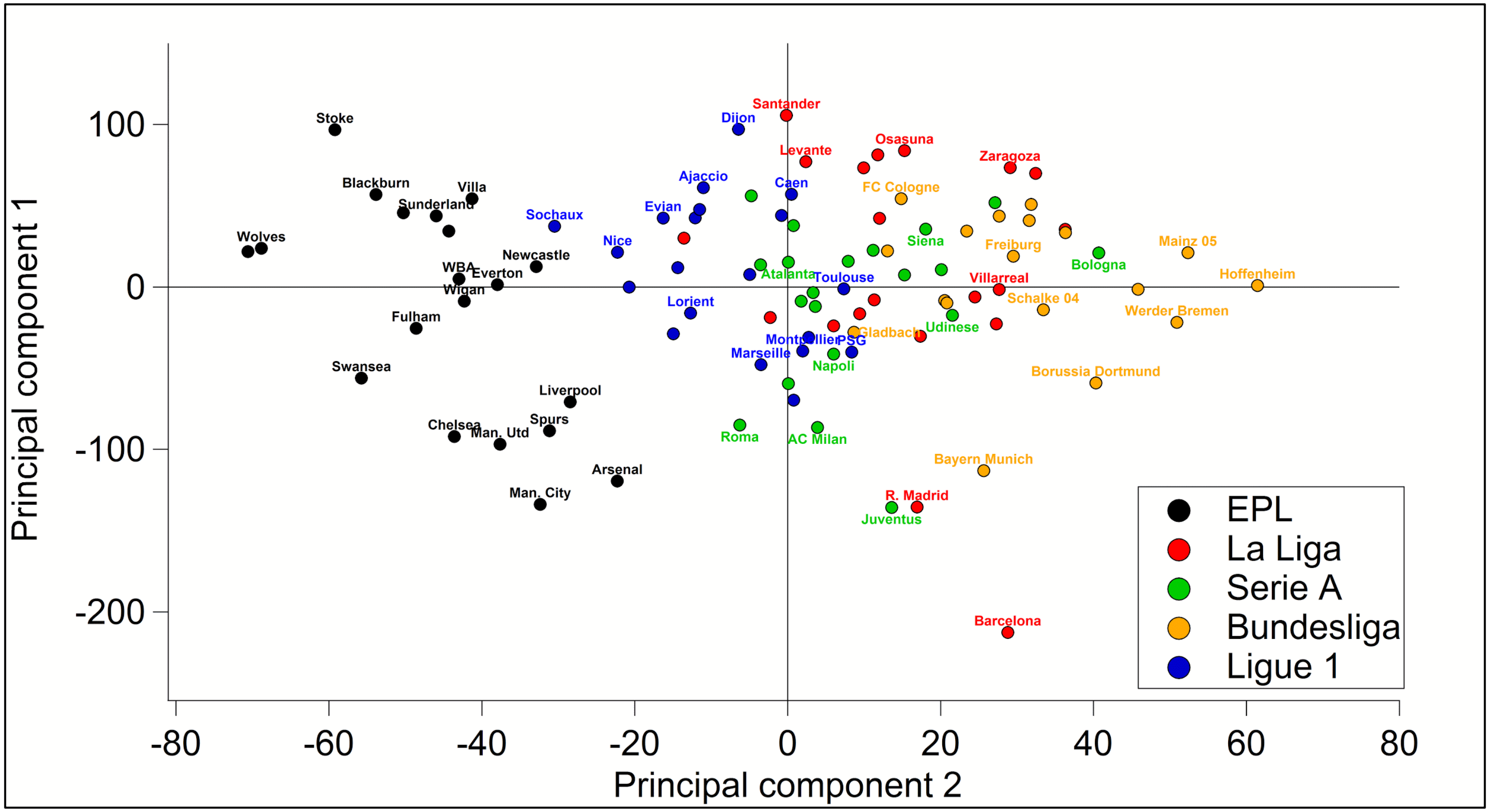


Distance Matrix

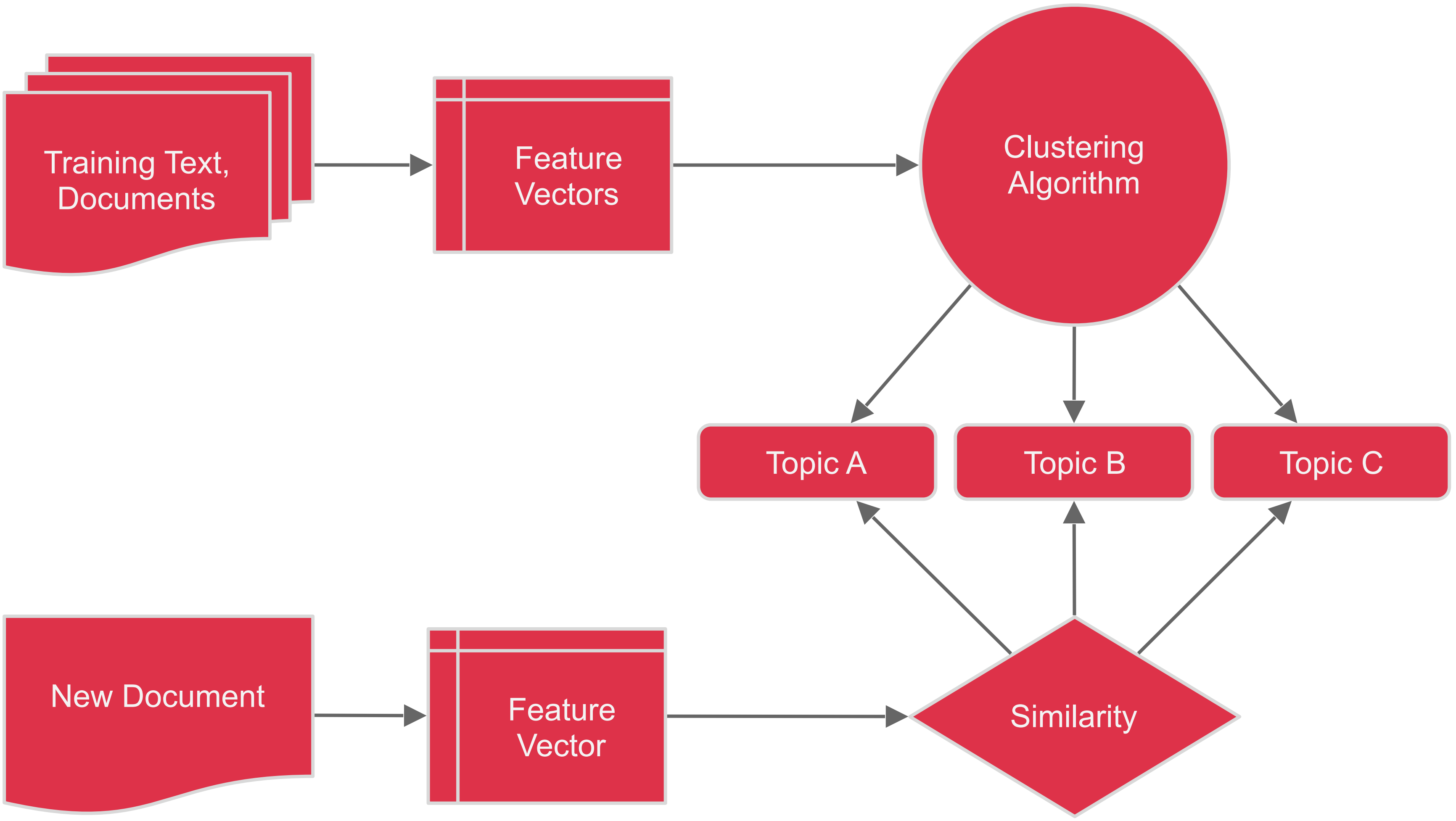


DistrictDataLabs

Principal Component Analysis (PCA)



Topic Modeling Pipeline



Clustering at Scale

Spark MLlib: Clustering

- K-means
- Gaussian mixture
- Power iteration clustering (PIC)
- Latent Dirichlet allocation (LDA)
- Bisecting k-means
- Streaming k-means