Yohanes Dimas Pratama

A11.2021.13254 - A11.4705

Klastering Dokumen dengan K-Means

1. Preprocessing

* D1: PSIS berburu juara Liga Indonesia

Tokenization: ['PSIS', 'berburu', 'juara', 'Liga', 'Indonesia']

Stemming: ['PSIS', 'buru', 'juara', 'Liga', 'Indonesia']

* D2: Hasil putusan Sidang Elit Politik

Tokenization: ['Hasil', 'putusan', 'Sidang', 'Elit', 'Politik']

Stemming: ['Hasil', 'putus', 'Sidang', 'Elit', 'Politik']

* D3: Partai politik berebut suara

Tokenization: ['Partai', 'politik', 'berebut', 'suara']

Stemming: ['Partai', 'politik', 'rebut', 'suara']

* D4: Manchester United Juara Liga Inggris

Tokenization: ['Manchester', 'United', 'Juara', 'Liga', 'Inggris']

Stemming: ['Manchester', 'United', 'Juara', 'Liga', 'Inggris']

* D5: Timnas Indonesia juara Liga AFC

Tokenization: ['Timnas', 'Indonesia', 'juara', 'Liga', 'AFC']

Stemming: ['Timnas', 'Indonesia', 'juara', 'Liga', 'AFC']

1. Term Weighting: TFIDF

* D1: PSIS berburu juara Liga Indonesia

TF-IDF: {'PSIS': 0.0, 'buru': 0.176, 'juara': 0.176, 'Liga': 0.176, 'Indonesia': 0.176}

* D2: Hasil putusan Sidang Elit Politik

TF-IDF: {'Hasil': 0.176, 'putus': 0.176, 'Sidang': 0.176, 'Elit': 0.176, 'Politik': 0.176}

* D3: Partai politik berebut suara

TF-IDF: {'Partai': 0.176, 'politik': 0.0, 'rebut': 0.176, 'suara': 0.176}

* D4: Manchester United Juara Liga Inggris

TF-IDF: {'Manchester': 0.176, 'United': 0.176, 'Juara': 0.176, 'Liga': 0.176, 'Inggris': 0.176}

* D5: Timnas Indonesia juara Liga AFC

TF-IDF: {'Timnas': 0.176, 'Indonesia': 0.176, 'juara': 0.176, 'Liga': 0.176, 'AFC': 0.176}

1. Menggunakan teknik K-Means Clustering

* Dokumen dan Label Cluster:

Dokumen 'PSIS berburu juara Liga Indonesia' dimasukkan ke cluster 1

Dokumen 'Hasil putusan Sidang Elit Politik' dimasukkan ke cluster 0

Dokumen 'Partai politik berebut suara' dimasukkan ke cluster 0

Dokumen 'Manchester United Juara Liga Inggris' dimasukkan ke cluster 1

Dokumen 'Timnas Indonesia juara Liga AFC' dimasukkan ke cluster 1

* Pusat Cluster (TF-IDF):

Cluster 0: [0.18, 0.18, 0.18, 0.18, 0.18]

Cluster 1: [0.18, 0.18, 0.18, 0.18, 0.18]

1. Menentukan Centroid D1 dan D3

* Centroid D1

TF-IDF D1: [0.0, 0.176, 0.176, 0.176, 0.176]

Centroid D1: [0.18, 0.18, 0.18, 0.18, 0.18] (nilai rata-rata)

* Centroid D3

TF-IDF D3: [0.176, 0.0, 0.176, 0.176]

Centroid D3: [0.18, 0.18, 0.18, 0.18] (nilai rata-rata)

1. Melakukan Similarity Measure menggunakan Euclidean Distance

Menggunakan rumus:

Perhitungan jarak Euclidean:

= 0