DOCUMENT CLUSTERING TERKAIT HEALTH NEWS PADA TWITTER DATA SET MENGGUNAKAN K-MEANS CLUSTERING

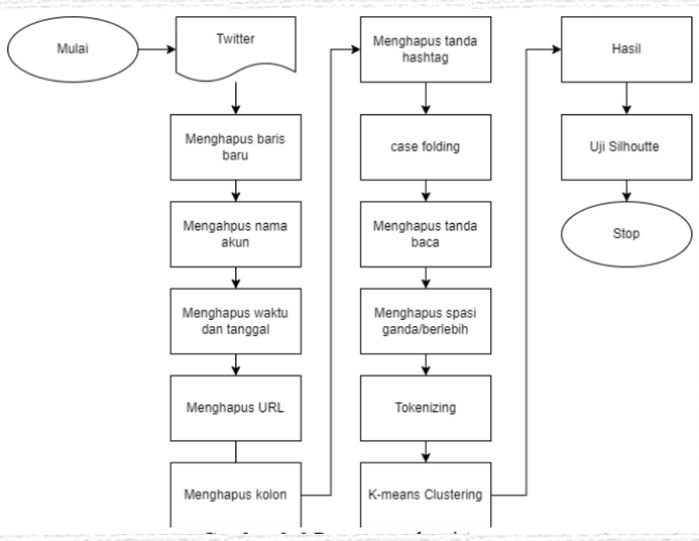
LATAR BELAKANG

Clustering merupakan salah satu metode analisis data yang bertujuan untuk mengelompokkan data-data yang memiliki karakteristik yang mirip pada area yang sama. Dalam konteks pengelompokan dokumen, metode ini digunakan untuk mengelompokkan dokumen berdasarkan karakteristik dan isi dari dokumen itu sendiri. Di akun Twitter, jumlah dokumen medis yang tersedia bisa sangat banyak, sehingga proses pencarian dokumen menjadi lambat dan tidak efisien. Saat ini, hasil pencarian dokumen biasanya ditampilkan berdasarkan klasifikasi dokumen, seringkali tidak menemukan dokumen tertentu. Hal ini dapat menyebabkan kesulitan dalam menemukan dokumen yang relevan dan penting bagi pengguna. Oleh karena itu, diperlukan pendekatan baru untuk meningkatkan akurasi dan efisiensi proses pencarian dokumen. Salah satu pendekatan yang dapat digunakan adalah dengan menggunakan clustering dalam klasifikasi dokumen medik. Dengan menerapkan pendekatan ini, dokumen-dokumen yang memiliki kesamaan karakteristik dapat dikelompokkan menjadi satu sehingga memudahkan pengguna untuk menemukan informasi yang sesuai dan relevan dengan kebutuhannya. Dengan bundling dokumen yang lebih baik, diharapkan proses pencarian dokumen di akun Twitter dapat dilakukan lebih cepat dan efisien. Hal ini secara signifikan akan memberikan kontribusi untuk meningkatkan akurasi penemuan dokumen dan memudahkan pengguna untuk mendapatkan informasi yang mereka butuhkan.

TUJUAN

Dapat menerapkan dan mengetahui performa
K-means Clustering pada Helath News pada Data Set

DIAGRAM ALIR



CLUSTERING

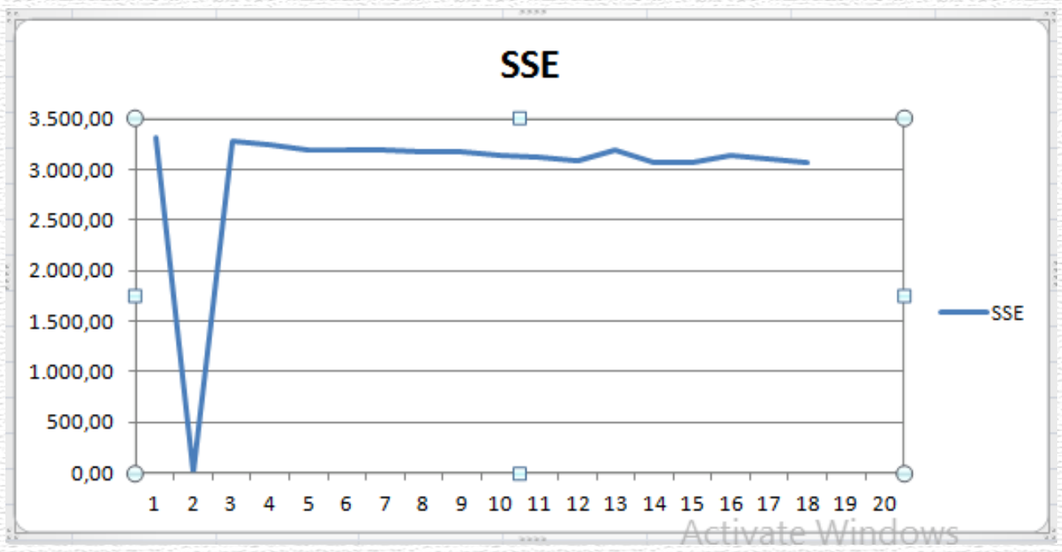
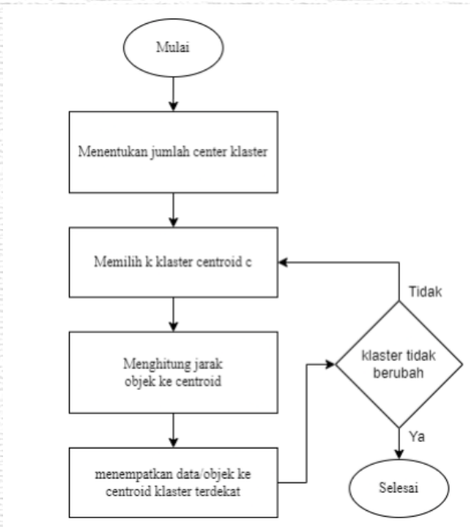


DIAGRAM ALIR K-MEANS



ANALISIS HASIL

Silhouette Score yang rendah menunjukkan bahwa klasterisasi tidak begitu baik. Dalam eksperimen ini, nilai Silhouette Score cenderung rendah dan hampir sama di setiap eksperimen, menunjukkan bahwa klasterisasi mungkin tidak optimal.

Berdasarkan hasil eksperimen, tidak ada nilai K yang memberikan klasterisasi yang sangat baik. Mungkin perlu eksperimen lebih lanjut dengan nilai K yang berbeda atau menggunakan algoritma klasterisasi yang berbeda untuk mendapatkan hasil yang lebih baik. Selain itu, penting untuk menganalisis ciri-ciri dan konteks data tweet yang digunakan untuk memahami mengapa hasil klasterisasi tidak optimal.

KESIMPULAN

Berdasarkan hasil penelitian yang sudah didapatkan, dapat disimpulkan bahwa penerapan Pyspark pada dataset yang dimiliki tidak menunjukan hasil yang lebih baik dikarenakan Nilai evaluasi SSE (Sum of Squared Error) yang bernilai 3288.3146163817337 hingga 3402.023678158529

SARAN

Eksplorasi nilai K yang berbeda, Praproses data yang lebih cermat, seperti menghilangkan kata-kata yang tidak relevan, memperbaiki kesalahan pengejaan, atau mengurangi dimensi data, untuk memperoleh representasi yang lebih baik. Menggunakan metode klasterisasi yang berbeda yang mungkin memberikan hasil yang lebih baik terhadap data tweet tersebut.

DIMAS YUDHISTIRA WIJAYA KOESUMA, ARI HERNAWAN S.KOM., M.SC. RAPHAEL BIANCO HUWAE S.T., M.T.
DEPT INFORMATICS ENGINEERING, MATARAM UNIVERSITY
JL. MAJAPAHIT 62, MATARAM, LOMBOK NTB, INDONESIA
EMAIL: DIMAS.YUDHISTIRA888@GMAIL.COM, ARIHERNAWAN@UNRAM.AC.ID, RAPHAEL.BIANCO.HUWAE@UNRAM.AC.ID