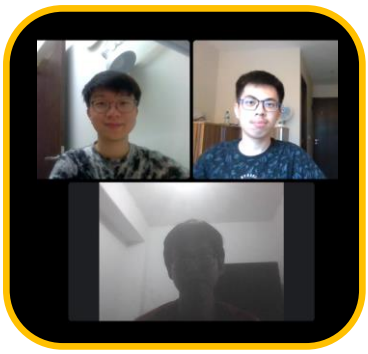


Analisis Pengelompokan Negara Berdasarkan *Folklore* Menggunakan Algoritma K-means

Bryan Ernestin (6162001097), Leonardo Alindra (6162001111), Wristopher (6162001147)

Mathematics Department, Parahyangan University

Abstract: Cerita rakyat (*folklore*) merupakan cerita yang berasal dari zaman dahulu dan diwariskan kepada generasi-generasi selanjutnya secara lisan. Salah satu manfaat dari cerita rakyat adalah untuk melestarikan budaya suatu negara. Pada penelitian ini, akan dilihat persebaran budaya dari negara-negara di benua Asia, Eropa, dan Afrika dengan cara mengelompokkan negara-negara tersebut menjadi beberapa bagian berdasarkan cerita rakyatnya dengan menggunakan *unsupervised machine learning*, yaitu metode *k-means clustering*. Metode *k-means clustering* akan membagi negara–negara tersebut menjadi k kelompok (*cluster*). Untuk mengetahui k yang optimal dalam penerapan *k-means clustering*, akan digunakan metode *elbow* dan metode *silhouette*. Pada penelitian ini, metode *elbow* kurang mampu untuk memberikan informasi k yang optimal, sehingga akan digunakan metode *silhouette*, di mana diperoleh k yang optimal adalah ketika k = 5. Setelah itu, akan dilakukan metode *5-means clustering* untuk membagi negara-negara menjadi 5 *cluster* berdasarkan cerita rakyatnya. Kemudian, akan ditampilkan kata-kata pada cerita rakyat yang paling sering muncul pada setiap *cluster*.beserta *wordcloud*-nya. Lebih lanjut, akan ditampilkan visualisasi dengan menggunakan peta dunia untuk melihat pengelompokkan negara-negara tersebut. Setiap negara yang berada di *cluster* yang sama disimpulkan memiliki kebudayaan yang serupa, dimana hal tersebut mungkin terjadi karena jarak antar negara yang berdekatan atau sejarah yang dialami tidak berbeda jauh.

Keywords: Cerita Rakyat, *Unsupervised Machine Learning, K-means Clustering,* Metode *Elbow,* Metode *Silhouette.*

Lecturer: Dr. Putu Harry Gunawan

1 Pendahuluan

Budaya suatu negara merupakan salah satu hal yang penting untuk mengetahui latar belakang dari suatu negara. Salah satu sarana untuk mengetahui budaya suatu negara adalah melalui cerita rakyat (*folklore*). Cerita rakyat merupakan cerita yang berasal dari zaman dahulu dan diwariskan kepada generasi-generasi selanjutnya secara lisan. Pengarang dari cerita rakyat umumnya tidak diketahui identitasnya. Cerita rakyat terbagi menjadi beberapa macam, yaitu epos, cerita jenaka, pararel, parabel, fabel, legenda, mite, dan sage.

Untuk mengetahui gambaran dari persebaran budaya dari negara-negara di Benua Asia, Eropa, dan Afrika, akan dilakukan pengelompokkan berdasarkan kata-kata yang digunakan pada cerita rakyat. Hal yang dilakukan untuk melakukan pengelompokkan tersebut adalah dengan menggunakan *unsupervised machine learning.* Metode *unsupervised machine learning* yang akan digunakan adalah metode *k-means clustering*, yaitu metode untuk mengelompokkan data menjadi k bagian berdasarkan kesesuaian karakteristik masing-masing objek. Penentuan k yang optimal akan dilakukan dengan memanfaatkan metode *elbow* dan metode *silhouette*[2].

2 Metode dan Data

Dataset yang digunakan dalam penelitian diperoleh melalui proses dokumentasi dari berbagai *website*. Data primer penelitian adalah data kualitatif tentang cerita rakyat (*folklore*) dari 50 negara, sedangkan data sekunder penelitian adalah data nama negara, kebangsaan, *stopwords* dalam bahasa Inggris, dan koordinat dari masing-masing negara. Data primer berupa dataset dengan format csv dan ditunjukkan pada Tabel 2.1.

Tabel 2.1: Dataset *folklore* dari 50 negara di Benua Asia, Eropa, dan Afrika

No	Country	Alpha-2	Alpha-3	Continent	Title	Folklore
0	Armenia	AM	ARM	Europe	Dakhanavar	There once dwelt in a cavern in this country a vampire ...
1	China	CN	CHN	Asia	The Great Race	An ancient folk story tells that Cat (猫)

						and Rat (鼠) …
2	Indonesia	ID	IDN	Asia	Malin Kundang	One day, when Malin Kundang was sailing, he saw a merchant's …
3	Finland	FI	FIN	Europe	Näkki	Näkki is the most well-known water spirit in Finnish mythology. You …
4	Thailand	TH	THA	Asia	Krasue	One of Thailand's most feared ghosts, Krasue was a lady who was …
…	…					
45	Brunei Darussalam	BN	BRN	Asia	The Golden Mountain and the Weeping Rice	Once upon a time, there roamed a mighty mountain covered with …
46	Estonia	EE	EST	Europe	The Young Man Who Would Have His Eyes Opened	Once upon a time there lived a youth who was never happy unless …
47	Timor Leste	TL	TLS	Asia	The Legend of Crocodile	Once upon a time in a land far far away lived a small crocodile. Like many …
48	Malta	MT	MLT	Europe	The Giant and the Bird Hunter	Once upon a time there was a hunter who ventured into the woods and …
49	Turkey	TR	TUR	Asia	The Creation	Allah, the most gracious God, whose dwelling, place is the seventh heaven …

Dataset pada Tabel 2.1 terdiri 50 baris dan 6 kolom, yaitu:

- Country : nama negara
- Alpha-2 : kode 2 huruf negara
- Alpha-3 : kode 3 huruf negara
- Continent : benua
- Title : judul cerita rakyat
- Folklore : isi cerita rakyat

Diambil masing-masing sebuah cerita rakyat dari 50 negara berbeda yang berada pada Benua Asia, Eropa, dan Afrika secara acak. Setiap baris data juga dilengkapi dengan kode Alpha-2 dan Alpha-3 setiap negara yang bersesuaian. Berdasarkan dataset tersebut, akan dilakukan pengelompokkan negara-negara menggunakan metode *k-means clustering*.

2.1 Metode *K-Means*

Metode *k-means* adalah metode dalam analisis *cluster* yang digunakan untuk mengelompokkan data menjadi beberapa klaster berdasarkan kesamaan atribut atau ciri-ciri yang dimiliki. Metode ini termasuk ke dalam *unsupervised machine learning* karena menggunakan dataset yang tidak memiliki variabel target. Tujuan utama dari metode *k-means* adalah meminimalkan variansi di dalam setiap *cluster* dan memaksimalkan perbedaan antara *cluster*.

Berikut adalah prosedur *k-means* secara formal. Pilih *k* buah titik awal secara acak sebagai pusat *cluster*. Lalu, dilakukan serangkaian proses iteratif. Pertama, hitung jarak antara setiap data dengan pusat *cluster* terdekat. Kedua, setiap data akan ditempatkan ke dalam *cluster* dengan pusat *cluster* terdekat yang meminimalkan jarak antara data dan pusat *cluster* yang sesuai. Terakhir, tentukan pusat *cluster* yang baru untuk setiap *cluster* berdasarkan data yang sudah ditempatkan di dalamnya dengan mengambil rata-rata dari data dalam klaster tersebut. Jika tidak ada lagi perubahan pusat klaster maupun penempatan data, maka proses *k-means* selesai.

2.2 Metode *Elbow*

Metode *elbow* adalah sebuah metode yang digunakan dalam analisis *cluster* untuk menentukan jumlah *cluster* yang optimal dalam sebuah data. Metode ini didasarkan pada pengamatan bahwa penambahan jumlah *cluster* yang lebih banyak akan mengurangi jumlah variansi di dalam setiap *cluster*. Namun, dengan bertambahnya jumlah *cluster*, manfaatnya akan berkurang secara signifikan. Metode *elbow* membantu pemilihan jumlah *cluster* yang optimal dengan mempertimbangkan *trade-off* antara meminimalkan variansi dalam *cluster* dan mencegah terlalu banyak klaster yang mungkin tidak bermakna.

Berikut adalah prosedur metode *elbow* secara formal. Lakukan suatu algoritma *clustering* (*k-means*) untuk beberapa nilai banyaknya *cluster* dimulai dari yang terkecil. Untuk setiap *cluster*, dihitung nilai *Sum of Squares Error* (SSE), yaitu jumlah jarak kuadrat antara setiap data dengan pusat *cluster* terdekat. Plot SSE sebagai fungsi dari jumlah klaster yang digunakan. Garis grafik tersebut akan menurun seiring peningkatan jumlah *cluster*. Banyaknya *cluster* yang optimal adalah nilai banyaknya *cluster* yang terletak pada siku atau *elbow point* pada grafik. Pada titik tersebut, penambahan *cluster* lebih lanjut tidak memberikan penurunan SSE yang signifikan.

2.3 Metode Silhouette

Metode *silhouette* adalah salah satu metode yang digunakan untuk mengevaluasi kualitas *clustering* yang dihasilkan oleh algoritma *k-means*. Metode ini memberikan ukuran numerik antara -1 hingga 1 yang menggambarkan sejauh mana setiap data berada dalam *cluster*-nya sendiri dibandingkan dengan *cluster* tetangga terdekatnya. Nilai *silhouette* yang tinggi menunjukkan bahwa data terklasifikasi dengan baik dalam *cluster*-nya dan memiliki jarak yang signifikan dengan *cluster* tetangga, sedangkan nilai yang rendah menandakan bahwa banyak data ditempatkan di *cluster* yang salah atau banyaknya *cluster* belum tepat..

Berikut adalah prosedur metode *silhouette* secara formal. Lakukan suatu algoritma *clustering* (*k-means*) untuk beberapa nilai banyaknya *cluster* dimulai dari yang terkecil. Untuk setiap data dalam *cluster*, hitung *a*, yaitu rata-rata jarak antara data tersebut dengan semua data lain yang ada di dalam *cluster*-nya dan *b*, yaitu rata-rata jarak antara data tersebut dengan semua data dalam *cluster* tetangga terdekat. Lalu, hitung koefisien *silhouette* dari setiap data dengan menggunakan rumus:

$$\frac{b - a}{max(a, b)}$$

Hitung rata-rata koefisien *silhouette* dari setiap data. Banyaknya *cluster* yang optimal adalah banyak *cluster* yang memiliki rata-rata koefisien *silhouette* tertinggi.

2.4 Pemrosesan Data

Langkah pertama yang dilakukan adalah membaca dataset dalam bentuk csv menggunakan *pandas*. Semua data pada kolom “folklore” akan digabungkan ke dalam sebuah *list*, yaitu corpus. Selanjutnya, dilakukan pembersihan dan penyederhanaan data. Tujuan dari proses ini adalah membuat *clustering* menjadi lebih efisien dan efektif karena data sudah memiliki format yang lebih teratur.

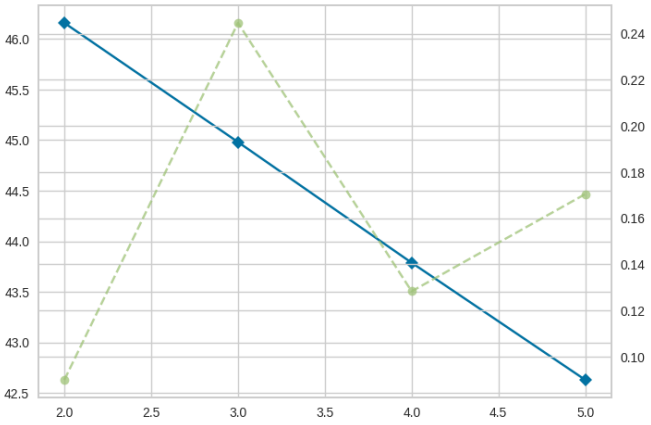
Pembersihan data dimulai dengan menghilangan kata-kata yang tidak berperan penting pada proses *clustering*. Himpunan kata-kata tersebut tersimpan pada *file stopwords*. Selanjutnya, dilakukan proses *stemming* untuk menyederhanakan kata ke dalam bentuk dasarnya. Berbagai angka, tanda baca, dan *link* juga dihapus dari dataset.

Dataset yang sudah bersih diberikan bobot menggunakan fungsi TF-IDF(*Term Frequency - Inverse Document Frequency*). Fungsi ini memetakan setiap kata ke suatu ukuran numerik [0,1] menggunakan frekuensi kata yang menunjukkan seberapa relevan kata tersebut terhadap dokumen.

Setelah itu, negara-negara dikelompokkan menggunakan *k-means* dengan nilai *k* yang bergerak mulai dari 2 hingga suatu nilai tetap. Banyaknya *cluster* (*k*) yang optimal diperoleh melalui metode *elbow* dan melihat nilai *silhouette* terbesar.

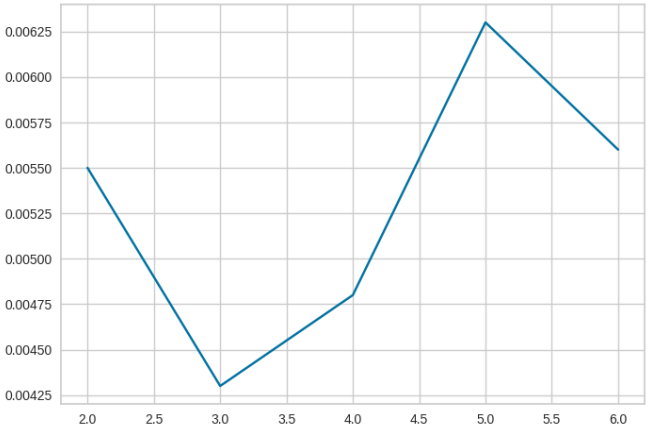
Misalkan diperoleh banyaknya *cluster* terbaik adalah *k**. Kemudian, akan ditunjukkan kata terbanyak dan *wordcloud* di setiap *cluster*. Pada *wordcloud*, ketebalan dan ukuran merepresentasikan frekuensi muncul suatu kata. Semakin sering suatu kata muncul, kata tersebut akan semakin tebal dan besar. Sebagai visualisasi akhir, akan ditampilkan peta dunia dengan *k** warna yang merepresentasikan *k* cluster* berbeda.

3 Analisis Hasil



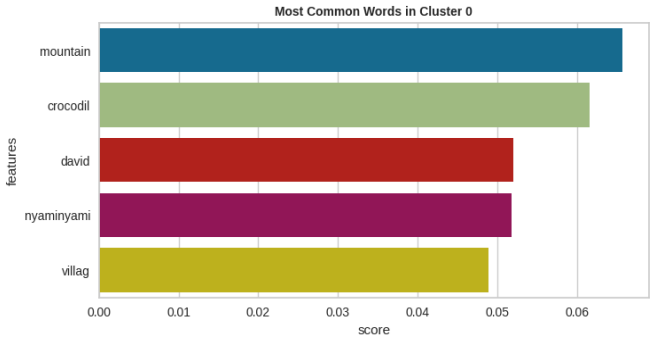
Gambar 3.1: Metode *elbow*

Dengan menggunakan metode *elbow*, sangat sulit untuk melihat berapa *cluster* yang paling optimal untuk mengelompokkan dataset *folklore*. Oleh karena itu, akan digunakan metode *silhouette* untuk melihat banyaknya *cluster* yang paling optimal dalam pengelompokkan dataset *folklore*, khususnya untuk penerapan algoritma *k-means clustering*.

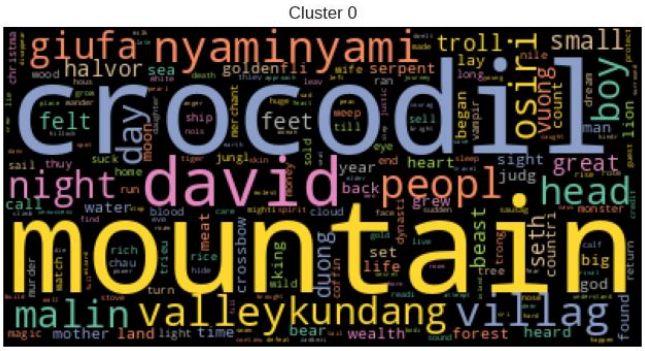


Gambar 3.2: Metode *silhouette*

Dapat dilihat pada gambar diatas, dengan menggunakan metode silhouette, rata-rata nilai *cluster* tertinggi terjadi ketika dipilih k = 5. Artinya, k yang paling optimal adalah 5, yaitu ketika dataset *folklore* dikelompokkan menjadi 5 *cluster*.

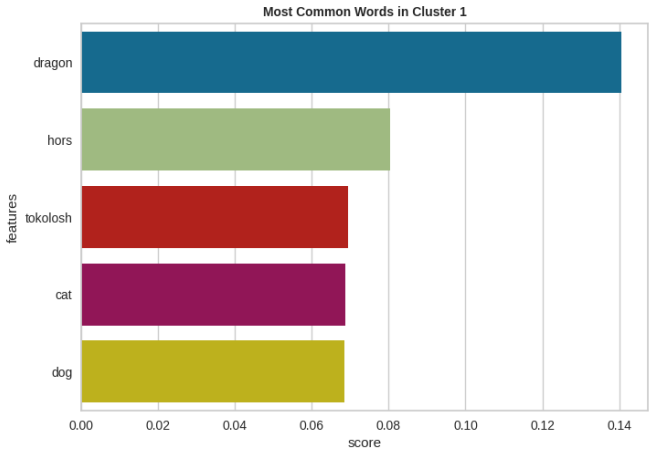


Gambar 3.3: Top 5 kata frekuensi terbanyak pada *cluster* 0

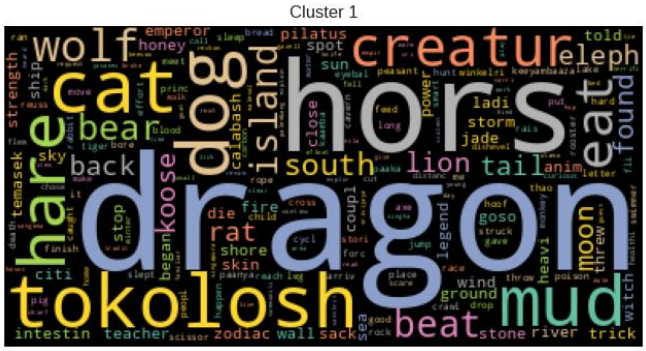


Gambar 3.4: *Wordcloud cluster* 0

Gambar 3.3 dan 3.4 menampilkan kata-kata dengan frekuensi paling banyak pada *cluster* 0. Pada *horizontal bar chart*, terlihat bahwa 5 kata yang paling sering muncul adalah “mountain”, “crocodil”, “david”, “nyaminyami”, dan “villag”. Hal tersebut juga terlihat pada *wordcloud* dimana kelima kata sebelumnya memiliki ketebalan dan ukuran yang paling besar. Dengan demikian, *cluster* 0 beranggotakan negara-negara dengan cerita rakyat tentang kehidupan di pinggir kota.

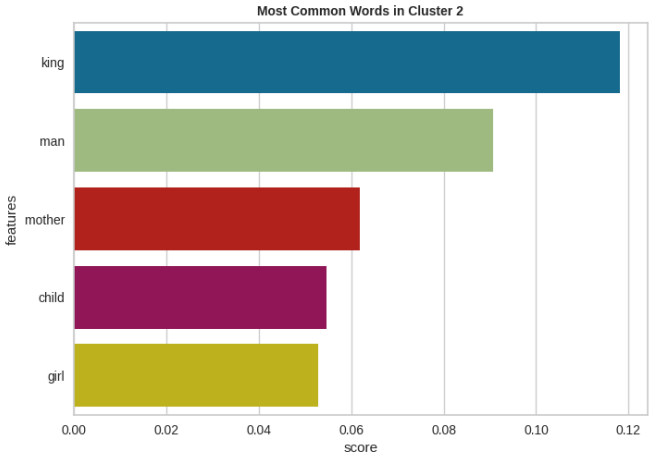


Gambar 3.5: Top 5 kata frekuensi terbanyak pada *cluster* 1

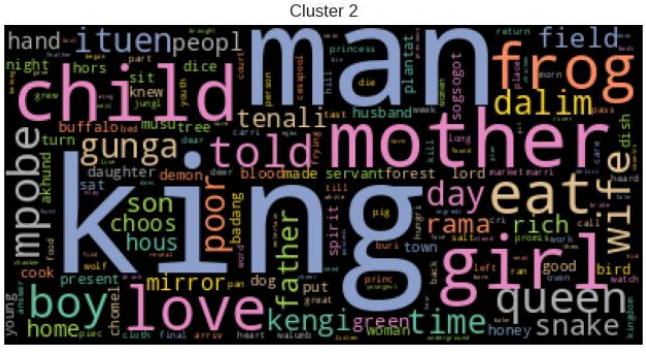


Gambar 3.6: *Wordcloud cluster* 1

Gambar 3.5 dan 3.6 menampilkan kata-kata dengan frekuensi paling banyak pada *cluster* 1. Pada *horizontal bar chart*, terlihat bahwa 5 kata yang paling sering muncul adalah “dragon”, “hors”, “tokolosh”, “cat”, dan “dog”. Hal tersebut juga terlihat pada *wordcloud* dimana kelima kata sebelumnya memiliki ketebalan dan ukuran yang paling besar. Berdasarkan kata-kata dengan frekuensi terbesar, *cluster* 1 didominasi oleh negara-negara dengan cerita rakyat berbentuk fabel yang bercerita tentang kehidupan binatang yang berperilaku seperti manusia.



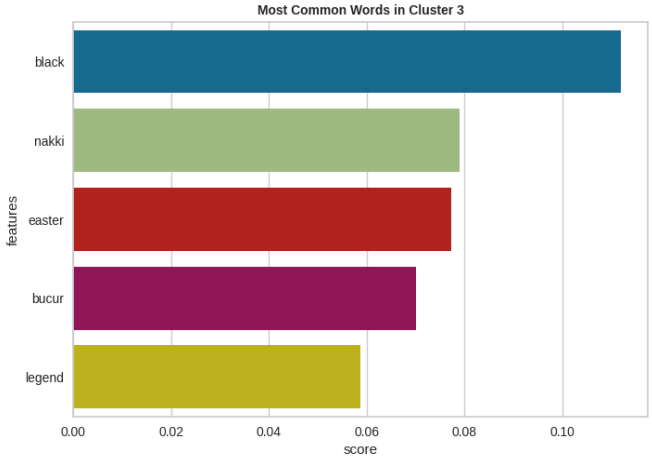
Gambar 3.7: Top 5 kata frekuensi terbanyak pada *cluster* 2



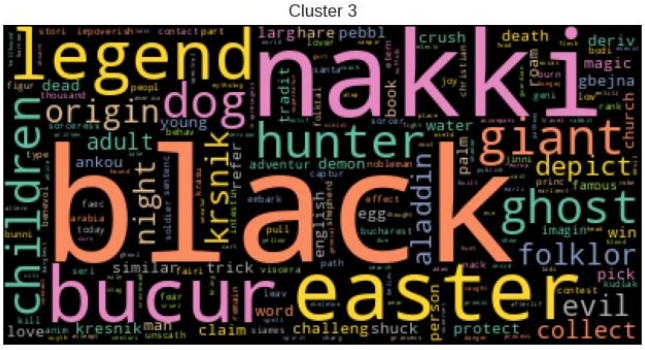
Gambar 3.8: *Wordcloud cluster* 2

Gambar 3.7 dan 3.8 menampilkan kata-kata dengan frekuensi paling banyak pada *cluster* 2. Pada *horizontal bar chart*, terlihat bahwa 5 kata yang paling sering muncul adalah “king”, “man”, “mother”, “child”, dan “girl”. Hal tersebut juga terlihat pada *wordcloud* dimana kelima kata sebelumnya memiliki ketebalan dan ukuran yang paling besar. Kata-kata

yang paling sering muncul menunjukkan bahwa *cluster* 2 terdiri dari banyak negara dengan cerita rakyat tentang kehidupan kerajaan dan keluarga.

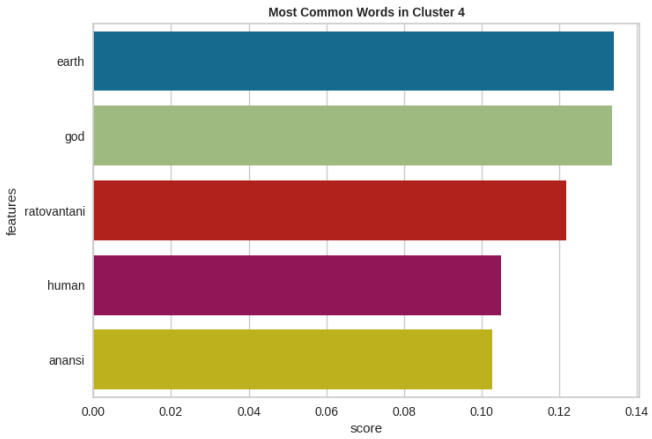


Gambar 3.9: Top 5 kata frekuensi terbanyak pada *cluster* 3

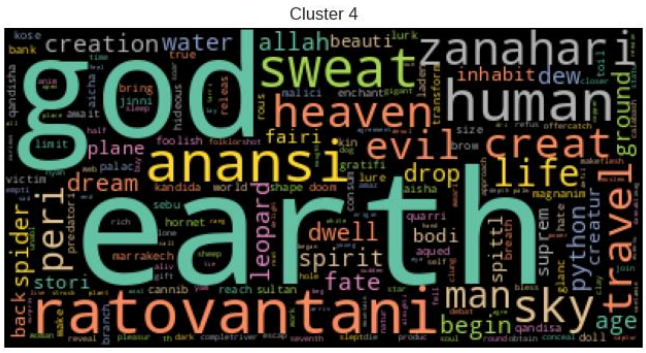


Gambar 3.10: *Wordcloud cluster* 3

Gambar 3.9 dan 3.10 menampilkan kata-kata dengan frekuensi paling banyak pada *cluster* 3. Pada *horizontal bar chart*, terlihat bahwa 5 kata yang paling sering muncul adalah "black", "nakki", "easter", "bucur", dan "legend". Hal tersebut juga terlihat pada *wordcloud* dimana kelima kata sebelumnya memiliki ketebalan dan ukuran yang paling besar. Hal ini menunjukkan bahwa negara-negara dengan cerita rakyat mengenai perayaan dan legenda mendominasi *cluster* 3.

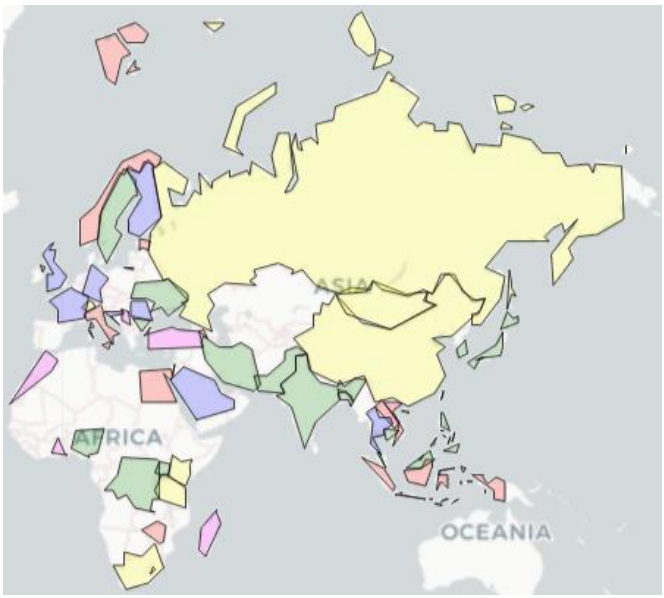


Gambar 3.11: Top 5 kata frekuensi terbanyak pada *cluster* 4



Gambar 3.12: *Wordcloud cluster* 4

Gambar 3.11 dan 3.12 menampilkan kata-kata dengan frekuensi paling banyak pada *cluster* 4. Pada *horizontal bar chart*, terlihat bahwa 5 kata yang paling sering muncul adalah "earth", "god", "ratovantani", "human", dan "anansi". Hal tersebut juga terlihat pada *wordcloud* dimana kelima kata sebelumnya memiliki ketebalan dan ukuran yang paling besar. Berdasarkan kelima kata dengan frekuensi terbesar, dapat diambil kesimpulan bahwa *cluster* 4 paling banyak berisikan negara-negara dengan cerita rakyat religius yang menceritakan hubungan dewa dan manusia.



Gambar 3.13. Hasil *clustering* menggunakan metode *k-means*

Gambar 3 memberikan visualisasi terkait hasil *clustering* berdasarkan cerita rakyat dengan menggunakan metode *k-means*. Negara-negara yang berada pada *cluster* yang sama akan diberikan warna yang sama. Misalnya Indonesia, Timor Leste, Vietnam, Mesir, Zimbabwe, dan Italia berada pada *cluster* 0, sehingga akan diberikan warna merah, dan seterusnya. Negara yang berada di *cluster* yang sama menunjukkan terdapat banyak kata-kata yang sama dari cerita rakyat negara-negara tersebut.

Terdapat banyak kata-kata yang sama pada cerita rakyat, artinya kebudayaan yang dimiliki negara tersebut juga sama. Rusia, Mongolia, China, dan Taiwan berada pada *cluster* yang

sama, yaitu *cluster* 1. Dapat dilihat bahwa keempat negara tersebut memiliki letak yang berdekatan satu sama lain, Keempat negara tersebut juga masih atau pernah menganut ideologi komunisme. Maka dari itu, semua faktor tersebut tercermin pada cerita rakyat yang diceritakan secara turun temurun.

4 Kesimpulan dan Saran

Budaya suatu negara dapat direpresentasikan dari cerita rakyat yang ada, sehingga jika terdapat beberapa negara yang memiliki cerita rakyat yang mirip maka negara-negara tersebut memiliki budaya yang mirip juga. Dengan menggunakan metode *k-means*, 50 negara dari dataset dapat dibagi menjadi 5 *cluster*. Setiap negara yang berada di *cluster* yang sama, artinya kebudayaan negara-negara tersebut juga serupa. Hal ini mungkin terjadi karena memiliki jarak antar negara yang berdekatan atau dikarenakan sejarah yang dialami tidak berbeda jauh.

Saran yang dapat penulis berikan untuk penelitian lebih lanjut adalah memperbanyak negara dari ketiga benua agar hasil *clustering* semakin akurat. Kemudian agar hasil semakin baik, cerita rakyat yang dikumpulkan dari setiap negara sebaiknya lebih dari satu sehingga informasi tentang budaya dari negara tersebut tidak hanya disimpulkan dari satu cerita rakyat.

5 Daftar Pustaka

1. Sarkar, D., Bali, R., dan Sharma, T. (2018). Practical Machine Learning with Python. Apress, New York.
2. Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). Effect of distance metrics in determining k-value in k-means clustering using *elbow* and silhouette method. In Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019) (pp. 341-346). Atlantis Press.