Text Processing Deskripsi Ibukota Negara Asia, Eropa, dan Afrika

Nama Kenneth Holivianto (6162001018), Farrel Valentio (6162001194)

Mathematics Department, Parahyangan University

Abstract: Ibukota merupakan kota dimana pemerintahan pusat berada dan sebagai ikon atau entitas nasional dalam suatu negara. Tentunya setiap negara pasti memiliki ibukota sebagai pusat administrasi, politik, perekonomian, dan kebudayaan negara tersebut. Pada penelitian ini, akan dilihat persebaran ibukota di setiap negara dari benua Asia, Eropa, dan Afrika yang akan dikelompokkan ke beberapa kelompok *(cluster)* berdasarkan kemiripan teritorial, fungsi bangunan pada ibukota tersebut, sejarah negara yang direpresentasikan sebagai ibukota, dan fungsi paling menonjol dari ibukota tersebut dengan menggunakan metode *k-means clustering*. Metode ini bertujuan untuk mengidentifikasikan setiap *k* klaster yang merupakan *k* kelompok berisi setiap ibukota di berbagai negara yang memiliki kesamaan. Untuk mencari *k* klaster yang optimal, maka akan digunakan metode *elbow* dan metode *silhouette*. Namun, pada penelitian ini, metode *elbow* dinilai kurang mampu memberikan nilai *k* yang optimal, sehingga digunakan metode *silhouette* saja. Jadi, diperoleh nilai *k* yang optimal sebesar. Kemudian, akan ditampilkan *wordcloud* dan grafik untuk kata yang paling sering muncul dari masing-masing klaster. Selain itu, juga akan ditampilkan visualisasi berupa peta geografik yang berisi negara-negara dengan warna yang berbeda sesuai dengan klasternya. Setiap negara dengan klaster yang sama memiliki keunikan yang sama juga dari gambaran ibukota negara-negara tersebut dalam berbagai aspek.

Keywords: Ibukota *(capital city), k-means,* Metode *Elbow,* Metode *Silhouette*

Lecturer Dr. Putu Harry Gunawan

1 Pendahuluan

Setiap negara tentu memiliki sebuah ibukota di mana pusat pemerintahan berada. Setiap ibukota memiliki keunikannya masing-masing namun tetap ada beberapa kemiripan satu dengan yang lainnya. Oleh karena itu, penulis ingin mencari tahu kemiripan-kemiripan yang ada dalam suatu kelompok ibukota dan apa yang secara umum mendefinisikan kelompok tersebut.

Penulis mengambil beberapa negara di benua Asia, Eropa, dan Afrika lalu daftarkan nama ibukotanya dan deskripsi singkat mengenai ibukota tersebut melalui situs ”Britannica”. Kemudian, akan dikelompokkan negara-negara tersebut berdasarkan kemiripan da akan dilakuukan visualisasi agar bisa lebih terlihat apa karakteristik setiap kelompok secara umum.

2 Metode dan Data

2.1 Data

Data yang digunakan adalah data 86 negara berbeda dari benua Asia, Eropa, atau Afrika (Britannica, 2023). Informasi yang terkandung dalam data adalah nama negara, nama ibukota, kode Alpha-3, nama benua, dan deskripsi singkat mengenai negara tersebut yang dikutip dari situs ”Britannica”, lalu disajikan dalam bentuk CSV.

Tabel 1. Contoh dataset deskripsi ibukota negara Asia, Eropa, dan Afrika

No	country	capital	alpha-3	continent	teritory
0	Afghanistan	Kabul	AFG	Asia	Kabul, Persian Kābol, city, capital of the province of Kabul and of Afghanistan. …
1	Albania	Tirana	ALB	Europe	Tirana, Albanian Tiranë, city, capital of Albania. It lies …
2	Algeria	Algiers	DZA	Africa	Algiers, French Alger, Arabic Al-Jazāʾir, capital and chief seaport of Algeria. It is the political, …
3	Andorra	Andorra la Vella	AND	Europe	Andorra la Vella, (Catalan: “Andorra the Old”) , French Andorre la Vieille, Spanish Andorra la Vieja, …
4	Angola	Luanda	AGO	Africa	Luanda, also spelled Loanda, formerly São Paulo de Luanda, city, capital of Angola. Located on the Atlantic

					coast of …
…	…	…	…	…	…
79	Slovenia	Ljubljana	SVN	Europe	Ljubljana, German Laibach , Italian Lubiana, capital city and economic, political, and cultural centre of Slovenia, …
80	South Korea	Seoul	KOR	Asia	Seoul, Korean Sŏul, formally Sŏul-t'ŭkpyŏlsi ("Special City of Seoul"), city and capital of South Korea (the Republic of Korea). …
81	Spain	Madrid	ESP	Europe	Madrid, city, capital of Spain and of Madrid provi ncia (province). Spain's arts and financial centre, …
82	Sweden	Stockholm	SWE	Europe	Stockholm, capital and largest city of Sweden. Stockholm is located at the junction of Lake Mälar (Mälaren) and Salt Bay (Saltsjön), …
83	Switzerland	Bern	CHE	Europe	Bern, also spelled Berne, ci ty, capital of Switzerland a nd of Bern canton, in the west-central part of the country. …

2.2 *K-Means*

Metode *k-means* merupakan salah satu teknik *clustering analysis* yang bertujuan untuk mengidentifikasikan setiap klaster yang merupakan kelompok berisi titik – titik data yang memiliki kesamaan berdasarkan pengukuran hasil *euclidean-based distance* atau *correlation-based distance*. Metode ini memiliki tujuan utama untuk memaksimumkan perbedaan antar klaster dan meminimumkan variansi di dalam setiap klaster.

Algoritma dari *k-means* mencoba membuat partisi dari dataset ke dalam klaster dimana setiap titik datanya hanya terdapat pada satu klaster saja agar tidak terjadi *overlapping.* Pertama, mendefinisikan *k* sebagai banyaknya klaster, kemudian menginisiasi *centroids* untuk dataset dan memilih secara acak *k* ke dalam *centroids* tanpa pengulangan. Hal ini bertujuan untuk menempatkan setiap data ke dalam klaster dengan *centroid* terdekat. Terakhir, menghitung rata – rata dari setiap data dalam klaster untuk menentukan *centroid* dari klaster baru.

2.3 Metode *Elbow*

Metode *elbow* merupakan metode yang digunakan untuk menampilkan grafik *within-cluster sum of squares* untuk menentukan nilai *k*, *k* adalah jumlah klaster optimal (Tomar, 2022). Grafik bisa berbentuk "*elbow*" atau lengan, lalu nilai k yang dipilih adalah di mana sendi siku terbentuk dalam grafik tersebut.

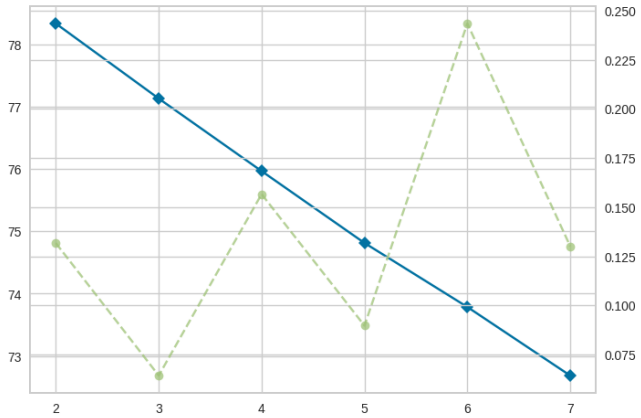
2.4 Metode *Silhouette*

Silhouette score merupakan metode untuk mengukur perbedaan setiap klaster antara rata-rata jarak observasi dengan klaster di dalamnya (*a*) dan rata- rata jarak observasi dengan klaster terdekatnya (*b*) (Bhardwaj, 2020). Nilai dari *silhouette score* berkisar -1 sampai 1, semakin mendekati 1, artinya klaster tersebut memiliki persebaran yang cukup baik dalam menjelaskan observasi-observasinya. Dan ketika nilai *silhouette score* mendekati -1, artinya klaster tersebut belum bisa menjelaskan observasi yang ada. Namun ketika *silhouette score* bernilai mendekati 0, artinya klaster kemungkinan mengalami *overlapping*.

3 Analisis Hasil

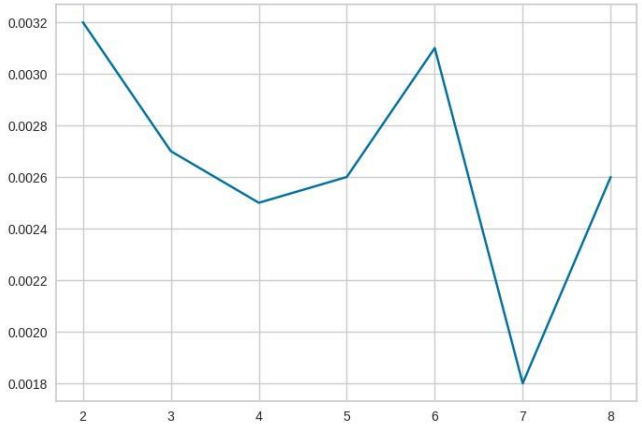
Dari dataset tersebut, akan digunakan bantuan "Google Colab". Pertama-tama akan digunakan metode *Elbow* untuk menentukan jumlah klaster, yaitu kelompok beberapa data, yang optimal sehingga kesimpulan yang bisa ditarik tidak terlalu spesifik maupun terlalu umum.

Gambar 1. Grafik Metode *Elbow*



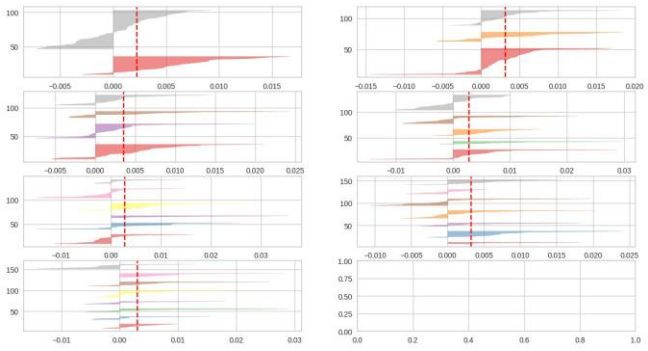
Berdasarkan gambar di atas, metode *Elbow* yang digambarkan kurang menyerupai lengan sehingga sulit untuk ditentukan di mana sendi siku berada. maka itu, tidak ada kesimpulan yang bisa diambil dari hasil metode *elbow* sehingga dibutuhkan metode lain untuk mencari nilai *k* optimal.

Gambar 2. Grafik *silhouette score* dataset



Berdasarkan grafik di atas, 2 klaster merupakan klaster yang paling optimal untuk menjelaskan observasi yang ada. Karena rata-rata *silhouette score* pada 2 klaster sebesar 0.0032 menjadi yang tertinggi daripada rata - rata *silhouette score* pada klaster lainnya. Namun, nilai *k* tersebut kemungkinan mengalami *overlapping*.

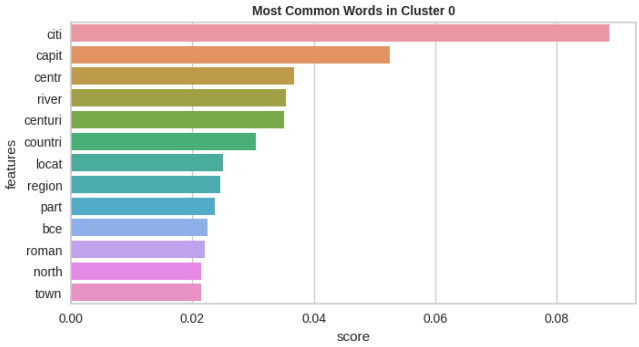
Gambar 3. Grafik *silhouette score* 2-9 jumlah klaster



Berdasarkan plot di atas, dapat dinyatakan bahwa 2 klaster merupakan nilai klaster yang optimal. Hal ini dikarenakan dari kedua klaster tersebut memiliki nilai postif yang cukup besar dan nilai negatif yang cukup kecil dibandingkan dengan klaster - klaster lainnya, sehingga memungkinkan juga mengalami *overlapping*. Ketebalan dari *silhouette score* dari masing - masing klaster merepresentasikan bahwa klaster tersebut saling berdistribusi *uniform*. Terlihat pada 3 klaster, meskipun memiliki ketebalan yang cukup baik antar klaster dan memiliki nilai positif yang cukup besar, namun ada nilai negatif *silhouette score* yang cukup besar diantara klaster tersebut, sehingga akan mengalami *overlapping* yang berlebihan, begitu pula dengan 4 klaster dan 5 klaster. Perhatikan juga pada 6 klaster, 7 klaster, dan 8 klaster, klaster - klaster tersebut memiliki *silhouette score* positif yang cukup besar, namun ketebalan antar klaster tersebut kurang merata serta memiliki *silhouette score* negatif yang cukup besar. Sehingga, meskipun beberapa klaster dapat menjelaskan sebagian observasi dengan baik, klaster - klaster tersebut juga tidak dapat menjelaskan sebagian observasi juga. Dan klaster - klaster tersebut tidak berdistribusi *uniform*. Ketika disimulasikan untuk 9 klaster, terlihat nilai *silhouette score* sudah tidak ada, yang artinya data ini tidak dapat dijelaskan lagi untuk klaster yang lebih besar dari 8. Jadi, jumlah klaster yang optimal untuk menjelaskan data ini berjumlah 2. Meskipun 2 klaster tidak memiliki nilai *silhouette score* yang paling positif, namun dapat meminimalisir *silhouette score* yang negatif, agar tidak terjadi *overlapping* yang berlebihan.

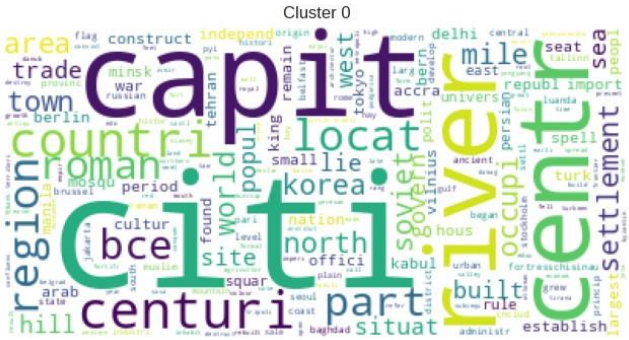
Setelah itu, akan dikelompokkan data ke sejumlah klaster yang telah ditentukan. Lalu, akan dibuat grafik untuk setiap klaster yang menunjukkan kata apa saja yang sering muncul dalam klaster tersebut serta *wordcloud* setiap klaster. Dari grafik ini dan *wordcloud* yang didapatkan akan disimpulkan apa saja kemiripan sekelompok negara dalam satu klaster yang sama secara umum.

Setelah data diklasterkan ke dalam dua klaster yaitu klaster 0 dan klaster 1, akan ditampilkan beberapa kata yang sering muncul dalam suatu klaster.



Gambar 4. Kata yang paling sering muncul dalam klaster 0

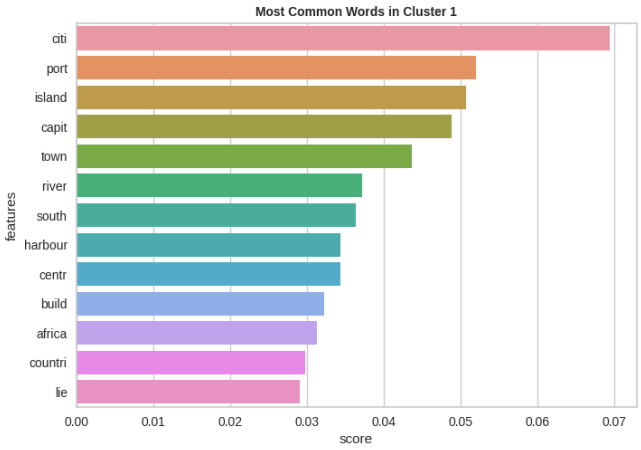
Kata signifikan yang ditampilkan di atas adalah ”river”, ”century”, ”bce”. ”roman”, dan ”north”. Dari lima kata tersebut, bisa disimpulkan bahwa pada umumnya ibukota klaster nol cenderung dekat atau memiliki sungai yang signifikan karena disebutkan dalam deskripsinya. Kemudian, klaster ini cenderung berada di bagian Utara Bumi. Lalu, berdasarkan kata ”century”, ”bce”, dan ”roman”, bisa diartikan bahwa negara di klaster 0 memiliki sejarah yang signifikan yang bisa ditelusuri sampai zaman sebelum masehi yang kemungkinan memiliki hubungan dan pengaruh yang besar dari kerajaan Romawi pada zaman itu.



Gambar 5. *Wordcloud* deskripsi klaster 0

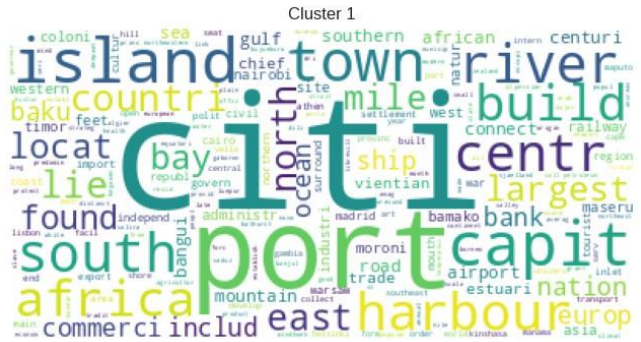
Selain dari kesimpulan dari Gambar 4, *wordcloud* menyajikan beberapa kata lain yang sering muncul. Beberapa kata tambahan yang signifikan adalah ”trade”, yang menunjukkan bahwa klaster 0 memiliki perdagangan yang cukup kuat dan ”hill” yang berarti klaster 0 cenderung

lebih banyak pergunungan. Selain itu, ada beberapa nama negara atau ibukota yang muncul dalam *wordcloud* di atas, diantaranya adalah Tokyo, Baghdad dan Delhi (bisa diasumsikan dari ibukota New delhi), sehingga ketiga ibukota tersebut merupakan contoh ibukota yang masuk ke dalam klaster 0.



Gambar 6. Kata yang paling sering muncul dalam klaster 1

Kata signifikan yang muncul pada gambar di atas adalah ”port”, ”island”, ”river”, ”south”, ”harbour”, dan ”africa”. Kata ”river” sering muncul di klaster 0 dan 1, namun nilai kata tersebut sedikit lebih tinggi di klaster 1 daripada klaster 0, sehingga relevansi adanya sungai sedikit lebih tinggi di klaster 1. Kemudian, klaster 1 cenderung memiliki pelauhan yang dekat atau signifikan, mungkin karena adanya pelabuhan di ibukota dan memiliki aktivitas yang banyak di pelabuhan. Kemudian, klaster 1 juga kemungkinan berbentuk kepulauan atau memiliki pulau yang banyak atau signifikan. Lalu, karena kata ”africa” sering muncul, maka kemungkinan besar negara dari benua Afrika masuk ke dalam klaster 1.



Gambar 7. *Wordcloud* deskripsi klaster 1

Beberapa kata signifikan lainnya dalam klaster 1 adalah ”build”, ”commerci” (dari kata commercial), dan ”east”. Dari ketiga kata tersebut, klaster 1 cenderung memiliki bangunan atau infrastruktur yang signifikan, sangat kuat dalam bidang komersial, dan cukup banyak juga negara yang berada di bagian timur. Beberapa nama ibukota yang muncul di *wordcloud* ini adalah Nairobi, Maseru, dan Moroni, sehingga ketiga ibukota tersebut masuk ke dalam klaster 1.

Gambar 8. Peta geografik untuk setiap klaster

Berdasarkan Gambar 8, diperoleh visualisasi berupa peta geografik yang berisi negara – negara yang berada dalam dari klaster 0 dan klaster 1. Warna merah muda merepresentasikan klaster 0 dan warna ungu muda merepresentasikan klaster 1. Klaster 0 cenderung berisi negara – negara yang berada di Asia dan Eropa. Yang juga dapat diartikan, sebagai negara – negara yang memiliki pengaruh dari negara – negara Eropa, entah pengaruh terhadap pusat pemerintahan, bangunan yang tertinggal dan lain sebagainya. Klaster 1 cenderung berisi negara – negara yang berada di Afrika. Sebagian besar negara – negara tersebut juga berada di dekat laut, yang artinya sektor perekonomian negara pada klaster 1 bertumpu pada kelautan dan perikanan.

4 Kesimpulan

Karakteristik suatu negara dapat direpresentasikan dari pusat pemerintahannya yang berada pada ibukota negara tersebut sebagai kota administrasinya. Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa sebagian besar negara-negara di Asia memiliki kemiripan dengan negara-negara di Eropa berdasarkan teritori dan karakteristik ibukota negaranya. Seperti lokasi ibukota dari negara-negara di kedua benua tersebut juga cenderung terdapat di dataran menegah hingga tinggi dan sebagian besar berada di utara ataupun tengah dari negaranya . Serta pada sektor perekonomiannya yang berhubungan erat dengan perdagangan antar negara-negara tersebut. Berbeda dengan negara-negara di benua Afrika dan sebagian kecil Eropa, yang ibukotanya berada di daerah selatan atau timur dari negaranya dan memiliki lokasi yang berdekatan dengan sungai. Sehingga, dapat dikatakan bahwa negara-negara tersebut memiliki sektor perikanan dan kemaritiman yang kuat sebagai sektor perekonomian negara tersebut.

Saran yang dapat penulis berikan untuk penelitian lebih lanjut adalah memberikan deskripsi lebih untuk masing-masing negara, agar hasil *clustering* lebih akurat dan memiliki *k* klaster yang lebih optimal.

5 Daftar Pustaka

1. Tomar, A. (2022). Stop using elbow method in k-means clustering, instead, use this!. Towards Data Science. https://towardsdatascience.com/elbow-method-is-not-sufficient-to-find-best-k-in-k-means-clustering-fc820da0631d
2. Bhardwaj, A. (2020). Silhouette Coefficient. Towards Data Science. https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c