JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Klasterisasi *Frequently Asked Question* menggunakan *K-means Clustering*

Khaerul Anwar[a1], I Ketut Gede Suhartana[a2]

[a]Program Studi Informatika, Universitas Udayana Bukit Jimbaran, Bali, Indonesia
[1]khaerulanwar2104@gmail.com
[2]ikg.suhartana2@unud.ac.id

Abstract

Frequently asked questions are an important part of providing good service to customers. information provided in the form of questions and answers related to products, applications, companies that are available in detail, concise and easily accessible .The determination of the format of the frequently asked question list should be based on the questions asked by the customer so that they are relevant to the customer's needs. clustering the list of questions using K-means and TF-IDF as the feature extraction method provides an optimal solution of 50000 list questions divided into 18 clusters with a silhoutte coefficients = 00. each cluster is taken 1 document which will be a question in that category provided that the document has at most the term frequncy of the features on the cluster.

Keywords: *Clustering, Frequently Asked Question, K-means,TF-IDF, Information*

1. Pendahuluan

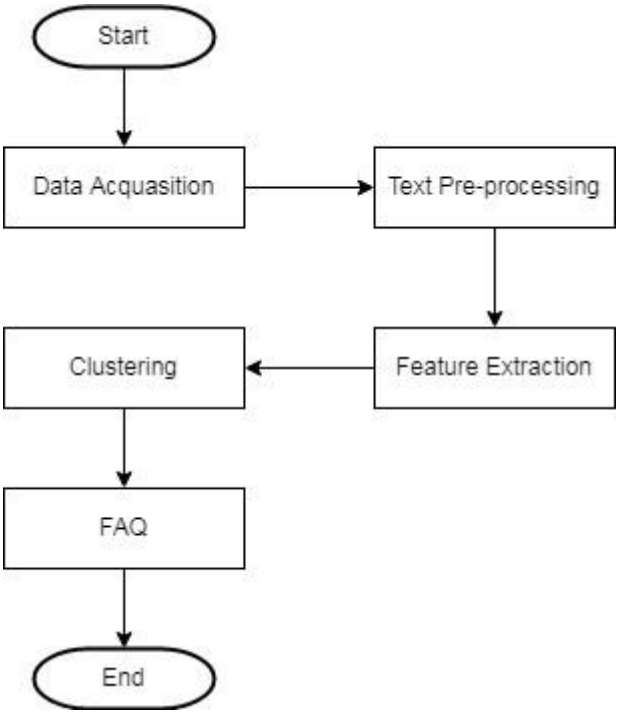
FAQ singkatan dari *frequently asked question* adalah daftar informasi pertanyaan yang sering diajukan pengguna dan jawaban terkait produk, apliaksi, perusahan yang tersedia secara detail, ringkas dan mudah diakses. FAQ tersusun atas topik kompleks dan diatur berdasarkan sub-topik yang berfungsi untuk meminimlakan pertanyaan berulang dari pengguna dari sosial media, *chat*, atau *email*. Selain itu FAQ juga dapat meningkatkan peringkat SEO pada mesin pencari untuk layanan berbasis website. Penentuan daftar pertanyaan yang sering diajukan dibuat berdasarkan asumsi perusahaan yang didapatkan dari tim customer service atau admin.

Penentuan seperti ini tidak sepenuhnya memuat informasi yang relevan dengan kebutuhan pengguna. Oleh karena itu, diperlukan suatu cara dalam penentuan format pertanyaan yang sering diajukan berdasarkan data pertanyaan yang telah diajukan pengguna. Terdapat berbagai macam cara peenntuan salah satunya dengan klasterisasi daftar pertanyaan menggunakan *Kmeans clustering*. Algoritma k-means adalah bagian dari metode non-hierarchical data clustering yang bertujuan untuk membagi-bagi data ke dalam bentuk satu atau lebih kelompok[1]. Metode ini menempatkan data ke dalam kelompok yang memiliki titik pusat terdekat dan mempunyai karakteristik sama ditempatkan pada kelompok yang sama, sedangkan jika data memiliki titik pusat yang jauh dan karakteristik berbeda maka dimasukkan ke kelompok yang berbeda.

Algoritma K-means clustering digunakan pada penenlitian [1] menggunakan 3 cluster mendapatkan *silhouette coefficient* dengan nilai 0,108690751 untuk klasterisasi kinerja akademik mahasiswa. Penelitian [2] membandingkan *K-means* dan DBSCAN dalam klasterisasi data kesehatan menghasilkan *K-means* berkerja lebih baik daripada DBSCAN. Sehingga dapat disimpulkan bahwa algoritma *K-means* dapat menjadi solusi yang optimal dalam menyelesaikan masalah klasterisasi.

2. Metode Penelitian

2.1. Desain Penelitian



Gambar 1. Desain Penelitian

2.2. Pengumpulan Data

Pengumpulan dataset pada penelitian ini menggunakan data *train* pasangan pertanyaan quora yang diperoleh dari website *Kaggle*, diakses pada tanggal 28 September 2022 pukul 09:10:30 WITA dengan alamat akses https://www.kaggle.com/competitions/quora-question-pairs/data?select=train.csv.zip. Data ini terdiri dari 6 kolom dan 404289 baris, namun penenlitian ini hanya menggunakan 50000 baris data dan kolom pertanyaan 1. Detail 5 dataset teratas dapat dilihat pada gambar berikut.

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0

Gambar 2. Detail dataset

2.3. *Pre-processing* Data

Data daftar pertanyaan quora yang telah dikumpulkan akan dipre-processing dengan tahapan *case folding, removing punctuation, tokenizing, stopword removal* dan *stemming*. *Case folding* dilakukan untuk mengubah semua huruf pada data menjadi huruf kecil, proses ini bertujuan untuk membuat kata yang sama jika ditulis dengan huruf berbeda akan menghasilkan nilai yang sama. Setelah semua kata menjadi huruf kecil maka

dilakukan penghapusan karakter-karakter yang tidak termasuk dalam ASCII melalui proses *removing punctuation*. Tokenizing merupakan proses untuk membagi data kalimat pertanyaan pada setiap baris menjadi kata – kata yang terpisah. Stop word removal adalah proses untuk menghapus kata - kata yang dianggap tidak penting dalam penelitian ini menggunakan stopword bahasa inggris maka yang dihapus kata-kata seperti "by", "the", "is", "do" dan lain lain. Proses text pre-processing terakhir adalah stemming yaitu mengembalikan kata menjadi kata dasar.

2.4. Ekstraksi fitur

Pada tahap ini ekstraksi fitur ini mengubah data tekstual menjadi vector yang dapat diamati jarak kedekatan pada proses klasterisasi. Proses ini menggunakan metode *Term Frequency Inverse Document Frequency.* Metode TF-IDF merupakan metode untuk menghitung bobot suatu kata (term) terhadap dokumen[3]. Keunggulan yang dimiliki metode yaitu mudah digunakan, efisien waktu, dan menghasilkan fitur yang akurat. Pada metode ini konsep perhitungan bobot frekuensi kemunculan kata dalam dokumen dan inverse frekuensi dokumen yang memiliki kata tersebut. Frekuensi kemunculan menunjukkan seberapa kuat pengaruh kata tersebut di dalam dokumen. Perhitungan TFIDF menggunakan rumus sebagai berikut.

$$tfidf_{dt} = tf_{dt}\ idf_t \quad (1)$$

Dengan idf_t diperoleh dari $idf_t = \log\left(\frac{N}{df}\right)$ (2)

2.5. Klasifikasi menggunakan *K-means*

K-means clustering dalam melakukan klasterisasi dokumen pertanyaan yang memiliki makna sama menggunakan beberapa tahapan sebagai berrikut: a. Menginisialisasi jumlah *cluster*

b. Menentukan nilai awal pusat *cluster* secara acak
c. Menentukan nilai kedekatan antara *vector* dan pusat *cluster*
d. Menempatkan *vector* ke *cluster* dengan jarak pusat terkecil
e. Menginisialisasi pusat *cluster* baru
f. Menentukan nilai kedekatan anatara *vector* dan pusat *cluster* baru sampai *vector* tidak berpindah *cluster*

2.6. Pengujian

Tahap pengujjian ini digunakan untuk menjadi tolak ukur keberhasilan Metode yang digunakan dalam mengelompokkan data. Dalam pengujian K-means clustering tolak ukur yang digunakan yaitu ketepatan kelompok dan kualitas kelompok. Untuk menentukan ketepatan pengelompokan dan kualitas kelompok menggunakan ketepatan kelompok deret waktu yaitu metode *silhouette coefficient*[1].

Perhitungan *silhoutte coefficient* memiliki rentang nilai -1 sampai 1. Ketepatan pengelompokan dikatan baik jika perhitungan bernilai positif yang menunjukkan data berada pada cluster yang sesuai. Sedangkan jika perhitungan bernilai negatif menunjukkan data berada pada *cluster* yang sesuai sehingga satu data dapat memiliki dua atau lebih *cluster*. Menurut teori Kaufman dan Rousseeuw[1] hasil perhitungan *silhouette coefficient* terbagi menjadi empat jenis yaitu:

1. Sangat Struktur 0,7 - 1
2. Terstruktur 0,5 - 0,7
3. Kurang terstruktur 0,25 - 0,5
4. Tidak terstruktur ≤ 0,25

3. Hasil dan Pembahasan

Pada penelitian ini, tahap awal adalah pre-processing data meliputi *case folding, removing punctuation, tokenizing, stopword removal* dan *stemming.* Hasil pre-processing dapat dilihat pada gambar 3.

	id	question1	Pre-processing
0	0	What is the step by step guide to invest in sh...	step step guid invest share market india
1	1	What is the story of Kohinoor (Koh-i-Noor) Dia...	stori kohinoor kohinoor diamond
2	2	How can I increase the speed of my internet co...	increas speed internet connect use vpn
3	3	Why am I mentally very lonely? How can I solve...	mental lone solv
4	4	Which one dissolve in water quikly sugar, salt...	one dissolv water quikli sugar salt methan car...
...	...	...	...
49995	49995	How do you take the derivative of [math]\frac{...	take deriv mathfracx22math
49996	49996	How much space does Mac OS X Yosemite take on ...	much space mac os x yosemit take new macbook
49997	49997	Why are criterium races pre-arranged and so lu...	criterium race prearrang lucr

Gambar 2. Hasil Preprocessing

Hasil pre-processing mengubah semua kata menjadi huruf kecil, menghilangkan karakterkarakter seperti (), -, ?, menghilangkan kata-kata tidak penting, dan mengembalikan kata ke kata dasar.

Data yang telah melewati proses pre-processing diubah ke vector menggunakan metode *Term Frequency Inverse Document Frequency.* Berikut hasil ekstraksi fitur.

	response1	response2	response3	response4	response5
kohinoor	0.000000	0.824032	0.000000	0.000000	0.0
step	0.711288	0.000000	0.000000	0.000000	0.0
solv	0.000000	0.000000	0.000000	0.595797	0.0
lone	0.000000	0.000000	0.000000	0.595797	0.0
mental	0.000000	0.000000	0.000000	0.538564	0.0
vpn	0.000000	0.000000	0.467608	0.000000	0.0
connect	0.000000	0.000000	0.441332	0.000000	0.0
speed	0.000000	0.000000	0.441332	0.000000	0.0
diamond	0.000000	0.412016	0.000000	0.000000	0.0
internet	0.000000	0.000000	0.396412	0.000000	0.0

Gambar 4. Hasil Ekstraksi fitur

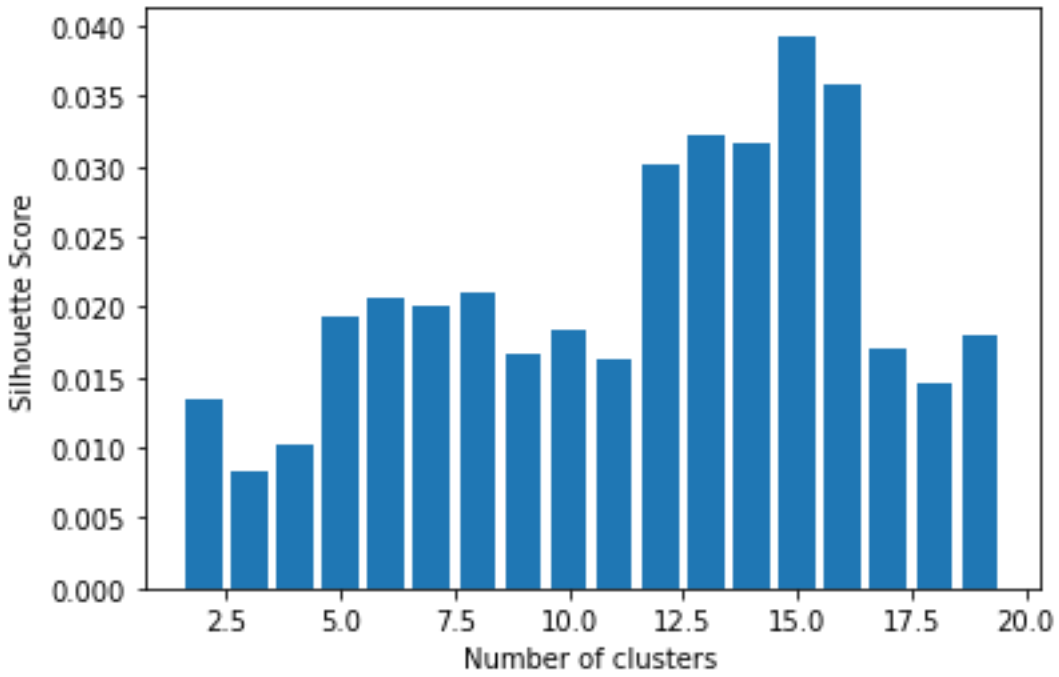
Untuk melakukan klasterisasi tahap awal menentukan jumlah klister, pada penelitian ini menggunakan 18 klaster kemudian melakukan pemodelan menggunakan algoritma K-means clustering. Berikut hasil klasterisasi.

	cluster	terms	jumlah
0	0	im,girl,leav,send,everi,month,ask,feel,faster,...	1
1	1	attend,colleg,like,univers,chines,graduat,pres...	6
2	2	caus,death,earli,reason,problem,reaction,war,w...	36
3	3	best,way,learn,book,india,movi,laptop,institut...	399
4	4	modi,narendra,mr,pm,meet,muslim,india,letter,c...	17
5	5	quora,question,answer,ask,improv,need,peopl,ma...	114
6	6	code,learn,app,start,googl,rule,wrong,creat,wa...	19
7	7	care,fast,black,mean,feel,featur,fear,favourit...	1
8	8	increas,height,concentr,21,traffic,skip,way,we...	31
9	9	expens,japan,cheap,colleg,track,becom,water,da...	6
10	10	good,learn,engin,car,make,compani,score,song,s...	123
11	11	note,1000,500,rupe,ban,rs,currenc,black,indian...	48
12	12	chang,life,time,believ,peopl,year,account,hair...	41
13	13	differ,use,india,peopl,make,life,mean,think,wo...	4023
14	14	like,work,guy,girl,live,feel,person,look,women...	135

Gambar 5. Hasil Klasterisasi

Berdasarkan hasil pada gambar 5, pertanyaan terkait *cluster* 0 hanya ditanyakan sekali, *cluster* 1 ditanyakan 6 kali, *cluster* 2 hanya ditanyakan 36 kali, *cluster* 3 hanya ditanyakan 399 kali, *cluster* 4 hanya ditanyakan 17 kali, *cluster* 5 hanya ditanyakan 114 kali, *cluster* 6 hanya ditanyakan 19 kali, *cluster* 7 hanya ditanyakan sekali, *cluster* 8 hanya ditanyakan 31 kali, *cluster* 9 hanya ditanyakan 6 kali, *cluster* 10 hanya ditanyakan 123 kali, *cluster* 11 hanya ditanyakan 48 kali, *cluster* 12 hanya ditanyakan 41 kali, *cluster* 13 hanya ditanyakan 4023 kali, *cluster* 14 hanya ditanyakan 135 kali.

Tahap akhir adalah pengujian model yang telah dibuat menggunakan metode *silhouette coefficients*, berikut hasil pengujian.



Gambar 6. Hasil Pengujian *Silhoutte Coefficients*

Berdasarkan grafik pengujian menggunakan rentang jumlah cluster antara 2 sampai 20 didapatkan *silhouette coefficients* = 0.04 pada jumlah *cluster* 15.

4. Kesimpulan

Berdasarkan penelitian yang dilakukan, dapat disimpulkan bahwa klasterisasi *frequently asked question* menggunakan algoritma *K-means* dan *TFIDF* mendapatkan silhouette coefficients = 0.04. Hasil klasterisasi tersebut memiliki nilai yang tidak baik karena termasuk dalam kategori *nostructure* atau data-data pada *cluster* masih terjadi overlapping.

References

[1] Aziz dkk, “Implementasi Algoritma K-Means untuk Klasterisasi Kinerja Akademik Mahasiswa”, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2, No. 6, Juni 2018, hlm. 2243-2251

[2] Godwin Ogbuabor, Ugwoke, F. N, “Clustering Algorithm For A Healthcare Dataset Using Silhouette Score Value”, International Journal of Computer Science & Information Technology (IJCSIT), Vol 10, No 2, April 2018.

[3] Ade Riyani , Muhammad Zidny Naf’an , Auliya Burhanuddin, “Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen”, Jurnal Linguistik Komputasional, Vol 2, No 1 Maret 2019.

Halaman ini sengaja dibiarkan kosong