LAPORAN

KLUSTERING PADA MEDIA SOSIAL TWITTER MENGGUNAKAN BAHASA PEMROGRAMAN "R"

Disusun Oleh :

Niko Fitrianto 16.01.63.0028

UNIVERSITAS STIKUBANK SEMARANG (UNISBANK)

FALKUTAS TEKNOLOGI INFORMASI

TEKNIK INFORMATIKA

2017

BAB I
PENDAHULUAN

1.1 Deskripsi Permasalahan

Media sosial adalah sebuah media online, dengan para penggunanya bisa dengan mudah berpartisipasi, berbagi, dan menciptakan isi meliputi twitter, jejaring sosial, wiki, forum dan dunia virtual. Twitter, jejaring sosial dan wiki merupakan bentuk media sosial yang paling umum digunakan oleh masyarakat di seluruh dunia. Andreas Kaplan dan Michael Haenlein, mendefinisikan media sosial sebagai sebuah kelompok aplikasi berbasis internet yang membangun di atas dasar ideologi dan teknologi.

Sebagai makhluk sosial, manusia tidak lepas dari kebutuhan dasar untuk bersosialisasi. Sosialisasi secara umum adalah proses belajar individu untuk mengenal dan menghayati norma-norma serta nilai-nilai sosial sehingga terjadi pembentukan sikap untuk berperilaku sesuai dengan tuntutan atau perilaku masyarakatnya. Salah satu cara bersosialisasi dapat dilakukan melalui komunikasi verbal maupun non verbal dan secara langsung ataupun tidak langsung. Melalui komunikasi antar individu dapat bertukar kabar atau berita yang menghasilkan suatu informasi.

R (juga dikenal sebagai GNU S) adalah bahasa pemrograman dan perangkat lunak untuk analisis statistika dan grafik. R dibuat oleh Ross Ihaka dan Robert Gentleman di Universitas Auckland, Selandia Baru, dan kini dikembangkan oleh R Development Core Team, di mana Chambers merupakan anggotanya. R dinamakan sebagian setelah nama dua pembuatnya (Robert Gentleman dan Ross Ihaka), dan sebagian sebagian dari permainan nama dari S. Bahasa R kini menjadi standar de facto di antara statistikawan untuk pengembangan perangkat lunak statistika, serta digunakan secara luas untuk pengembangan perangkat lunak statistika dan analisis data. R merupakan bagian dari proyek GNU. Kode sumbernya tersedia secara bebas di bawah Lisensi Publik Umum GNU, dan versi biner prekompilasinya tersedia untuk berbagai sistem operasi. R menggunakan antarmuka baris perintah, meski beberapa antarmuka pengguna grafik juga tersedia. R menyediakan berbagai teknik statistika (permodelan linier dan nonlinier, uji statistik klasik, analisis deret waktu, klasifikasi, klasterisasi, dan sebagainya) serta grafik. R, sebagaimana S, dirancang sebagai bahasa

komputer sebenarnya, dan mengizinkan penggunanya untuk menambah fungsi tambahan dengan mendefinisikan fungsi baru. Kekuatan besar dari R yang lain adalah fasilitas grafiknya, yang menghasilkan grafik dengan kualitas publikasi yang dapat memuat simbol matematika. R memiliki format dokumentasi seperti LaTeX, yang digunakan untuk menyediakan dokumentasi yang lengkap, baik secara daring (dalam berbagai format) maupun secara cetakan.

Clustering adalah metode penganalisaan data, yang sering dimasukkan sebagai salah satu metode Data Mining, yang tujuannya adalah untuk mengelompokkan data dengan karakteristik yang sama ke suatu 'wilayah' yang sama dan data dengan karakteristik yang berbeda ke 'wilayah' yang lain. Ada beberapa pendekatan yang digunakan dalam mengembangkan metode clustering. Dua pendekatan utama adalah clustering dengan pendekatan partisi dan clustering dengan pendekatan hirarki. Clustering dengan pendekatan partisi atau sering disebut dengan partition-based clustering mengelompokkan data dengan memilah-milah data yang dianalisa ke dalam cluster-cluster yang ada. Clustering dengan pendekatan hirarki atau sering disebut dengan hierarchical clustering mengelompokkan data dengan membuat suatu hirarki berupa dendogram dimana data yang mirip akan ditempatkan pada hirarki yang berdekatan dan yang tidak pada hirarki yang berjauhan. Di samping kedua pendekatan tersebut, ada juga clustering dengan pendekatan automatic mapping (Self-Organising Map/SOM).

Twitter adalah jejaring sosial berupa blog ukuran kecil yang didirikan oleh Jack Dorsey pada bulan Maret 2006. Melalui Twitter pengguna dapat mengirim dan membaca pesan, berbagi informasi, menjalin relasi bisnis, menuangkan isi hati dan pikiran dalam bentuk tulisan (sering disebut tweet), dengan kapasitas kata yang bisa diunggah dan ditampilkan pada timeline penggunatwitter mencapai 140 karakter. Sama halnya dengan situs jejaring sosial lain dalam Twitter disediakan suatu mesin pencarian (search engine) yang berguna untuk mempermudah pengguna dalam menemukan informasi menggunakan kata kunci. Melalui search engine pengguna dapat menemukan lebih banyak informasi yang dibutuhkan terkait topik yang ingin dicari, yaitu lebih dari satu akun yang ada di twitter. Twitter sebagai hasil dari perkembangan teknologi informasi memungkinkan setiap waktu untuk menghasilkan kumpulan data yang banyak, dimana setiap detiknya pada saat kehidupan normal rata-

rata jumlah tweet yang ada dalam twitter adalah 600 tweet. Hal tersebut tidak berlaku jika suatu waktu terjadi peristiwa-peristiwa tertentu yang menyebabkan peningkatan atau penurunan rata-rata jumlah tweet perdetiknya. Dengan adanya kumpulan data yang terus meningkat setiap waktunya yaitu berupa data tweet perlu dilakukan suatu penangan menggunakan metode khusus untuk mengalisis data pada twitter sehingga menghasilkan suatu informasi yang bermanfaat.

K-Means adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode k-means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas maka rumusan masalahnya adalah bagaimana penjelasan mengenai Klustering K-means dan implementasinya menggunakan bahasa “R” dengan menggunakan data yang disediakan pada media sosial yaitu twitter.

BAB II
TINJAUAN PUSTAKA

1.1 Kajian Deduktif

K-Means adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode k-means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Data Clustering merupakan salah satu metode *Data Mining* yang bersifat tanpa arahan (*unsupervised*). Ada dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) data clustering dan *non-hierarchical* (non hirarki) data clustering. *K-Means* merupakan salah satu metode data clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster/kelompok. Metode ini mempartisi data ke dalam cluster/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan dari data clustering ini adalah untuk meminimalisasikan *objective function* yang diset dalam proses clustering, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu cluster dan memaksimalisasikan variasi antar cluster.

Data clustering menggunakan metode K-Means ini secara umum dilakukan dengan algoritma dasar sebagai berikut (Yudi Agusta, 2007) :

1. Tentukan jumlah cluster
2. Alokasikan data ke dalam cluster secara random
3. Hitung centroid/ rata-rata dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid/ rata-rata terdekat
5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan.

Beberapa alternatif penerapan K-Means dengan beberapa pengembangan teori-teori penghitungan terkait telah diusulkan. Hal ini termasuk pemilihan:

1. Distance space untuk menghitung jarak di antara suatu data dan centroid.
2. Metode pengalokasian data kembali ke dalam setiap cluster.

Beberapa permasalahan yang sering muncul pada saat menggunakan metode K-Means untuk melakukan pengelompokan data adalah:

1. Ditemukannya beberapa model clustering yang berbeda
2. Pemilihan jumlah cluster yang paling tepat
3. Kegagalan untuk converge
4. Pendeteksian outliers
5. Bentuk masing-masing cluster
6. Masalah overlapping

Keenam permasalahan ini adalah beberapa hal yang perlu diperhatikan pada saat menggunakan K-Means dalam mengelompokkan data. Permasalahan 1 umumnya disebabkan oleh perbedaan proses inisialisasi anggota masing-masing cluster. Proses initialisasi yang sering digunakan adalah proses inisialisasi secara random. Dalam suatu studi perbandingan, proses inisialisasi secara random mempunyai kecenderungan untuk memberikan hasil yang lebih baik dan independent, walaupun dari segi kecepatan untuk converge lebih lambat.

Permasalahan 2 merupakan masalah laten dalam metode K-Means. Beberapa pendekatan telah digunakan dalam menentukan jumlah cluster yang paling tepat untuk suatu dataset yang dianalisa termasuk di antaranya Partition Entropy (PE) dan GAP Statistics. Satu hal yang patut diperhatikan mengenai metode-metode ini adalah pendekatan yang digunakan dalam mengembangkan metode-metode tersebut tidak sama dengan pendekatan yang digunakan oleh K-Means dalam mempartisi data item ke masing-masing cluster. Permasalahan kegagalan untuk converge, secara teori memungkinkan untuk terjadi dalam kedua metode K-Means. Kemungkinan ini akan semakin besar terjadi untuk metode Hard K-Means, karena setiap data di dalam dataset dialokasikan secara tegas (hard) untuk menjadi bagian dari suatu cluster tertentu. Perpindahan suatu data ke suatu cluster tertentu dapat mengubah karakteristik model clustering yang dapat menyebabkan data yang telah dipindahkan tersebut lebih

sesuai untuk berada di cluster semula sebelum data tersebut dipindahkan dan demikian juga dengan keadaan sebaliknya. Kejadian seperti ini tentu akan mengakibatkan pemodelan tidak akan berhenti dan kegagalan untuk converge akan terjadi. Untuk Fuzzy K-Means walaupun ada, kemungkinan permasalahan ini untuk terjadi sangatlah kecil, karena setiap data diperlengkapi dengan membership function (Fuzzy K-Means) untuk menjadi anggota cluster yang ditemukan.

Permasalahan 4 merupakan permasalahan umum yang terjadi hampir di setiap metode yang melakukan pemodelan terhadap data. Khusus untuk metode K-Means hal ini memang menjadi permasalahan yang cukup menentukan. Beberapa hal yang perlu diperhatikan dalam melakukan pendeteksian outliers dalam proses pengelompokan data termasuk bagaimana menentukan apakah suatu data item merupakan outliers dari suatu cluster tertentu dan apakah data dalam jumlah kecil yang membentuk suatu cluster tersendiri dapat dianggap sebagai outliers. Proses ini memerlukan suatu pendekatan khusus yang berbeda dengan proses pendeteksian outliers di dalam suatu dataset yang hanya terdiri dari satu populasi yang homogen.

Permasalahan kelima adalah menyangkut bentuk cluster yang ditemukan. Tidak seperti metode data clustering lainnya termasuk Mixture Modelling, K-Means umumnya tidak mengindahkan bentuk dari masing-masing cluster yang mendasari model yang terbentuk, walaupun secara alamiah masing-masing cluster umumnya berbentuk bundar. Untuk dataset yang diperkirakan mempunyai bentuk yang tidak biasa, beberapa pendekatan perlu untuk diterapkan.

Masalah overlapping sebagai permasalahan terakhir sering sekali diabaikan karena umumnya masalah ini sulit terdeteksi. Hal ini terjadi untuk metode Hard K-Means dan Fuzzy K-Means, karena secara teori metode ini tidak diperlengkapi feature untuk mendeteksi apakah di dalam suatu cluster ada cluster lain yang kemungkinan tersembunyi.

K-Means merupakan metode data clustering yang digolongkan sebagai metode pengklasifikasian yang bersifat unsupervised (tanpa arahan). Pengkategorian metodemetode pengklasifikasian data antara supervised dan unsupervised classification didasarkan pada adanya dataset yang data itemnya sudah sejak awal mempunyai label kelas dan dataset yang data itemnya tidak mempunyai label kelas.

Untuk data yang sudah mempunyai label kelas, metode pengklasifikasian yang digunakan merupakan metode supervised classification dan untuk data yang belum mempunyai label kelas, metode pengklasifikasian yang digunakan adalah metode unsupervised classification.

Selain masalah optimasi pengelompokan data ke masing-masing cluster, data clustering juga diasosiasikan dengan permasalahan penentuan jumlah cluster yang paling tepat untuk data yang dianalisa. Untuk kedua jenis K-Means, baik Hard KMeans dan Fuzzy K-Means, yang telah dijelaskan di atas, penentuan jumlah cluster untuk dataset yang dianalisa umumnya dilakukan secara supervised atau ditentukan dari awal oleh pengguna, walaupun dalam penerapannya ada beberapa metode yang sering dipasangkan dengan metode K-Means. Karena secara teori metode penentuan jumlah cluster ini tidak sama dengan metode pengelompokan yang dilakukan oleh KMeans, kevalidan jumlah cluster yang dihasilkan umumnya masih dipertanyakan.

1.2 Kajian Induktif

Tinjauan Pustaka tersebut adalah hasil penelitian terdahulu tentang informasi hasil penelitian yang telah dilakukan sebelumnya dan menghubungkan dengan masalah yang sedang diteliti.

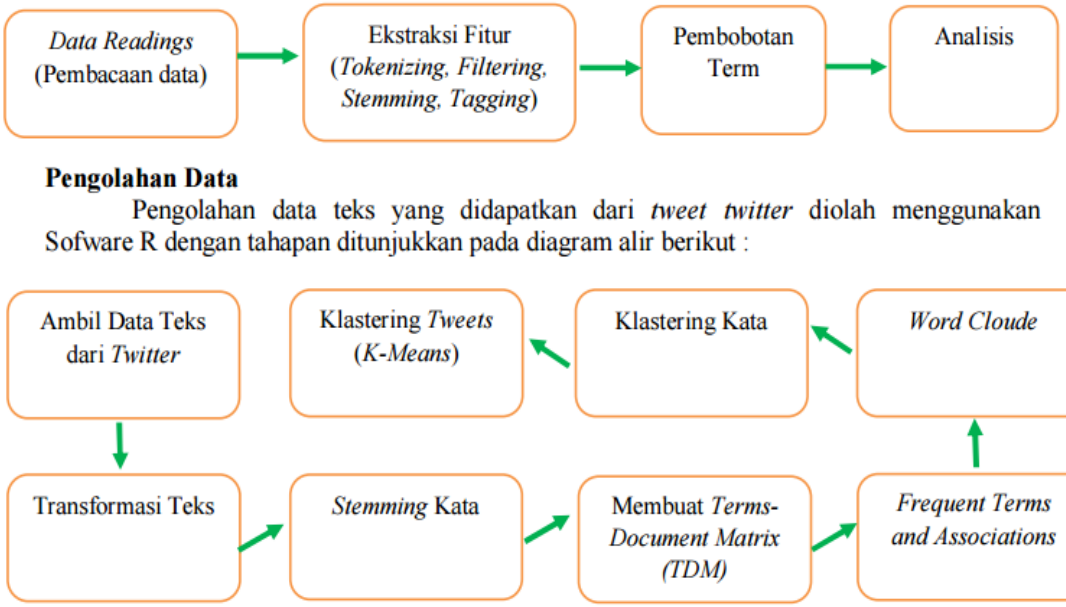
Implementasi text mining pada mesin pencarian twitter untuk mengetahui karakter seseorang menggunakan algoritma Naïve Bayes Classifier.

Menurut Mohammad Zoqi Sarwani, Wayan Firdaus Mahmudy (2015) dalam jurnal dengan judul Analisis Twitter Untuk Mengetahui Karakter Seseorang Menggunakan Algoritma Naïve Bayes Classifier, maka dapat ditarik kesimpulan yang menghasilkan sebuah informasi bahwa twitter dapat digunakan sebagai salah satu media untuk mengetahui kepribadian seseorang melakui posting atau tweets mereka. Selain itu, proses pengklasifikasian twitter menggunakan metode Naïve Bayess Classificaation juga mampu memberikan tingkat akurasi yang baik dengan membandingkan hasil klasifikasi dari sistem dengan hasil dari pakar. Untuk pengembangan dari penelitian ini, perlu untuk melakukan percobaan dengan menggunakan jumlah data latih dan data uji yang besar untuk menghitung keakurasian metode yang digunakan dalam penelitian ini.

BAB III
METODE PENELITIAN

Penelitian ini menggunakan data-data teks pada media sosial twitter yang berkaitan dengan kata kunci “startup”. Tujuan akhir penelitian ini adalah dapat mendeskripsikan topik utama dan kata - kata yang melekat pada “startup”, serta mengelompokkan topik - topik lain yang saling berkaitan. Populasi dari penelitian ini adalah berita/dokumen (teks) pada sosial media twitter yang termasuk dalam kategori microblogging tentang “startup”. Tweet-tweet pengguna pada twitter dapat dijadikan sebagai topik-topik pada media online. Sampel dan Teknik Pengambilan Sampel Sampel yang diambil adalah data teks dari twitter yang diambil dengan permintaan kepada sistem sebanyak 250 tweet pada periode waktu tertentu.

Sumber Data dan Metode Pengumpulan Data yang digunakan adalah data primer yang dikumpulkan dari media sosial twitter. Data diambil dengan cara mendownload dan mengumpulkan data tweet mengenai “startup” dari hasil pencarian search engine twitter (mesin pencarian twitter) dan menyimpannya menggunakan suatu program dengan software “R”. Tahapan Analisis Data secara umum, tahapan melakukan analisis text mining dapat digambarkan dalam diagram alir berikut :



Gambar 3.1 Diagram Alir

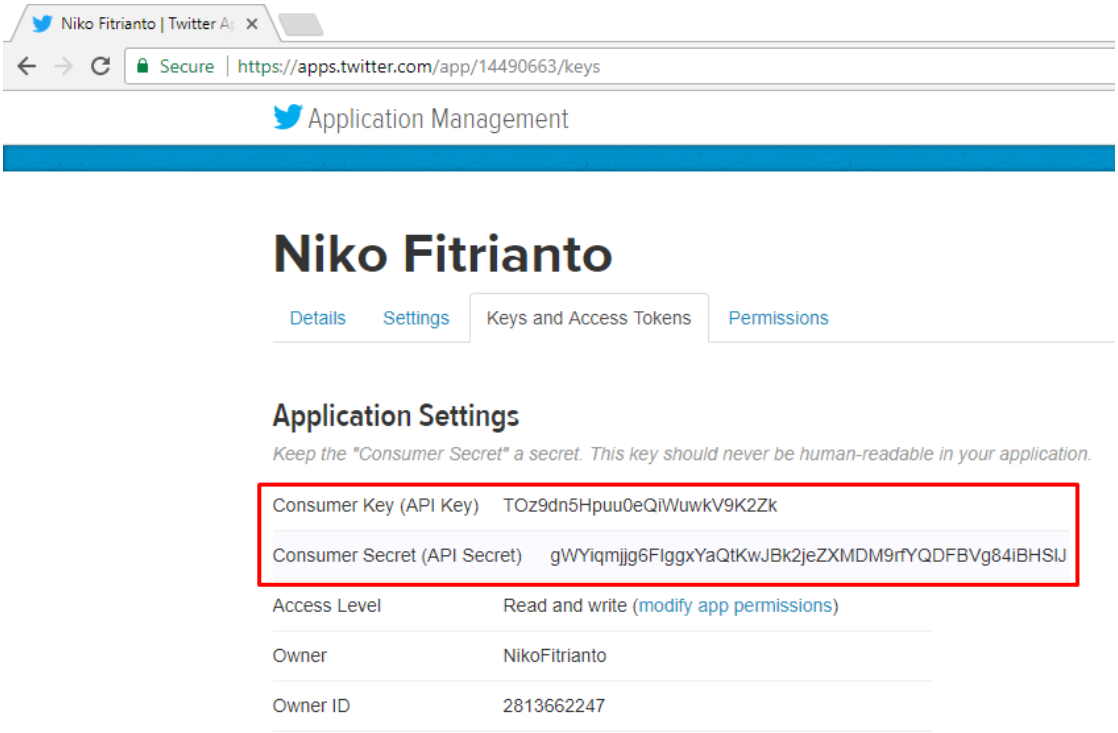
BAB IV
HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

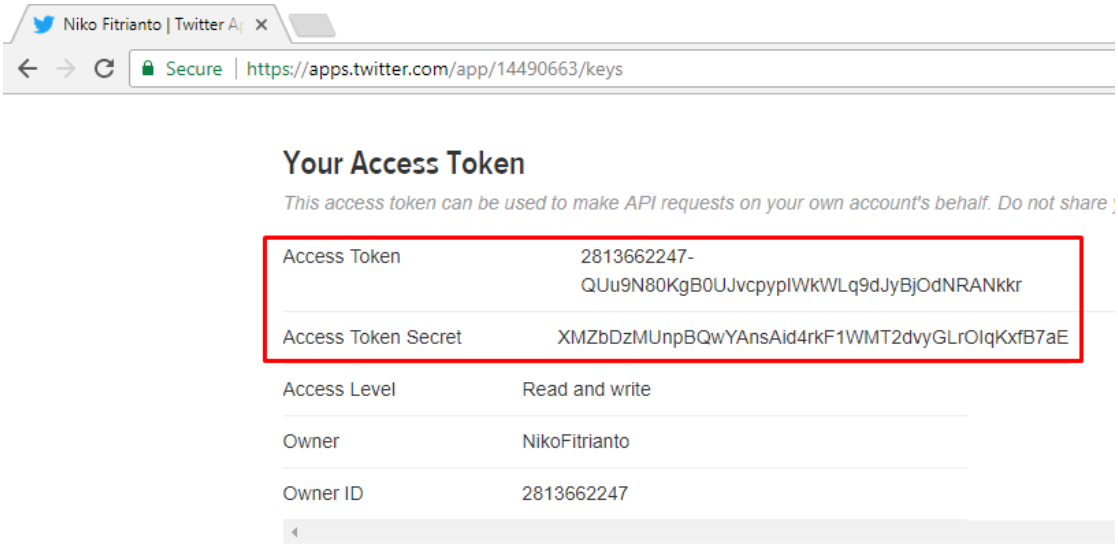
4.1.1. Pembuatan Akun Twitter

Sebelum mengakses twitter terlebih dahulu membuat akun twitter sendiri, selain untuk mengakses data di dalamnya tetapi juga untuk mendapatkan token twitter, Token atau Access Token sendiri, dalam arsitektur Windows NT adalah sebuah objek sistem operasi (yang diberi nama "*Token*") yang merepresentasikan subjek dalam beberapa operasi pengaturan akses (*access control*). Objek Token umumnya dibuat oleh layanan logon (*logon service*) untuk merepresentasikan informasi keamanan yang diketahui mengenai sebuah pengguna yang lolos proses autentikasi (*authenticated user*). Token yang disediakan dalam twitter, seperti :

1. Consumer Key
2. Consumer Secret Key
3. Access Token Key
4. Access Token Secret Key



Gambar 4.1 Consumer Key dan Consumer Secret Key



Gambar 4.2 Access Token Key dan Access Token Secret Key

4.2. Pengolahan Data

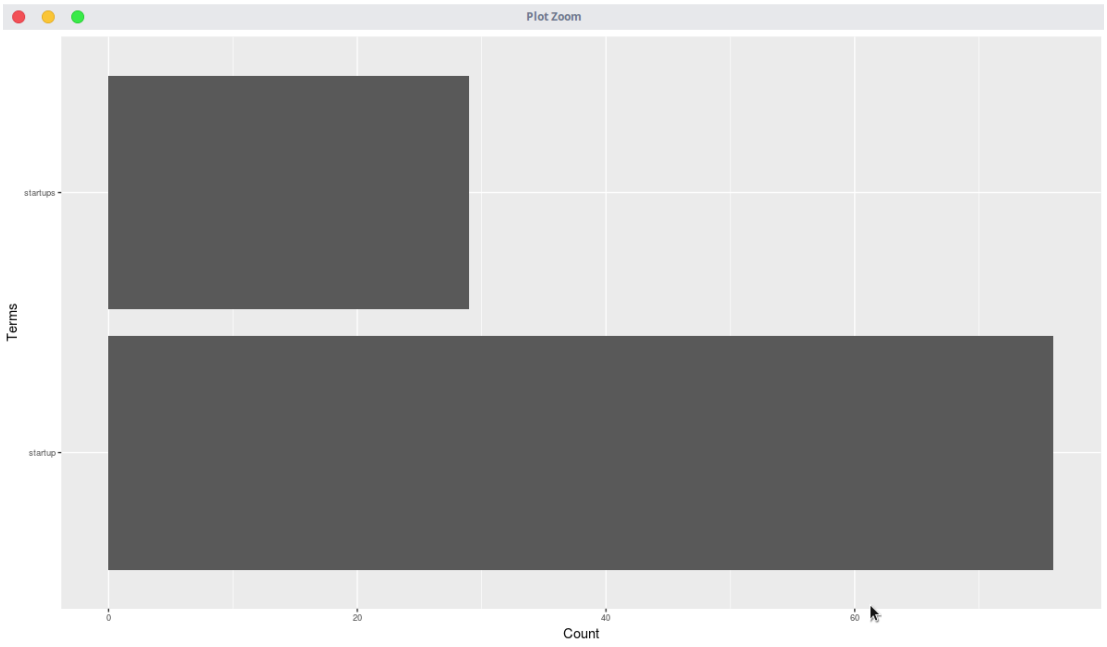
Term of Frequency "startup" berikut ini merupakan gambar term of frequency topik-topik terkait "startup".

```
tweets <- userTimeline("startup", n = 250)
show(tweets)
n.tweet <- length(tweets)
# convert tweets to a data frame
tweets.df <- twListToDF(tweets)
myCorpus <- Corpus(VectorSource(tweets.df$text))
# convert to lower case
myCorpus <- tm_map(myCorpus, content_transformer(tolower))
# remove URLs
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeURL))
# remove anything other than English letters or space
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))
# remove stopwords
myStopwords <- c(setdiff(stopwords('english'), c("r", "big")),"use", "see", "used", "via", "amp")
```

```
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
# remove extra whitespace
myCorpus <- tm_map(myCorpus, stripWhitespace)
# keep a copy for stem completion later
myCorpusCopy <- myCorpus

term.freq <- rowSums(as.matrix(tdm))
tdm <- TermDocumentMatrix(myCorpus)
term.freq <- subset(term.freq, term.freq >= 10)
df <- data.frame(term = names(term.freq), freq = term.freq)

ggplot(df, aes(x=term, y=freq)) + geom_bar(stat="identity") +
  xlab("Terms") + ylab("Count") + coord_flip() +
  theme(axis.text=element_text(size=7))
```



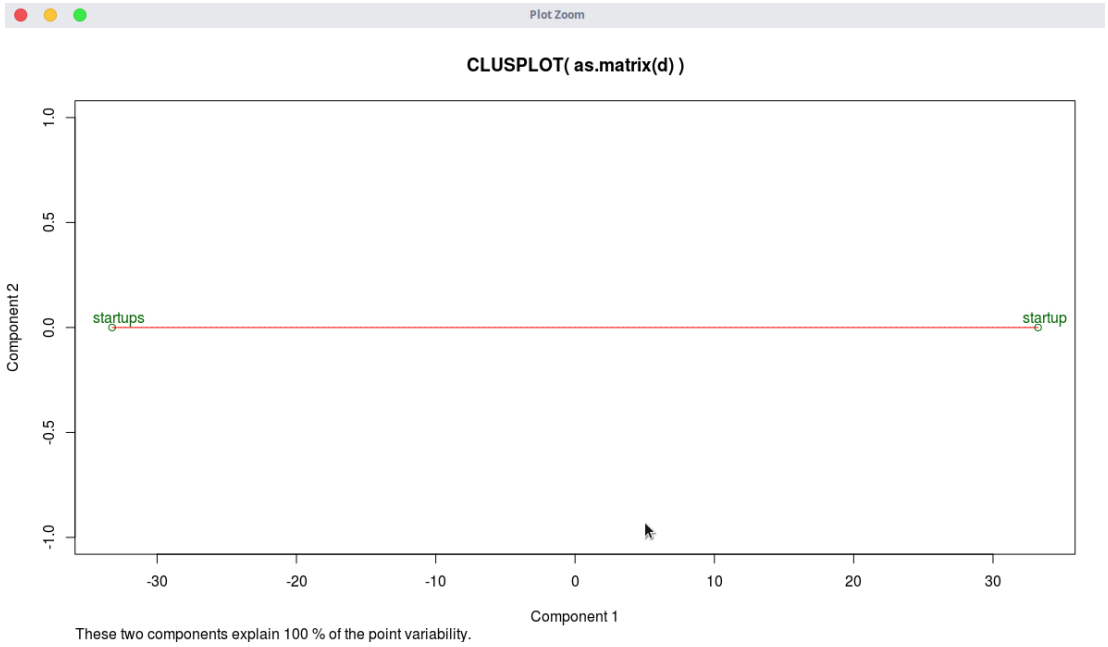
Gambar 4.1. Barchart yang menunjukan kata-kata yang menjadi topik utama dan sering muncul pada “startup”

Berdasarkan *barchart* (diagram batang) pada gambar diatas dapat diketahui *term of frequency* dari kata-kata (topik) lain yang sering muncul bersamaan dengan kata startup. *Term of frequency* adalah kata-kata yang sering muncul dari data teks yang dianalisis,

yang ditampilkan dalam bentuk diagram batang dimana topik utama terkait kata kunci ditampilkan dengan gambar batang yang lebih panjang dibanding lainnya.

Gambar 4.1. barchart menunjukan kata-kata yang menjadi topik utama dan sering muncul saat pengguna twitter mencari informasi terkait startup pada search engine twitter adalah kata startup. Sedangkan kata- kata lain seperti startups, json, funding, good, python, ideas merupakan topik menarik lain yang cukup sering digunakan terkait pencarian topik startup.

```
#k-means clustering
d <- dist(term.freq, method="euclidian")
carsCluster <- kmeans(term.freq, 3)
clusplot(as.matrix(d), carsCluster$cluster, color=T, shade=T, labels=3, lines=0)
```

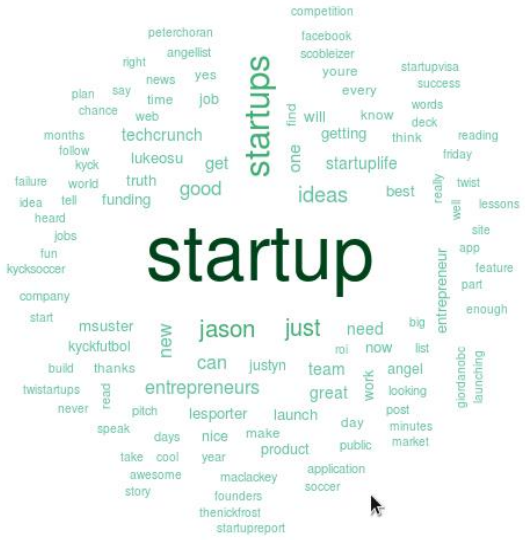


Gambar 4.2. Hasil klasifikasi tweets berdasarkan plotcluster

Berdasarkan plot diatas, data dapat mengelompok dengan baik sesuai dengan kesamaan karakteristiknya, cluster 1 ditunjukkan dengan plot berwarna merah, cluster 2 warna biru, dan cluster 3 berwarna merah muda.

```
m <- as.matrix(tdm)
# calculate the frequency of words and sort it by frequency
```

```
word.freq <- sort(rowSums(m), decreasing = T)
# colors
pal <- brewer.pal(9, "BuGn")[-(1:4)]
# plot word cloud
wordcloud(words = names(word.freq), freq = word.freq, min.freq = 3,
 random.order = F, colors = pal)
```



Gambar 4.3. Wordcloud

Dari Gambar 4.3, terlihat bahwa Word cloud (awan kata) adalah kumpulan kata-kata yang paling banyak muncul dalam data teks yang dianalisis. Kata-kata tersebut terkumpul seperti sebuah gumpalan awan yang berisi kata-kata sehingga disebut awan kata. Intensitas keseringan kata yang digunakan, ditunjukan dengan ukuran huruf pada kata. Semakin besar huruf dari kata yang terdapat di awan kata menunjukkan semakin sering kata tersebut muncul. Tampilan awan kata ini lebih menarik serta cepat untuk menemukan kata-kata yang sering muncul, akan tetapi awan kata memiliki kekurangan yaitu tidak dapat menunjukkan frekuensi kata-kata muncul dalam suatu teks yang dianalisis.

Berdasarkan gambar 4.3. dapat diketahui kata (topik) yang sering muncul dalam word cloud bersamaan dengan kata **startup** adalah kata yang menjadi topik utama dan

sering muncul saat pengguna twitter mencari informasi terkait startup pada search engine twitter adalah kata startup. Sedangkan kata- kata lain seperti startups, json, funding, good, python, ideas merupakan topik menarik lain yang cukup sering digunakan terkait pencarian topik startup.

BAB V
KESIMPULAN DAN REKOMENDASI

Klustering adalah mengelompokkan variabel ke dalam kelompok berdasarkan kesamaan tertentu. Sedangkan R adalah suatu perangkat lunak yang digunakan untuk manipulasi data, perhitungan, simulasi, penayangan grafik, dan sekaligus sebagai bahasa pemograman yang bersifat interpreter R. Dari hasil pembahasan diatas mengenai aplikasi teks mining untuk penanganan data besar hasil pencarian topik-topik terkait pada search engine twitter dengan studi kasus topik **startup** maka dapat ditarik kesimpulan yang menghasilkan sebuah informasi atau pemberitaan yang sedang hangat dibicarakan dan diberitakan oleh media mengenai **startup** bahwa terdapat topik utama yang digunakan pada tweet-tweet ketika pengguna twitter mencari informasi seputar topik startup pada search engine dapat diketahui kata (topik) yang sering muncul dalam word cloud bersamaan dengan kata **startup** adalah kata yang menjadi topik utama dan sering muncul saat pengguna twitter mencari informasi terkait **startup** pada search engine twitter adalah kata **startup**. Sedangkan kata- kata lain seperti startups, json, funding, good, python, ideas merupakan topik menarik lain yang cukup sering digunakan terkait pencarian topik **startup**.

DAFTAR PUSTAKA

Mohammad Zoqi Sarwani, Wayan Firdaus Mahmudy. 2015. Analisis Twitter Untuk Mengetahui Karakter Seseorang Menggunakan Algoritma Naïve Bayes Classifier. Jurnal Prosiding Pendidikan Matematika UMS

Agusta Yudi, 2007. *K-Means – Penerapan Permasalahan dan Metode Terkait*, Jurnal Sistem dan Informatika Vol. 3 (Pebruari 2007), 47-60.

Zhao, Yanchang 2011. R and Data Mining: Examples and Case Studies. Elsevier. sumber: http:// http://www.rdatamining.com/examples/kmeans-clustering (29 November 2017)