TEXT PROCESSING

K-Means Clustering

OF CONSULTATION TEXT DATA OF DIABETIC

YOHANNES ALEXANDER AGUSTI SINAGA

WHAT IS THE PROJECT ABOUT ?

This project is a simple search engine that uses text data. This project consists of breaking down the document data into several clusters that are similar in characteristics. The scope of the data is diabetes consultation.

This project adopts several methods, such as Natural Languange Processing and Machine Learning (Clustering).

METHODS

1. Data Gathering
2. Preprocessing
3. Determining The Number of Cluster
4. Visualizing The Clusters
5. Clustering K-Means
6. Result

DATA GATHERING

We got the data from the site which is https://www.sehatq.com/cari?filter=forum&q=diabetes. We got 50 respondents on the site by doing the copy manually with 4 columns and 50 rows. This data is about consulting between doctor's answers and patient complaints through the website with diabetes keywords. the data was retrieved in the middle of 2022. The dataset is stored in CSV format.

The description of Columns:

1. **NAMA DOKTER** : The name of the doctor who answered the question of respondent.
2. **JUDUL** : The title of the question
3. **TULISAN** : The answer of the doctor to the question.
4. **LINK** : The site link of the doctor to the question.

PREPROCESSING

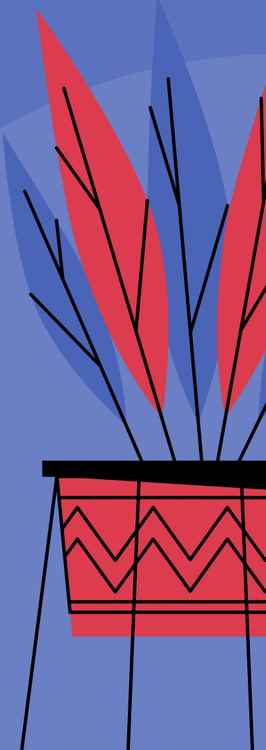
1. Filtering (StopWords)
2. Stemming
3. Tokenizing
4. Bag Of Words
5. Filtering keyword
6. Handling Missing Values

```
diabetes melitus kelompok penyakit metabolik d...
diabetes gula darah penyakit kronis dikontrol ...
fenomena fajar penderita diabetes disebabkan h...
diabetes penyakit kronis gangguan organ pankre
diabetes kondisi gula darah puasa melebihi bat
diabetes kondisi medis berkaitan hormon insuli
penglihatan berbayang kabur nyaman penderitany
diabetes peningkatan gula darah gula darah 200
oats berasal biji gandum kering , mengandung z
diabetes mellitus kondisi dimana sesoerang mem
diabetes melitus kelainan kelenjar insulin men
penyakit diabetes memiliki faktor keturunan ,
diabetes penyakit berkaitan gula darah . pasie
diabetes melitus penyakit kronis dimana gula d
diabetes penyakit sembuh gula darah dikotrol m
prinsipnya , penderita diabetes menjalankan ib
komplikasi penyakit diabetes pria satunya impo
diabetes penyakit kronis lansia , tandanya pen
kondisi gangguan ereksi ereksi tahan ereksi ko
diabetes peningkatan gula darah akibat ganggua
diabetes gula darah penyakit kronis identik fa...
panas dingin , badan pegal , batuk , hilang in...
diabetes penyakit alami . diabetes dimana pank...
```

Filtering (StopWords)

Stopword is a technique of removing words that do not have a significant meaning. Most of the words removed are conjunctions and question words, like "apakah", "atau", and etc

Each document is cleaned of words that have no significant meaning. These words are taken from a combination of the NLTK and Sastrawi Python libraries.



Stemming

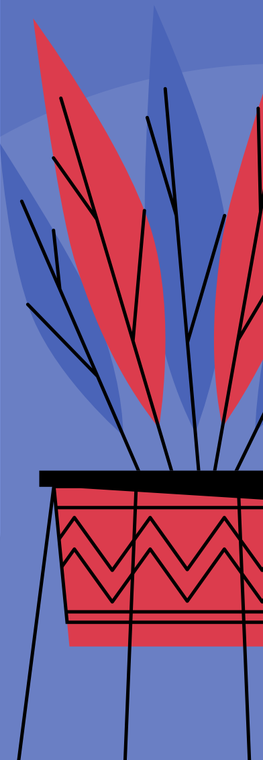
diabetes sakit mana gula temu darah muncul keluh mudah lapar mudah haus gatal kencing turun berat badan ukur gula darah perik
sa gula darah periksa gula acak waktu puasa gula darah puasa periksa puasa 8 jam gula darah 2 jam makan hba1c gula sel darah
merah rata gula darah 3 periksa istilah diabetes gula darah periksa gula darah perihal keluh kepala kliyengan ringan jatuh ta
nda gula rendah batas makan ketat kontrol gula makan minum obat atur laku cek gula darah ulang jalan ibadah puasa tunjuk hasi
l dokter obat diabetes dosis atur pola makan

diabetes sakit mana tubuh produksi insulin sebab tingkat glukosa tubuh tipe diabetes diabetes tipe 1 akibat genetik muncul an
ak diabetes tipe 2 akibat gaya hidup muncul dewasa gejala diabetes cepat haus lapar buang air malam lemas semut berat badan t
urun pandang kabur mulut kering dll mendiagnosa sakit periksa hba1c gula darah gula darah puasa tes toleransi glukosa oral de
teksi sakit hrs obat obat beri insulin sunti obat anti diabetes metformin glimepirid dll aktivitas fisik pola makan sehat oba
t diabetes timbul komplikasi gagal ginjal saraf buluh darah jantung obat rutin dokter spesialis sakit obat tuju gula darah st
abil sakit ginjal parah rawat inap selagi kondisi badan stabil ajak suami tani kontrol suami jelas dokter kondisi harap depan
suami erti

nak milik orangtua diabetes takut kena diabetes milik faktor risiko obesitas obesitas sebab diabetes pasien obesitas kena dia
betes apakah anak obesitas orangtua diabetes alami diabetes umur 13 berat badan 63 kg milik riwayat sakit pradiabetes alami p
utih kuku kuning gigi goyang rasa keluh lemas kencing haus kena obesitas putih kuku kuning gigi goyang gejala diabetes obesit
as sebab diabetes tambah riwayat orangtua diabetes juga namun obesitas putih kuku kuning gigi goyang gejala khas diabetes gej
ala khas diabetes 4 poliuri kemih polidipsi haus terus polifagi makan turun berat badan sebab putih remaja putrikeputihan seb

Stemming is a technique of removing words that do not have a significant meaning. Most of the words removed are conjunctions and question words, like "apakah", "atau", and etc

Each document is cleaned of words that have no significant meaning. These words are taken from a combination of the NLTK and Sastrawi Python libraries.



Tokenizing

```
['diabetes', 'sakit', 'mana', 'gula', 'temu', 'darah', 'muncul', 'keluh', 'mudah', 'lapar', 'mudah', 'haus', 'gatal', 'kencin
g', 'turun', 'berat', 'badan', 'ukur', 'gula', 'darah', 'periksa', 'gula', 'darah', 'periksa', 'gula', 'acak', 'waktu', 'puas
a', 'gula', 'darah', 'puasa', 'periksa', 'puasa', 'jam', 'gula', 'darah', 'jam', 'makan', 'hbac', 'gula', 'sel', 'darah', 'me
rah', 'rata', 'gula', 'darah', 'periksa', 'istilah', 'diabetes', 'gula', 'darah', 'periksa', 'gula', 'darah', 'perihal', 'kel
uh', 'kepala', 'kliyengan', 'ringan', 'jatuh', 'tanda', 'gula', 'rendah', 'batas', 'makan', 'ketat', 'kontrol', 'gula', 'maka
n', 'minum', 'obat', 'atur', 'laku', 'cek', 'gula', 'darah', 'ulang', 'jalan', 'ibadah', 'puasa', 'tunjuk', 'hasil', 'dokte
r', 'obat', 'diabetes', 'dosis', 'atur', 'pola', 'makan']

['diabetes', 'sakit', 'mana', 'tubuh', 'produksi', 'insulin', 'sebab', 'tingkat', 'glukosa', 'tubuh', 'tipe', 'diabetes', 'di
abetes', 'tipe', 'akibat', 'genetik', 'muncul', 'anak', 'diabetes', 'tipe', 'akibat', 'gaya', 'hidup', 'muncul', 'dewasa', 'g
ejala', 'diabetes', 'cepat', 'haus', 'lapar', 'buang', 'air', 'malam', 'lemas', 'semut', 'berat', 'badan', 'turun', 'pandan
g', 'kabur', 'mulut', 'kering', 'dll', 'mendiagnosa', 'sakit', 'periksa', 'hbac', 'gula', 'darah', 'gula', 'darah', 'puasa',
'tes', 'toleransi', 'glukosa', 'oral', 'deteksi', 'sakit', 'hrs', 'obat', 'obat', 'beri', 'insulin', 'sunti', 'obat', 'anti',
'diabetes', 'metformin', 'glimepirid', 'dll', 'aktivitas', 'fisik', 'pola', 'makan', 'sehat', 'obat', 'diabetes', 'timbul',
'komplikasi', 'gagal', 'ginjal', 'saraf', 'buluh', 'darah', 'jantung', 'obat', 'rutin', 'dokter', 'spesialis', 'sakit', 'oba
t', 'tuju', 'gula', 'darah', 'stabil', 'sakit', 'ginjal', 'parah', 'rawat', 'inap', 'selagi', 'kondisi', 'badan', 'stabil',
'ajak', 'suami', 'tani', 'kontrol', 'suami', 'jelas', 'dokter', 'kondisi', 'harap', 'depan', 'suami', 'erti']

['nak', 'milik', 'orangtua', 'diabetes', 'takut', 'kena', 'diabetes', 'milik', 'faktor', 'risiko', 'obesitas', 'obesitas', 's
```

Tokenizing is a word processing technique to perform word separation in a sentence.

Each document is separated word by word.

Bag Of Words

Word Document	abortus	acak	acarbose	adang	agam	air	ajak	akantosis	akar	akibat	...	warna	waspada	wawancara	x	yaa	yang	yogurt	zaitun	
Document 1	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
Document 10	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
Document 11	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
Document 12	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
Document 13	NaN	NaN	NaN	NaN	1.0	2.0	NaN	NaN	NaN	NaN	...	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	N
Document 14	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN	NaN	1.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
Document 15	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	NaN	1.0	...	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	N

bag of words is a word processing technique to count the number of each word in a sentence.

After tokenizing, the words are counted separately and converted into a data frame.

Filtering keyword

Word Document	diabetes	sakit	gula	darah	obesitas	insulin	glukosa	obat	glikogen	makan	minum	ginjal	hormon	saraf	manis	melitus	kalori	pankreas
Document 1	3.0	1.0	13.0	10.0	NaN	NaN	NaN	2.0	NaN	4.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Document 10	5.0	5.0	2.0	2.0	NaN	2.0	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	5.0	NaN	1.0
Document 11	4.0	1.0	8.0	7.0	1.0	NaN	NaN	2.0	NaN	2.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Document 12	1.0	NaN	7.0	7.0	NaN	7.0	1.0	1.0	NaN	NaN	1.0	NaN	4.0	NaN	NaN	NaN	NaN	NaN
Document 13	1.0	1.0	1.0	3.0	NaN	NaN	NaN	2.0	NaN	NaN	NaN	1.0	NaN	1.0	NaN	NaN	NaN	1.0
Document 14	4.0	1.0	3.0	3.0	NaN	1.0	NaN	3.0	NaN	5.0	2.0	NaN	1.0	NaN	NaN	NaN	NaN	NaN
Document 15	5.0	3.0	6.0	4.0	NaN	4.0	NaN	3.0	NaN	1.0	NaN	NaN	2.0	5.0	NaN	NaN	NaN	1.0

Because there are so many words and some words cannot be found in other documents, the next step is to filter words related to diabetes.

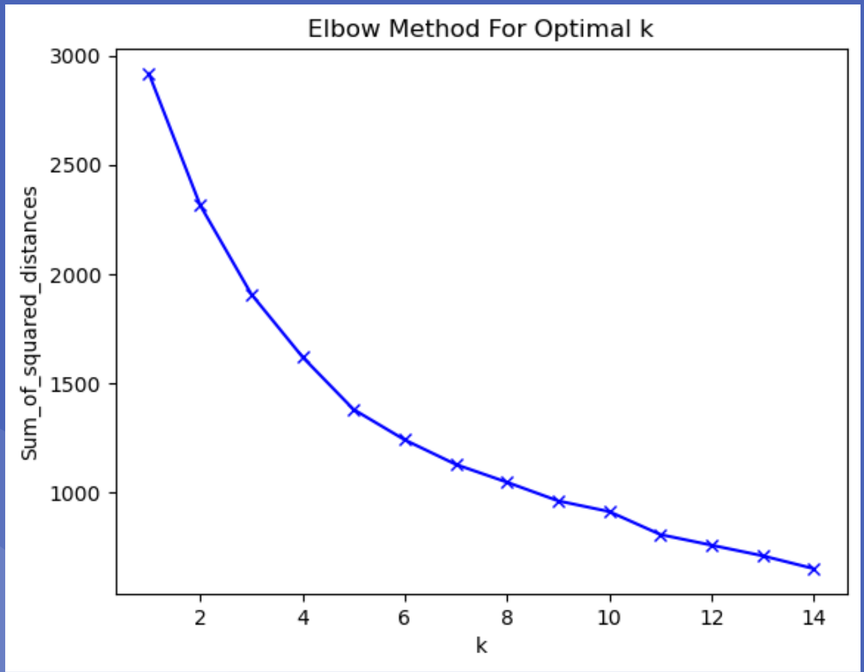
There were 19 words related to diabetes, namely diabetes, sakit, gula , darah , obesitas, insulin, glukosa, obat, glikogen, makan, minum, ginjal, hormon, saraf, manis, melitus, kalori, pankreas, dm .

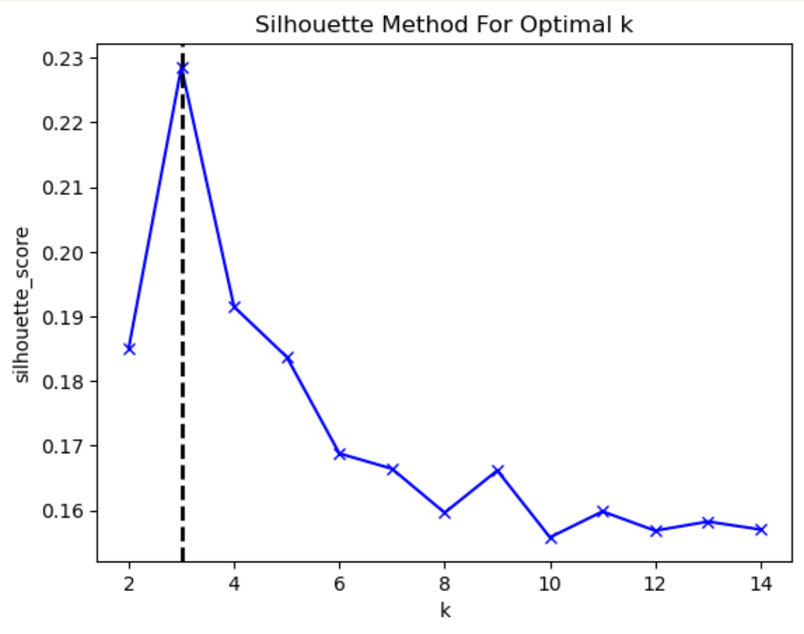
Handling Missing Values

Word Document	diabetes	sakit	gula	darah	obesitas	insulin	glukosa	obat	glikogen	makan	minum	ginjal	hormon	saraf	manis	melitus	kalori	pankreas
Document 1	3.0	1.0	13.0	10.0	NaN	NaN	NaN	2.0	NaN	4.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Document 10	5.0	5.0	2.0	2.0	NaN	2.0	NaN	NaN	NaN	1.0	NaN	NaN	NaN	NaN	NaN	5.0	NaN	1.0
Document 11	4.0	1.0	8.0	7.0	1.0	NaN	NaN	2.0	NaN	2.0	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Document 12	1.0	NaN	7.0	7.0	NaN	7.0	1.0	1.0	NaN	NaN	1.0	NaN	4.0	NaN	NaN	NaN	NaN	NaN
Document 13	1.0	1.0	1.0	3.0	NaN	NaN	NaN	2.0	NaN	NaN	NaN	1.0	NaN	1.0	NaN	NaN	NaN	1.0
Document 14	4.0	1.0	3.0	3.0	NaN	1.0	NaN	3.0	NaN	5.0	2.0	NaN	1.0	NaN	NaN	NaN	NaN	NaN
Document 15	5.0	3.0	6.0	4.0	NaN	4.0	NaN	3.0	NaN	1.0	NaN	NaN	2.0	5.0	NaN	NaN	NaN	1.0

Some documents have missing values for these words. Therefore, the input value is 0 for words that have a missing value

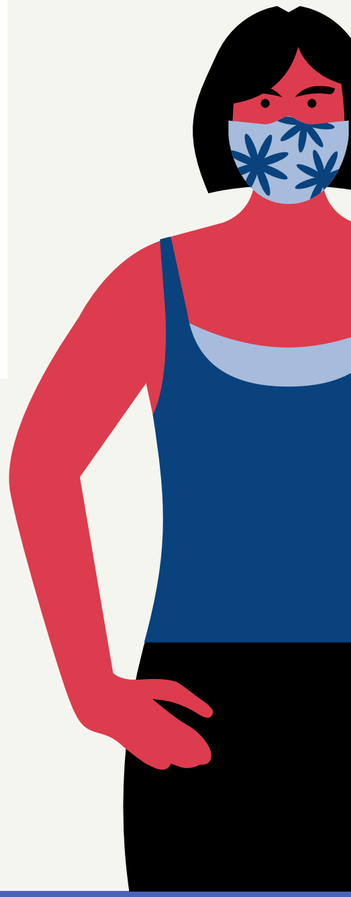
Determining The Number of Cluster

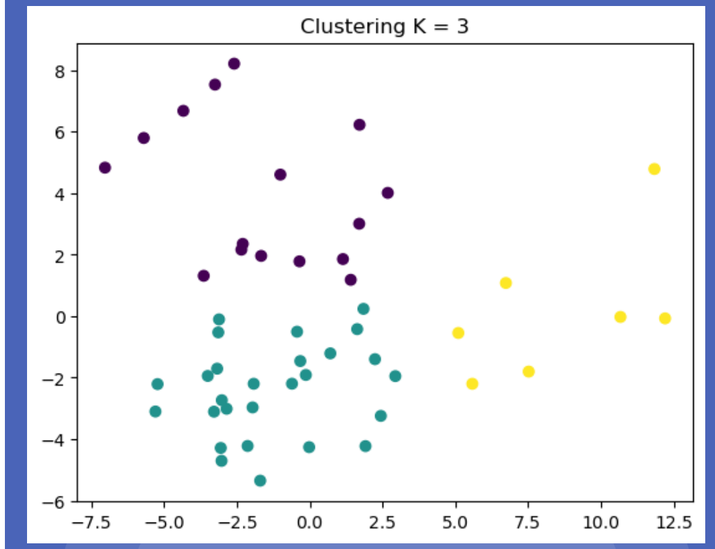




Based on Elbow Method, the elbow on the graph is not clearly described. Therefore, the most optimal number of cluster can not determine on this method.

Because the most optimal of cluster has not been clearly descibed, Silhouette Method shows that 3 is the most optimal number of cluster.





Because the data has many dimensions, the PCA method must be used to reduce dimensions to facilitate visualization.

After we made the K-Means Clustering Model, we cluster the data into 3 cluster.

The graph shows that the data is perfectly separated.

K-MEANS CLUSTERING

RESULT

The number of cluster 0 is 16 documents. The number of cluster 1 is 27 documents. The number of cluster 2 is 7 documents. It means that the member of cluster. It means that the members in the cluster have the same characteristics.

For instances, on the cluster 0, Document 1 and Document 10 are similar characteristics.

Therefore, this model describes like a simple search engine document based on keywords.

THANK YOU

https://github.com/Yohannes-Alexander/machine_learning_project/tree/main/Project_Diabetes_Text%20Processing_K-Means