

# Unsupervised Text Clustering using NLP

Emma Goh, DSIF 5

# Table of Contents

1. Problem Statement
2. Exploratory Data Analysis
3. Preprocessing, Text Feature Extraction
4. Clustering Algorithm Evaluation
5. Data Abstraction
6. Final Evaluation

01

# PROBLEM STATEMENT



- Aims to democratise giving
- helps patients with hospital bills
- enables children to study in rural areas
- helps organisers raise funds for causes
- provides support for humanitarian effort for countries in need



✓ VERIFIED

### Help Jayden Live A Normal Life

Give Jayden a CHANCE to Grow Just Like Any other Child. Jayden was born with a rare genetic bone disorder known as achondroplasia, also known as dwarfism. Jayden is currently 9 mon...

by Amanda Tan

RAISED S\$427,362 OF S\$1,423,728

SGD 114.2M

Raised

2,082,373

Givers

18,725

Campaigns

849

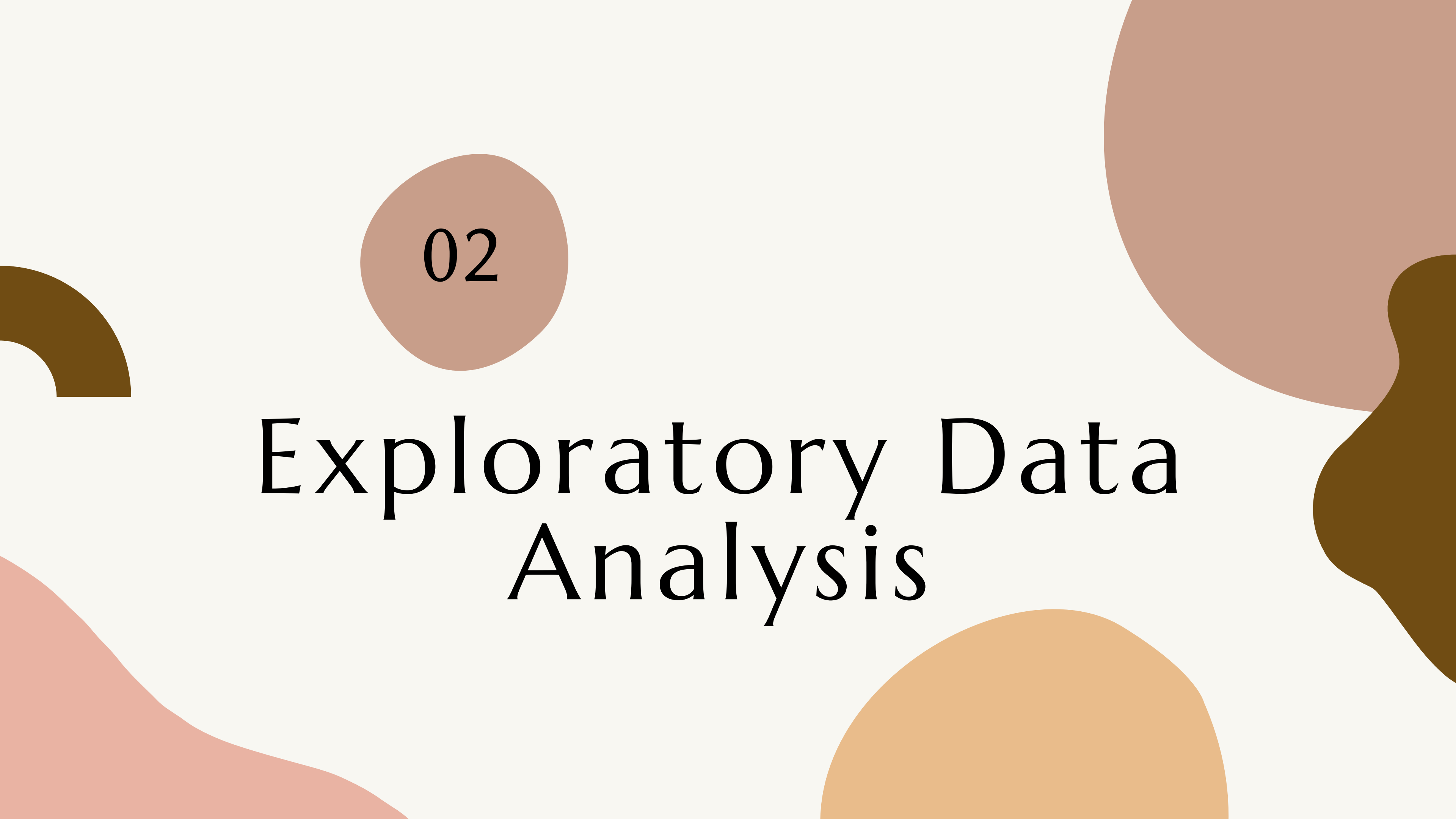
Charities

**KINDNESS**  
*in* ACTION



# Problem Statement

- To gather and group insights from the comments that donors left after donating to campaigns on Give.asia - 'why i gave?'
- To identify the types of donors and using these insights (if any) to design future donor engagement



02

# Exploratory Data Analysis

# Quick Overview (Dataset)

01

'COMMENT'

1,000 rows of  
comments  
collected

02

'CURRENCY'

7 currencies captured

- SGD, MYR, VND, USD,  
INR, THB, HKD

03

'AMOUNT'

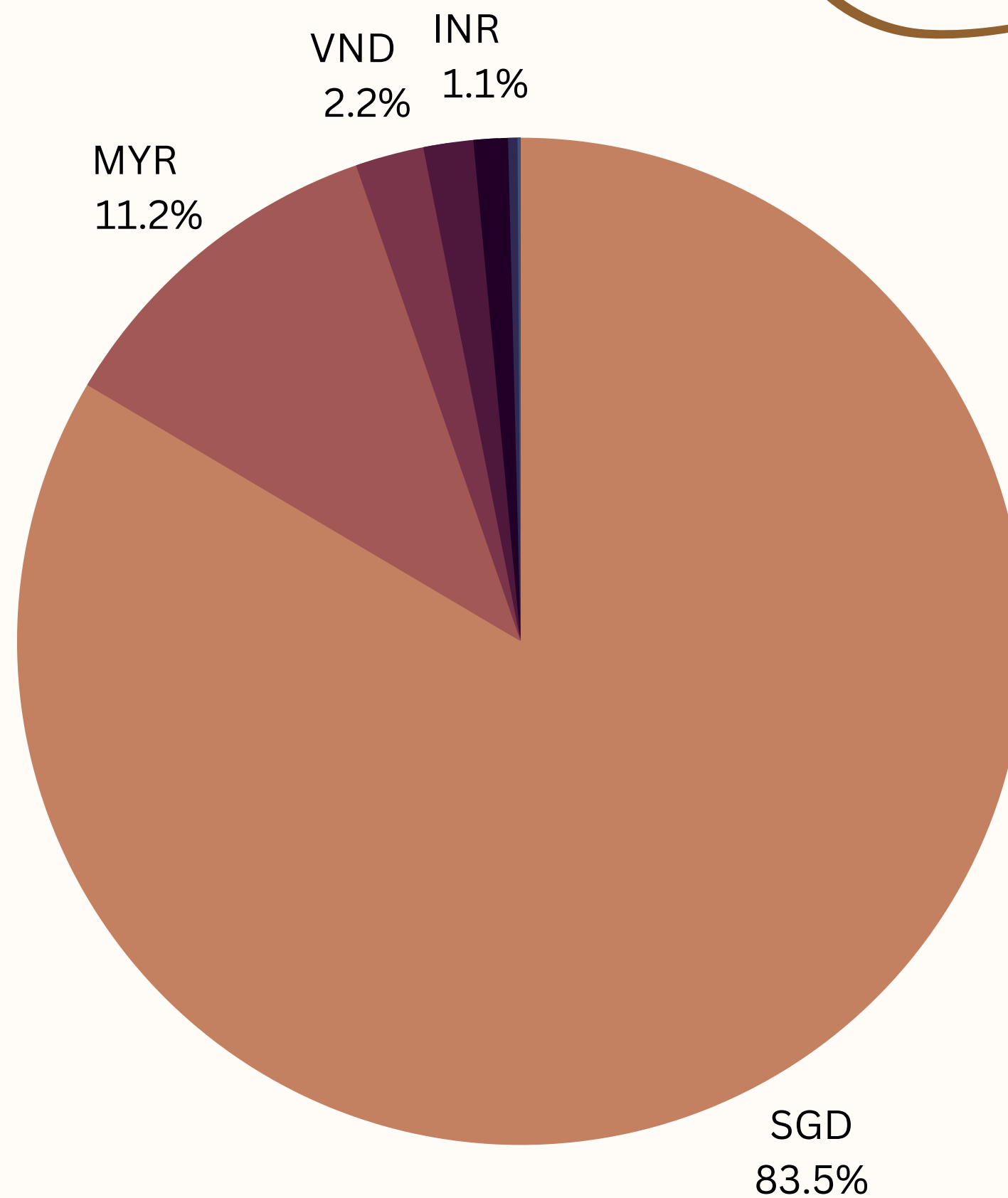
- Average amount for  
SGD - SGD\$8, 673
- Average amount for  
MYR - RM\$24,130

# 02

'CURRENCY'

7 currencies captured

- SGD, MYR, VND,USD, INR, THB, HKD

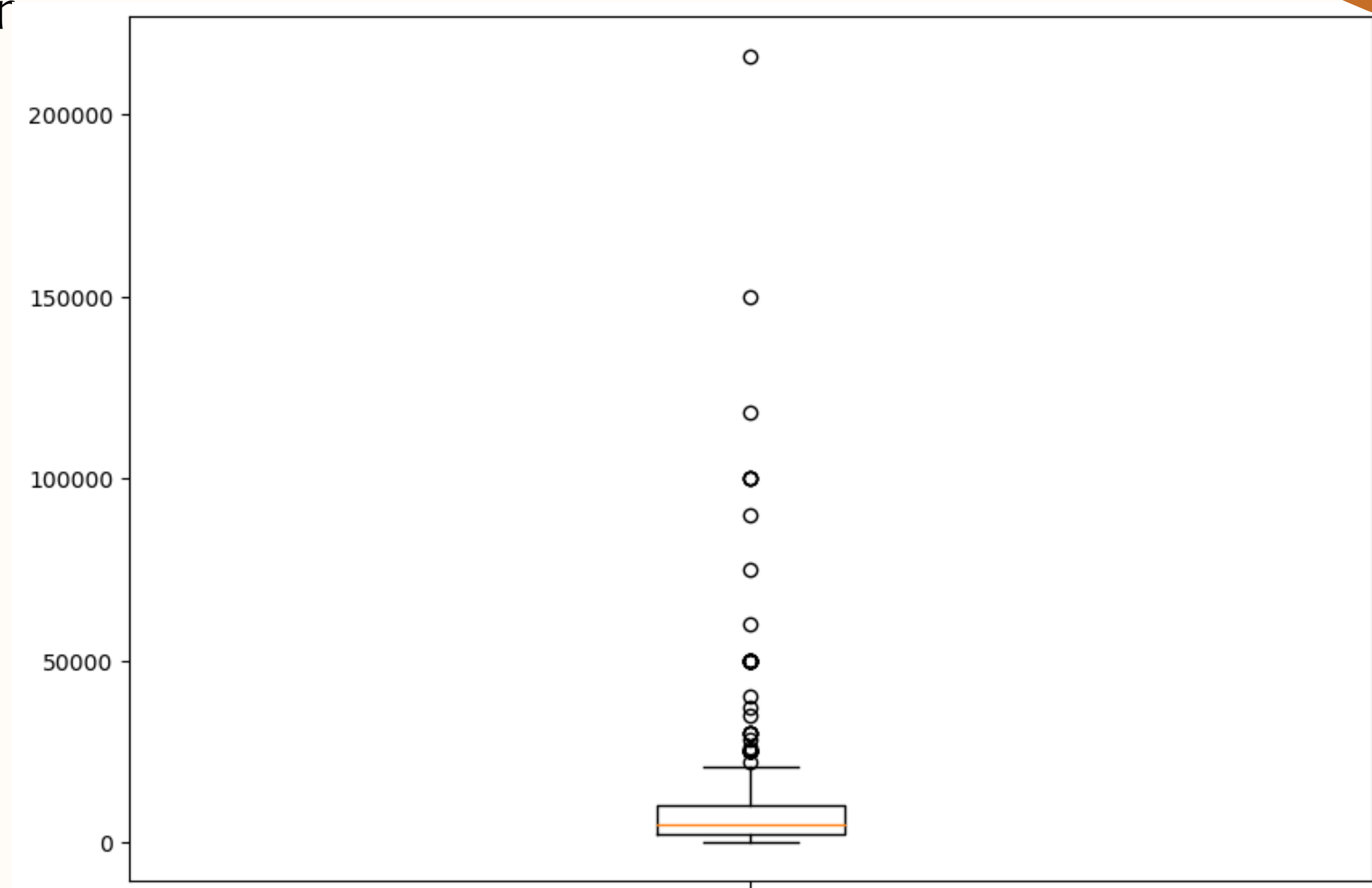




# 03

'AMOUNT

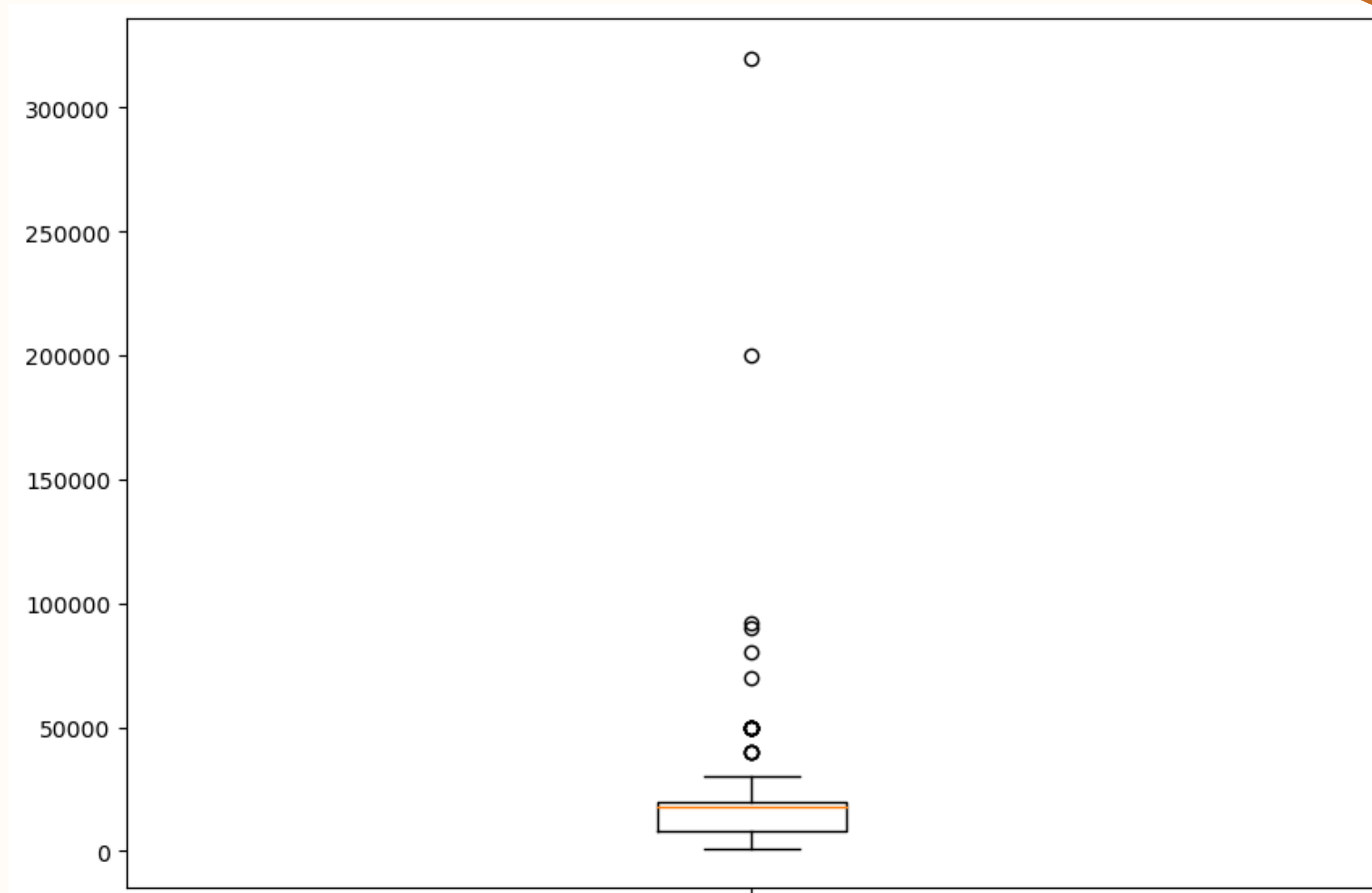
- Average amount for SGD - SGD\$8, 673



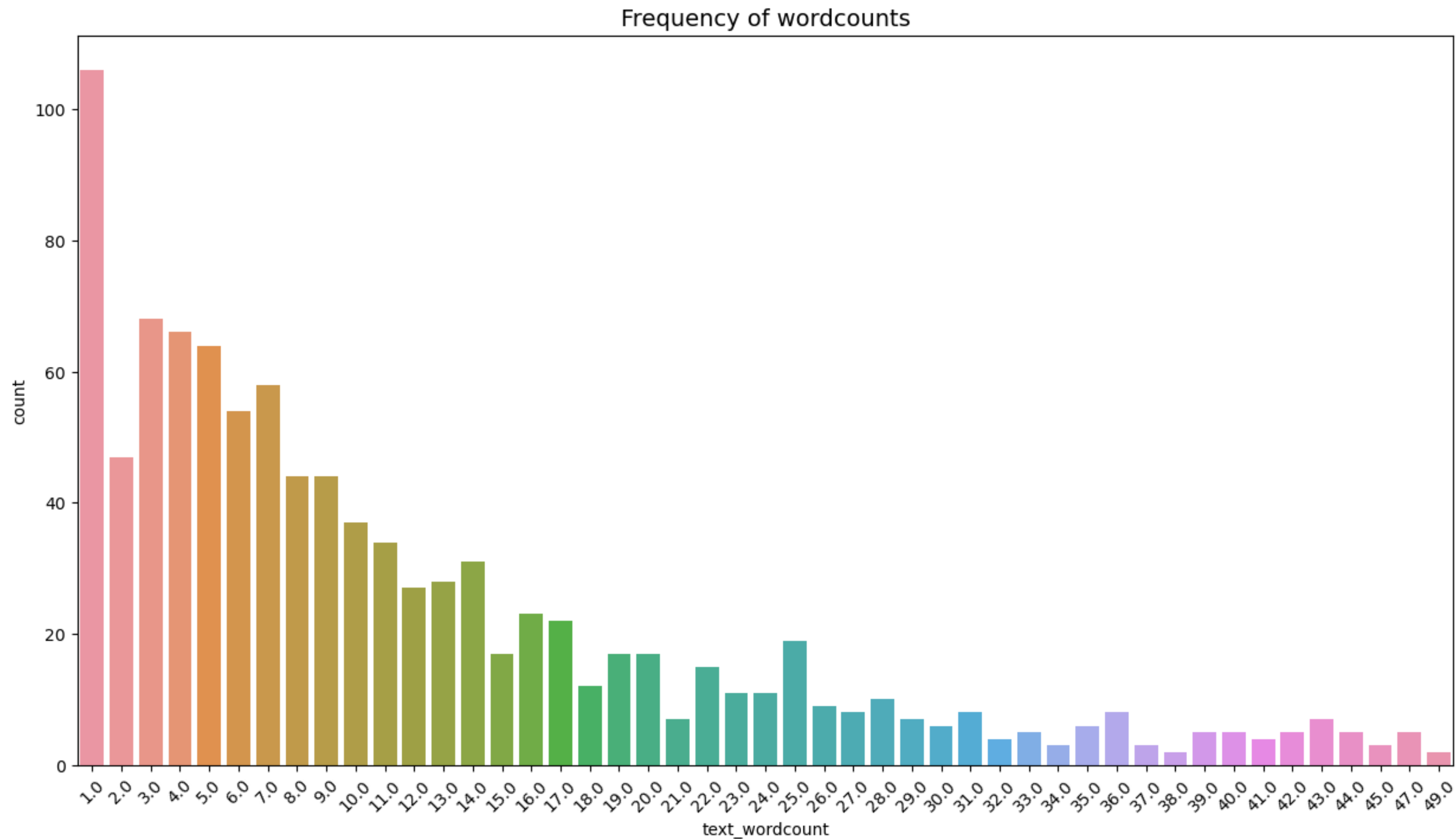
# 03

'AMOUNT

- Average amount for MYR - RM\$24,130



# Frequency of wordcounts





03

# Preprocessing, Text Feature Extraction

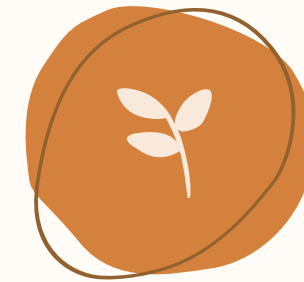
# Preprocessing steps



Language  
Detection



Remove non  
english rows



Remove special  
characters &  
white spaces



Tokenizing



Lemmatizing

# Text Feature Extraction

01

## Countvectorizer

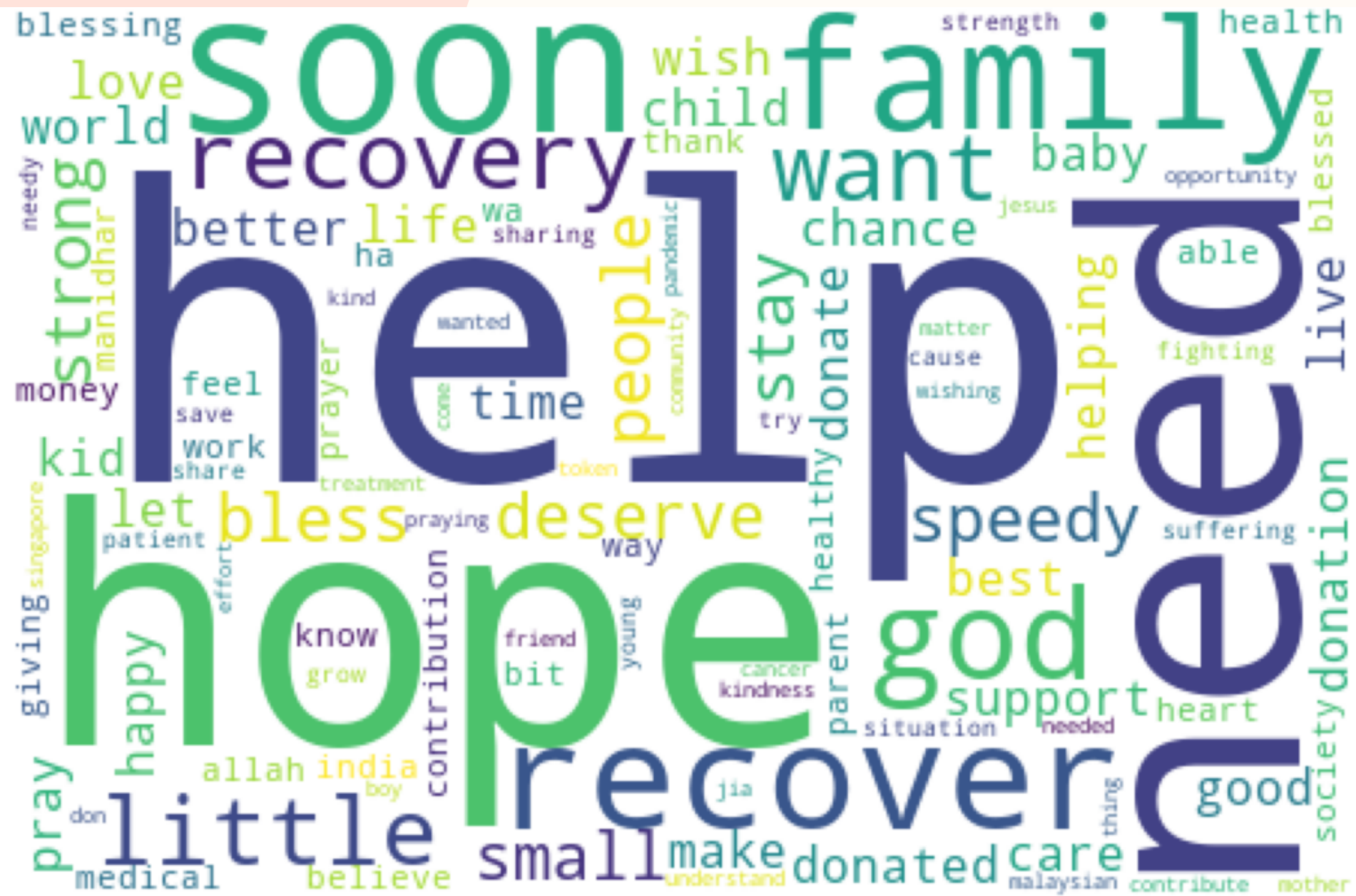
Fill the vector with frequency of each word as it appears in the document

02

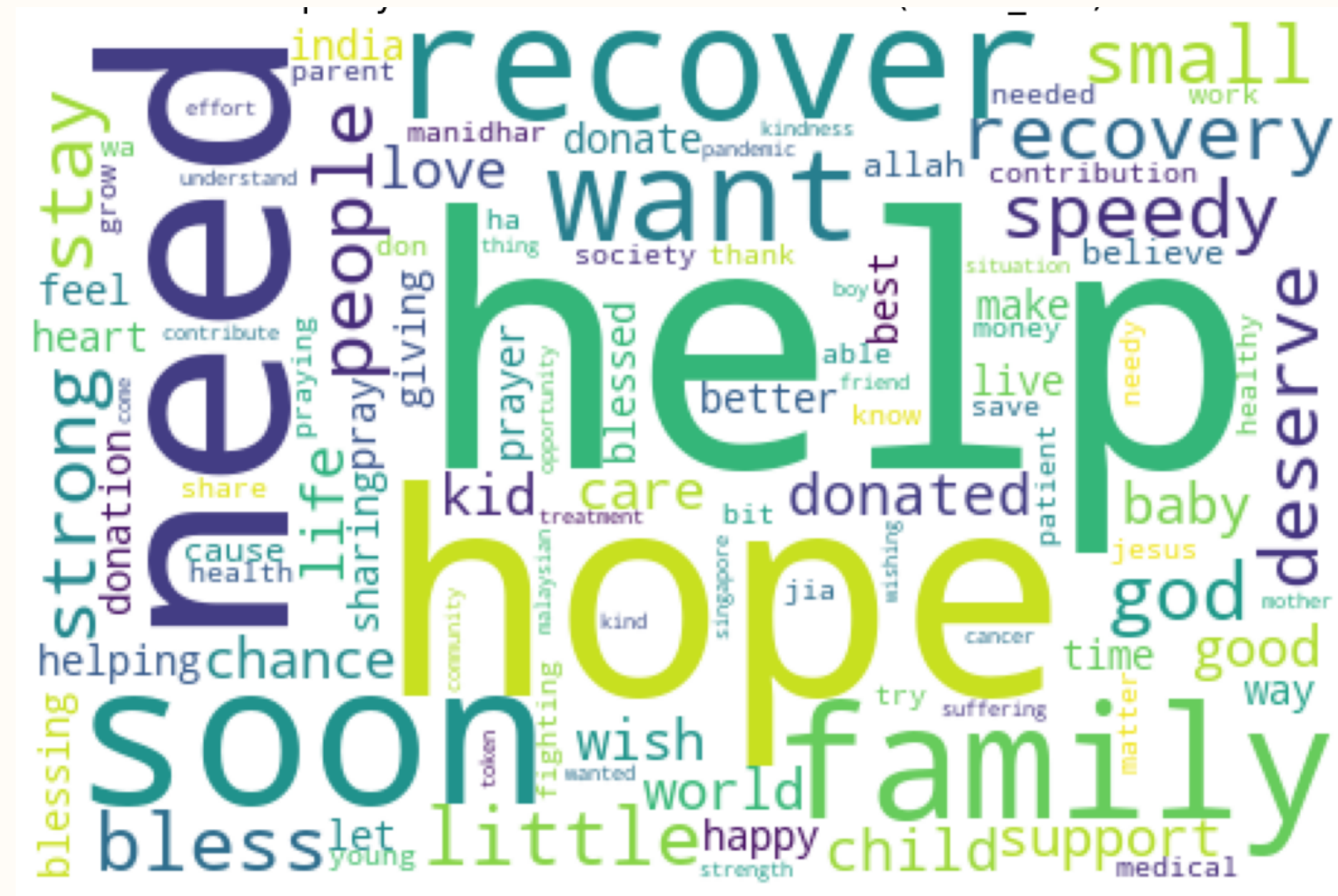
## Term-Frequency - Inverse Document Frequency (TF-IDF)

Assigns more weight to less frequently occurring words rather than frequently occurring ones

# Word Cloud



# Countvectorizer



# TF-IDF

04

# Clustering Algorithm Evaluation





# Algorithms



K-means

DSCAN

Affinity Propagation

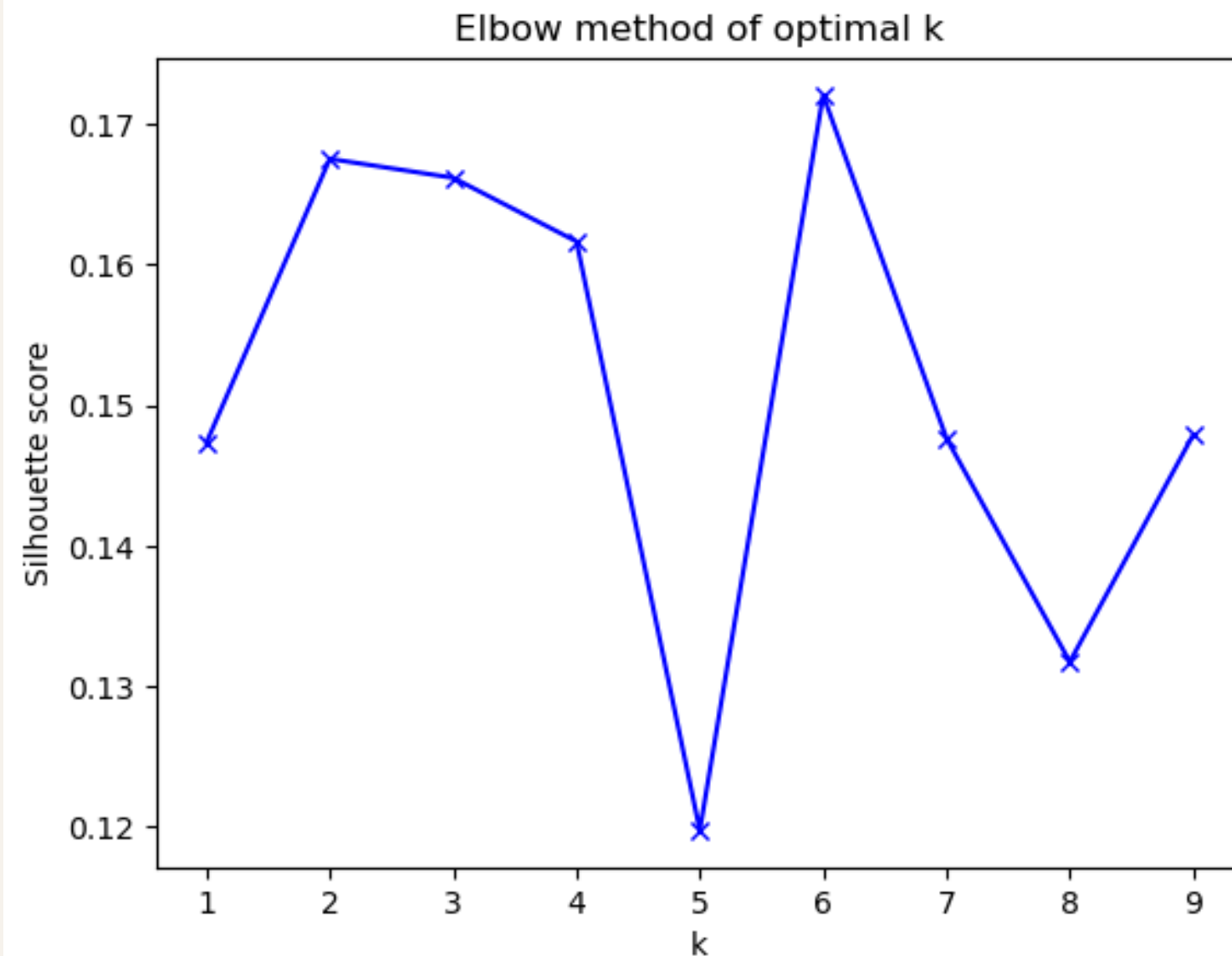


# Evaluation Metrics

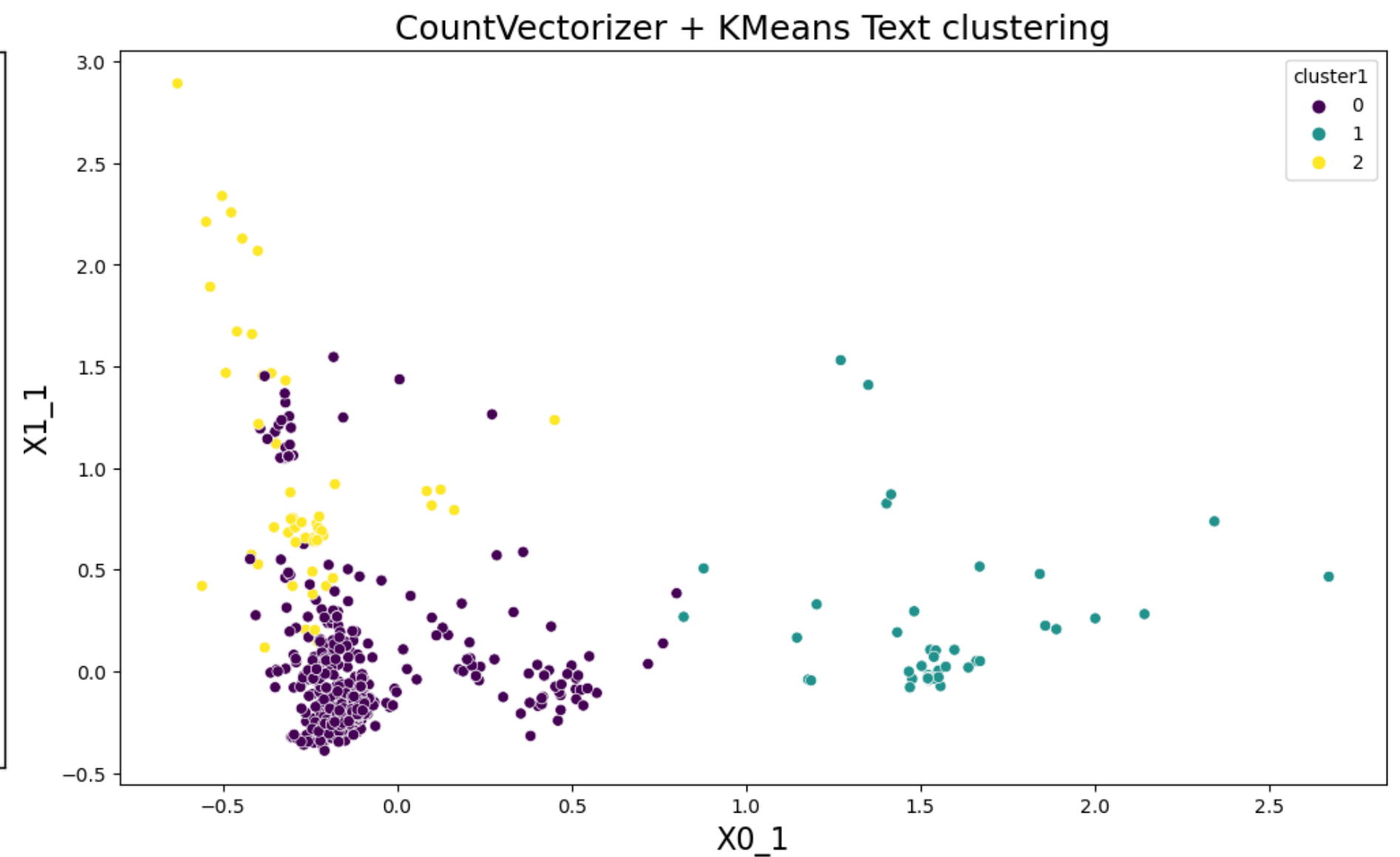
- Number of optimal clusters,  $k$  (using elbow method)
- Silhouette Coefficient
- Interpretability of clusters

# Selected model

## KMeans & CountVectorizer



Silhouette Coefficient: 0.1676



k=3

05


# Data Abstraction



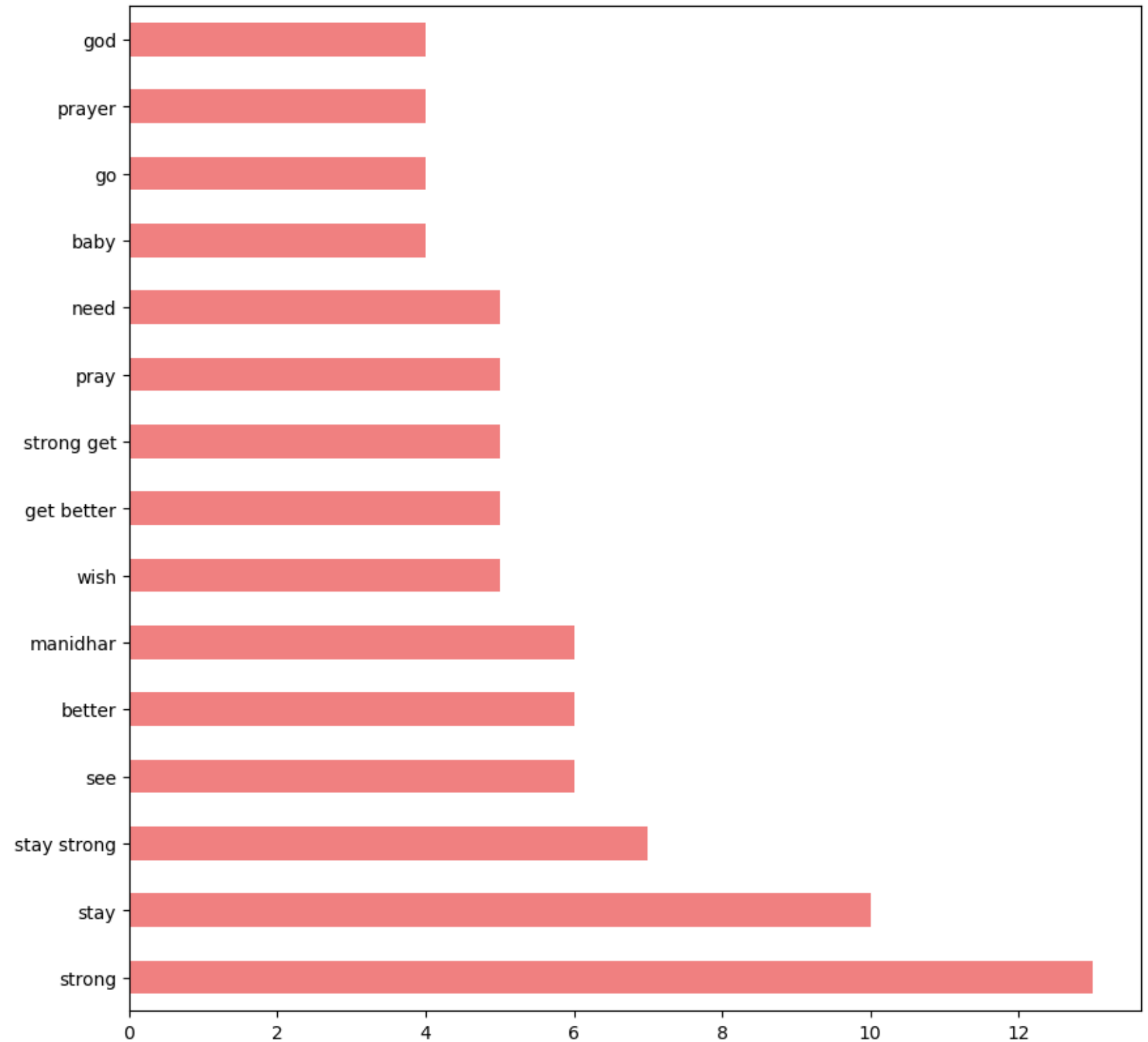
# Cluster 1

Guess what's the  
common theme?



- Imagine you are Joceline,'s parent, you'll definitely wish Joceline to get treatments, you wish someone can help you, someone can donate to you. Pray for Joceline will get better.
  - My prayers will always be with you. Stay strong, Get Well Soon ,
  - Be strong! Get well soon
  - I hope she'll get well soon ,
  - I am saddened to see a young boy like Manidhar suffer from blood cancer. I hope that with the little token - he can be given a new hope, an opportunity to get the treatment he needs to recover. I say "Stay strong and get well soon!!"
  - Get well soon...
  - Get well soon
  - Please get well soon, bro
  - Get well soon. God bless.
  - I know this is just a small amount but i know this can go a long way, my prayers are with you. I am also an ofw here and it is tough to be one, especially with your case, Get better soon! Stay strong!
- 

# Word Freq for Cluster 1





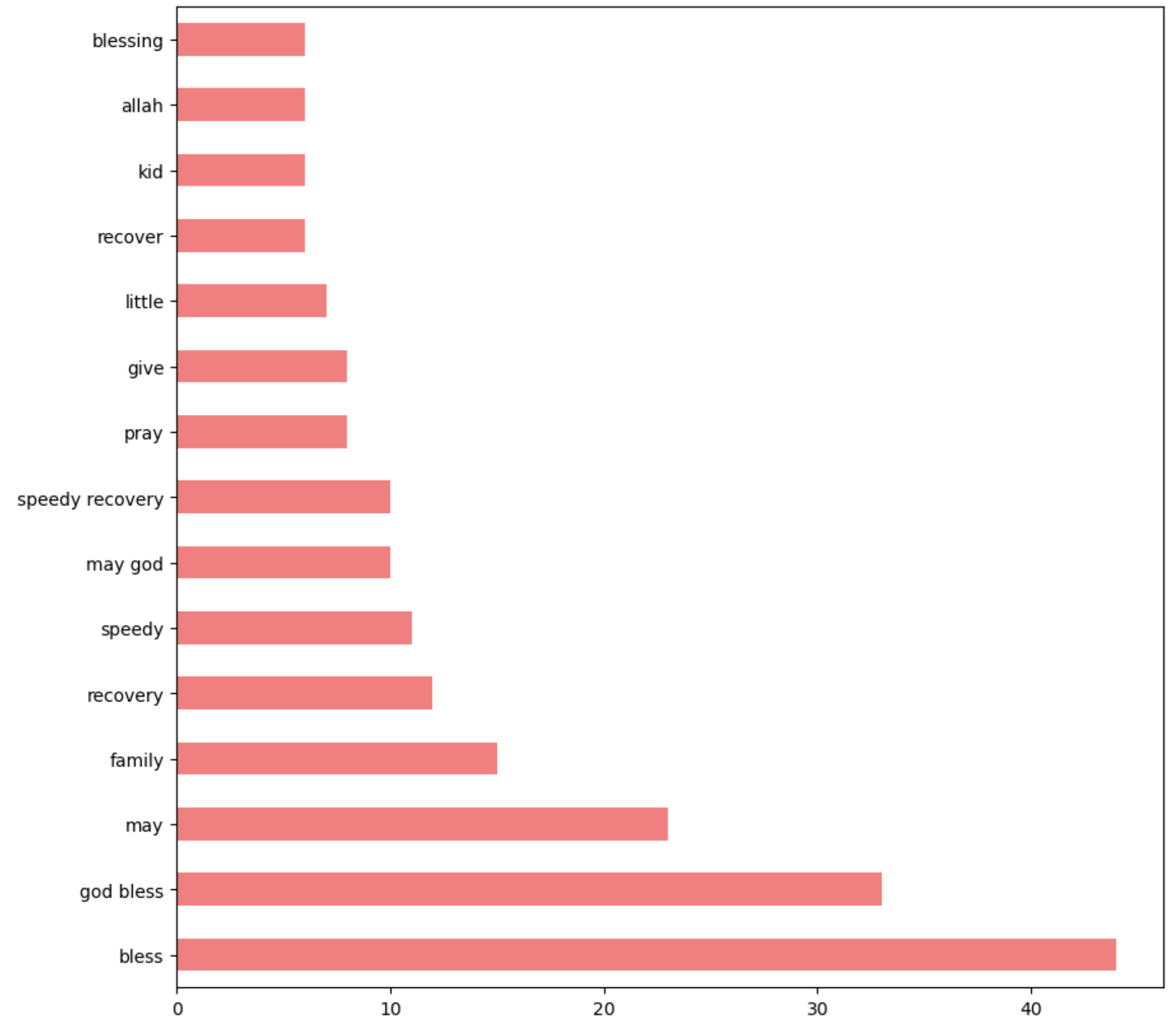
# Cluster 2

Guess what's the  
common theme?



- I believe all human beings deserved a second chance. Babies are a gift from Allah. Hopefully my small amount can help. May Allah protect this child from harm and may Allah also give strength to this family.
- May God's mercy be with this family and the kind doctor who proceeded the operation to save this little baby's life.
- May Allah bless every each one of us with health and hapiness in life.
- May God make it easy for Baby Niyaz and his parents.
- May god shower his blessings to Mani,speedy recovery to you dear.
- I donated because I want the family to know that there is still hope & do not give up. May Allah S.W.T give you strength to go through this difficult times & may you be rewarded for your patience in HIM.
- every bit helps. hoping the money i've donated and others have donated will greatly benefit Manidhar :) God bless!
- All living things are God creations and we must love them. Thank you to all the kind and loving people caring and taking care of these animals. I pray that God will bless them and all the animals.
- God bless you all I will try to help more when I can
- I will try to help when I can God bless you all

# Word Freq for Cluster 2








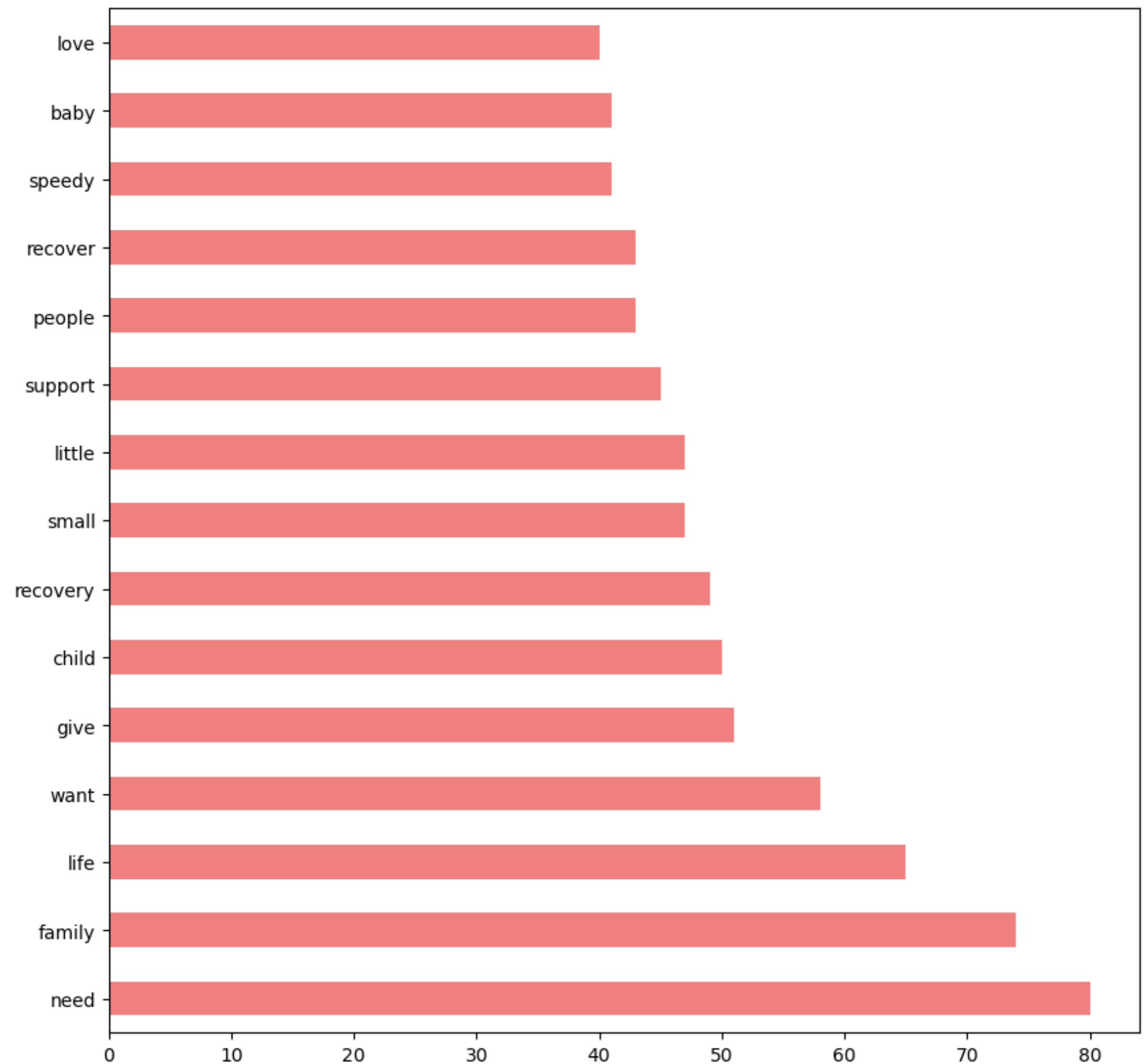
# Cluster O


Guess what's the  
common theme?




- Everyone, please donate any amount that you can. Every bit counts. Thank you!
  - This breaks my heart seeing a young baby like this as I have a child too and can't imagine if he's my own child. I hope with the small amount makes a different to you and will keep baby Niyaz in my prayers too.
  - Because it is the right thing to do.
  - I donated because I hope this small amount helps the family financially in the medical needs of the child.
  - I donated because every child deserves to lead a normal life
  - Doing what I can to help
  - Because I have children too.
  - I donated because I wanted to help this poor child of 10 months old suffering from liver disease and have about 6 months to live if there is no donor available. He was born as a normal child and has not even seen or chance to see the world.
  - My dad had the same condition
- 

# Word Freq for Cluster O





# What type of donors do we have?



## Cluster 1

Religious, faithful to their belief

## Cluster 2

'Get well soon' group

## Cluster O

Donors with relatable personal experiences, altruistic

06

# Final Evaluation

# How useful are the obtained results?

- Two minority groups were distinct in characteristics
- Great potential in using text clustering for targetted marketing purposes
- Downside is that biggest cluster can be broken into smaller clusters
- BERT (more advanced NLP technique) - contextual model can help to improve proper language representation for general-purpose language understanding by machines



# Thank You!

Do you have any questions for me before we go?