Capstone project 2 Milestone Report- Personality Prediction MBIT

Problem statement

Have you ever been curious about the personality type of yourself and people around you? Is there any simple test to help us find the answer? The answer is yes. You can figure out your personality type by simply answering these four questions:

- Are you outwardly or inwardly focused ?
- How do you prefer to take in information?
- How do you prefer to make decisions?
- How do you prefer to live your outer life?

For each question, there are two options to choose from. You choose the side that seems most natural for you. The options are: Introversion (**I**) or Extraversion (**E**), Sensing (**S**) or Intuition (**N**), Thinking (**T**) or Feeling (**F**) and Judging (**J**) or Perception (**P**) respectively. The introspective self-report questionnaire we introduced above is known as Myers-Briggs -Type- Indicator (MBTI). It is a personality type system that divides everyone into 16 distinct personalities based one their answers to the previous questions. For example, someone who chooses extraversion, sensing, thinking and judging would be labeled an ESTP in the MBTI system.

This project focuses on using machine learning algorithms in supervised machine learning to predict the personality of a person from the type of posts they put on social media.

Data Source

The data was collected through the PersonalityCare forum and is available on Kaggle Personality Prediction Dataset. This dataset consists of over 8675 rows representing 8675 different people and 2 columns representing a person's MBTI personality type and the things they have posted.

Exploratory Data Analysis

We first load the data to a DataFrame using **pd.read_csv** and look at the structure of the data set.

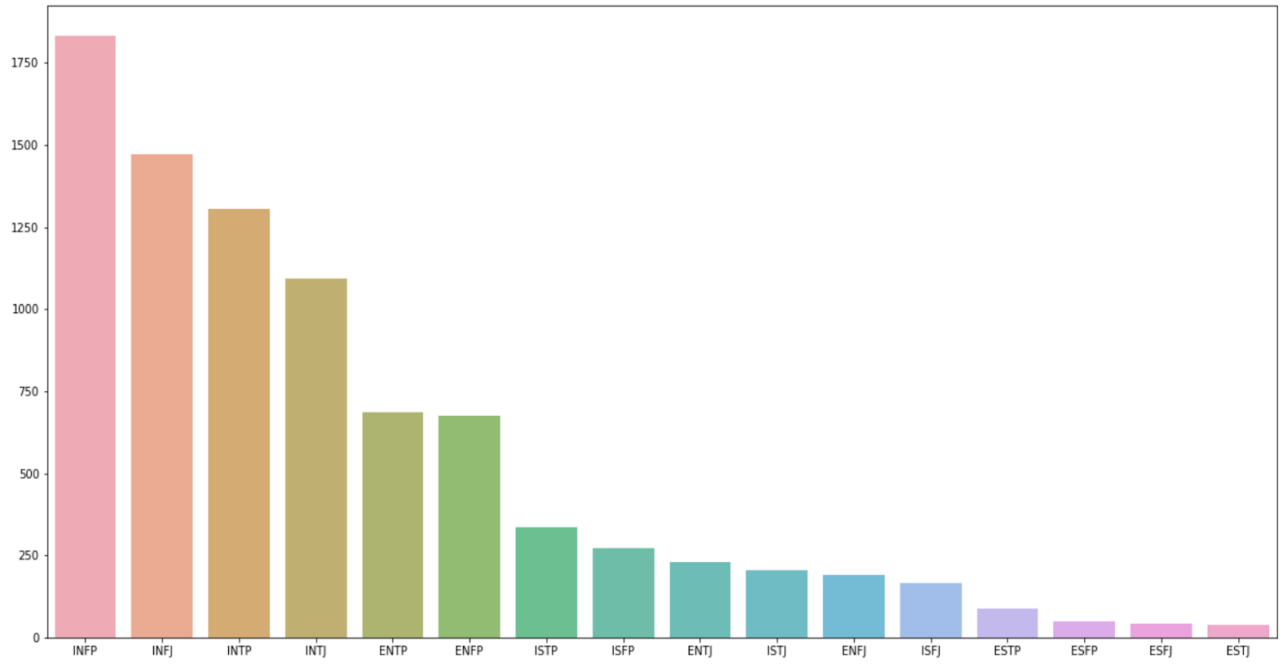
	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

The data has two columns: **type** and **posts** and no missing data. The two columns represent the personality type of 8675 different people and their corresponding posts on social media. The personality type is expressed by 4 letters representing the answers for 4 questions above. The first letter can either be **I**(Introversion) or **E**(Extraversion). The second one can either be **S**(sensing) or **N**(Intuition). The third one can either be **T**(Thinking) or **F**(Feeling). The last letter can either be **J**(Judging) or **P**(Perception). The first person in our table has personality type **INFJ** implying Introversion-Intuition-Feeling-Judging.

We are going to analyze the data set using common sense. It is natural to think that the length of a post, number of question marks, exclamation, ellipsis or images in a post somehow reveal the personality of the owner. We are going to find these numbers and add columns **post_length, question_per_post, excl_per_post, ellipsis_per_post, img_per_post** to our data **df**.

	type	posts	post_length	question_per_post	img_per_post	excl_per_post	ellipsis_per_post
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...	556	18	6	3	15
1	ENTP	'I'm finding the lack of me in these posts ver...	1170	5	1	0	19
2	INTP	'Good one _____ https://www.youtube.com/wat...	836	12	0	4	13
3	INTJ	'Dear INTP, I enjoyed our conversation the o...	1064	11	0	3	26
4	ENTJ	'You're fired.|||That's another silly misconce...	967	10	2	1	21

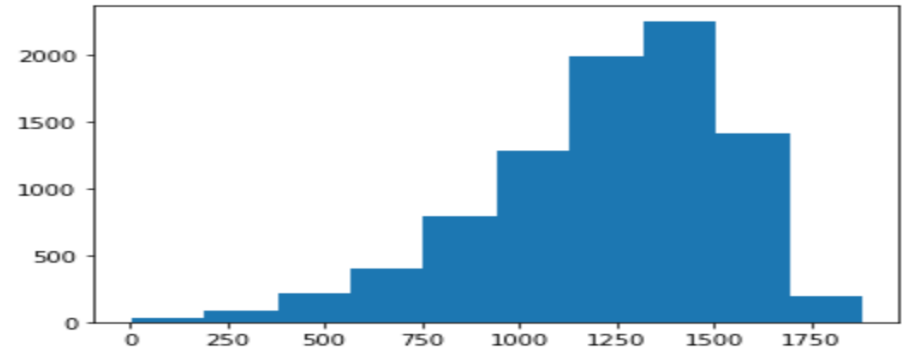
There are 16 personality types. How are they distributed in our data? We first want to explore the distribution of different personality types.



It is obvious that the distribution is unbalanced between different classes. The group **INFP(Introversion-Intuition-Feeling-Perception)**has the highest frequency while the group **ESTJ (Extraversion-Sensing-Thinking-Judging)** has the lowest frequency.

The largest group is **INFP (Introversion-Intuition-Feeling-Perception)** with 1832 posts, while the smallest group is **ESTJ (Extraversion-Sensing-Thinking-Judging)** with only 39 posts. We observe that the majority of posts (5697) in our data belong to **IN** (Introversion-Intuition) people while the **ES**(Extraversion_Sensing) people contribute only 218 posts. The difference in frequencies is really significant. The group with highest frequency is almost triple the group with 5th highest frequency, and 46 times more than the group with the lowest frequency.

We also want to look at the distribution of the length of posts.



The majority of posts in our data are more than 500 words. The models based on this data will work best for posts or texts with more than 500 words.

We now look at the relations between the type of personality and the length of post, as well as the number of question marks, exclamations, ellipsis and images per post.

type	post_length
ESFJ	1290.476190
ENFJ	1286.584211
INFJ	1278.431973
ENFP	1260.770370
INFP	1244.552948
ISFJ	1241.295181
ESTJ	1229.538462
ENTJ	1218.086580
ISTJ	1213.224390
ENTP	1205.995620
INTP	1197.763037
INTJ	1194.577452
ISTP	1165.563798
ESTP	1162.595506
ISFP	1136.383764
ESFP	1022.125000

type	question_per_post
ESTP	12.292135
ENTJ	12.000000
ESFP	11.854167
ENFP	11.362963
ISTP	11.080119
INTP	11.078988
ENTP	11.048175
ISFP	10.819188
ISTJ	10.746341
INTJ	10.742438
ENFJ	10.652632
ESTJ	10.487179
INFJ	10.448299
INFP	10.126638
ISFJ	10.048193
ESFJ	8.809524

type	excl_per_post
ENFP	16.939259
ENFJ	13.800000
ESFJ	11.523810
ESFP	10.958333
ISFJ	10.475904
ISFP	10.398524
INFP	9.322052
INFJ	9.077551
ENTJ	8.298701
ESTJ	8.205128
ESTP	8.000000
ISTJ	7.980488
ENTP	7.129927
INTJ	5.262145
INTP	5.004601
ISTP	4.629080

The **ESFJ** (Extraversion-Sensing-Feeling-Judging) tends to have the longest post(average 1290 words per post) while the **ESFP** (Extraversion-Sensing-Feeling-Perception) tends to have the shortest post (average 1022 words per post).

People from the **ESTP** (Extraversion-Sensing-Thinking-Perception) tend to use more question marks in their post (average 12.3 question marks per post) while people in the **ESFJ** (Extraversion-Sensing-Feeling-Judging) tend to use less question marks in their posts (average 8.8 question marks per post). The difference in question marks between groups is not significant compared to the length of these posts.

The group **ENFP** (Extraversion-Intuition-Feeling-Perception) use the most exclamation (Average 16.9 exclamation per post) while the group **ISTP** (Introversion-Sensing-Thinking-Perception) use the least exclamation (average 4.6 exclamation per post).

Data Preprocessing

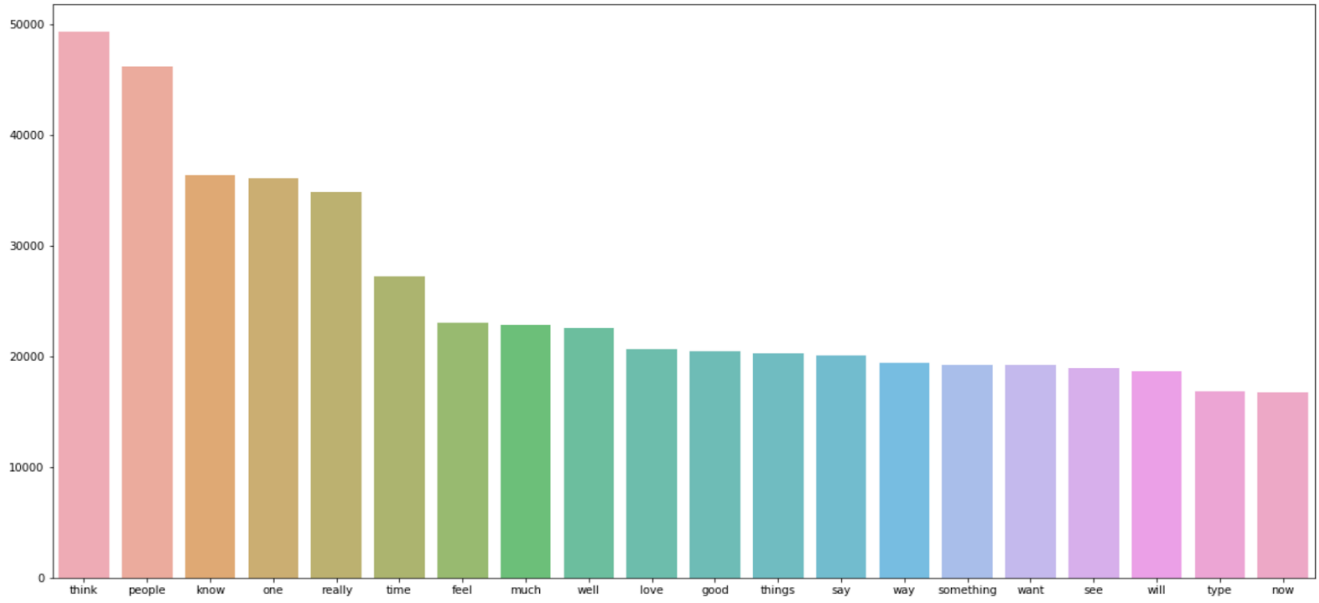
We are going to create 3 functions which play an important role in preprocessing text. They are : **clean_text**, **combine_text** and **text_preprocessing**.

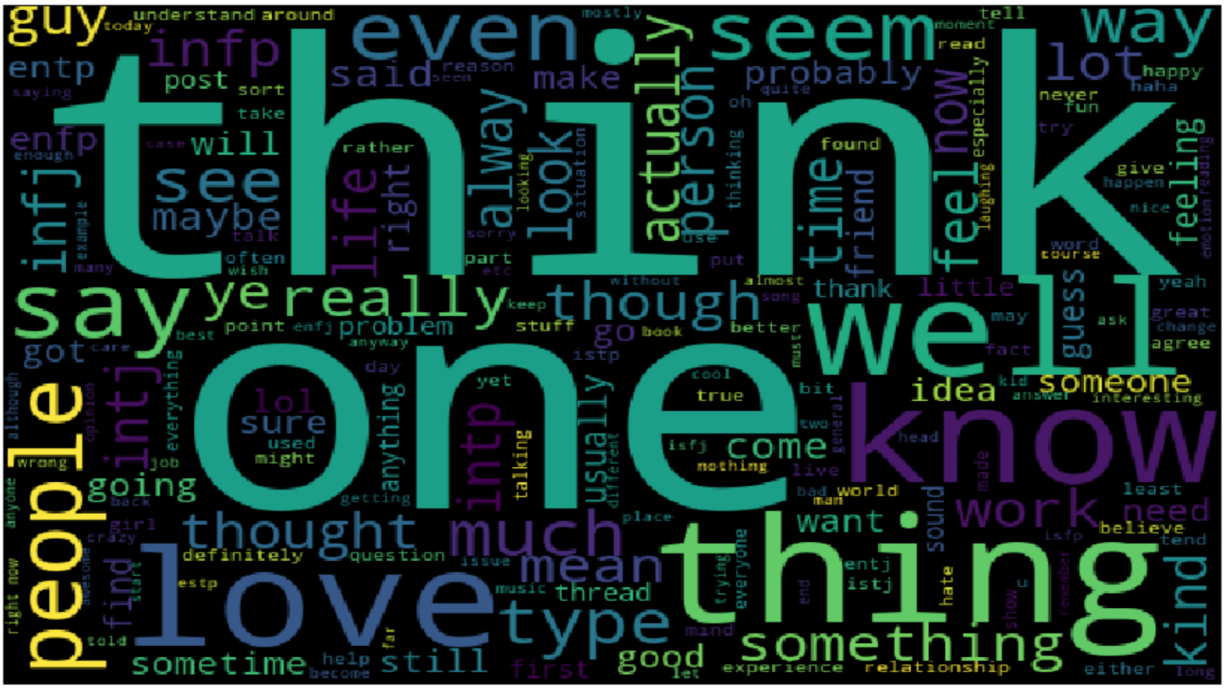
- **clean_text** function takes the argument text, the make the text lowercase, remove text in square brackets, remove links, remove punctuation and remove words containing numbers.
- **text_preprocessing** function takes text as an argument. It first applies the function clean_text on the argument, tokenizes the result, removes all stop words before combining all tokens.
- **combine_text** function combine single words into text.

The **clean_posts** column is added which contains the corresponding posts which have been preprocessed.

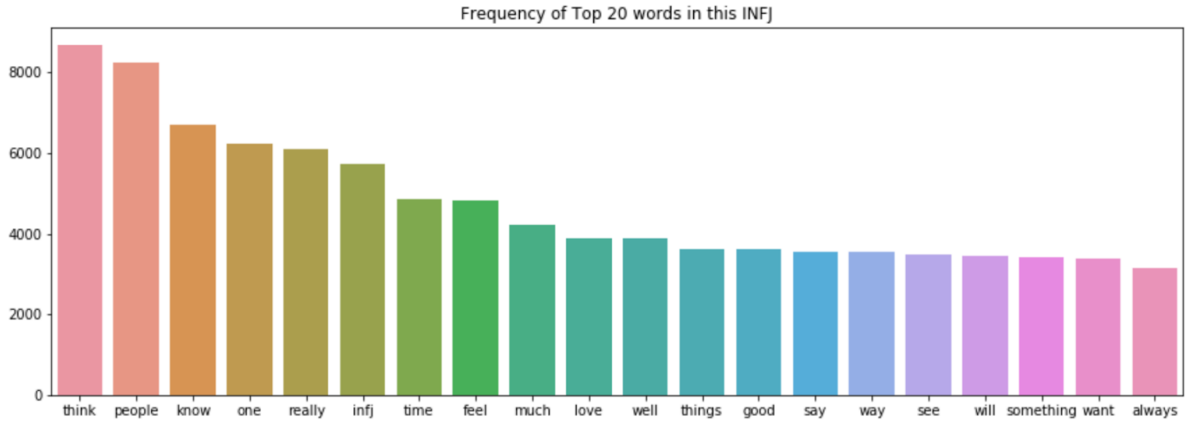
Most Used Words and Word Cloud.

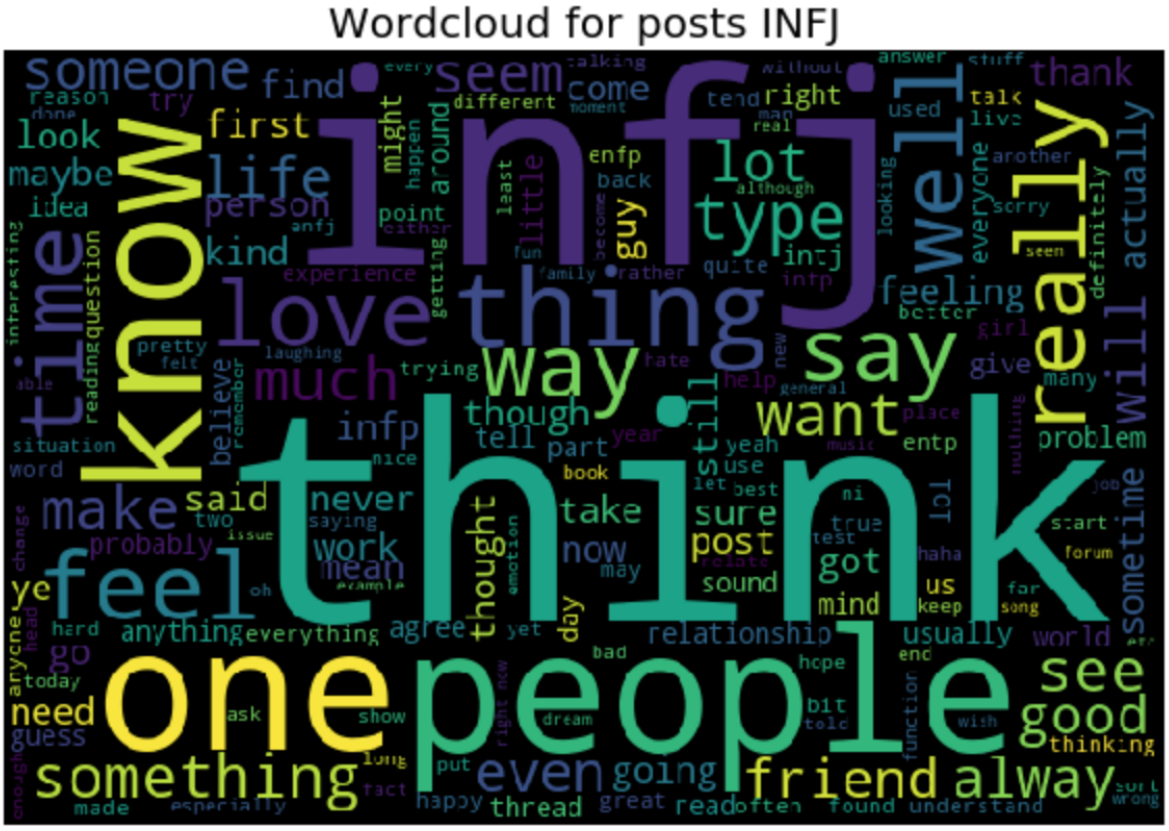
- Most Used Words and Word Cloud in all posts.



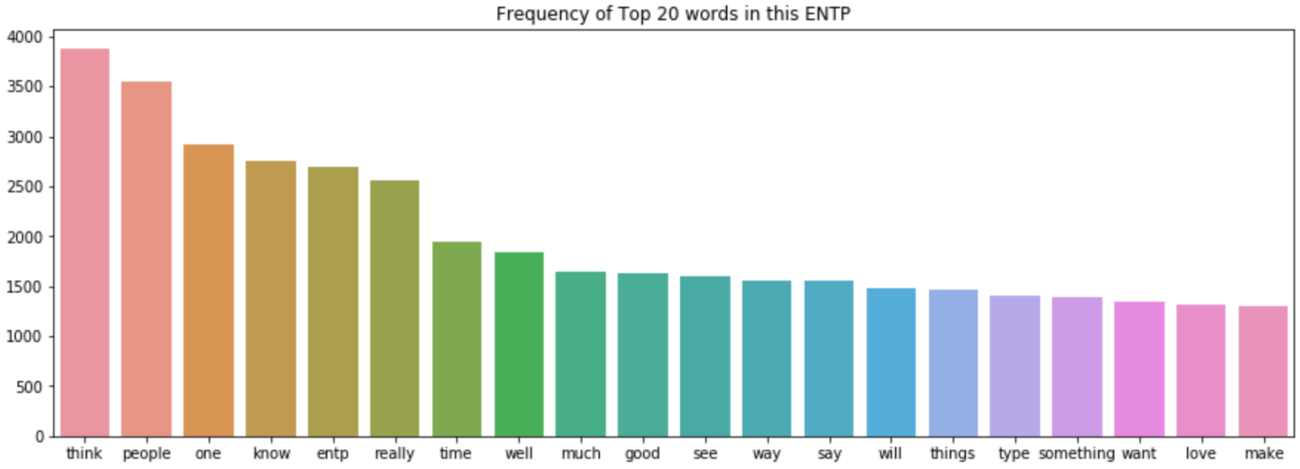


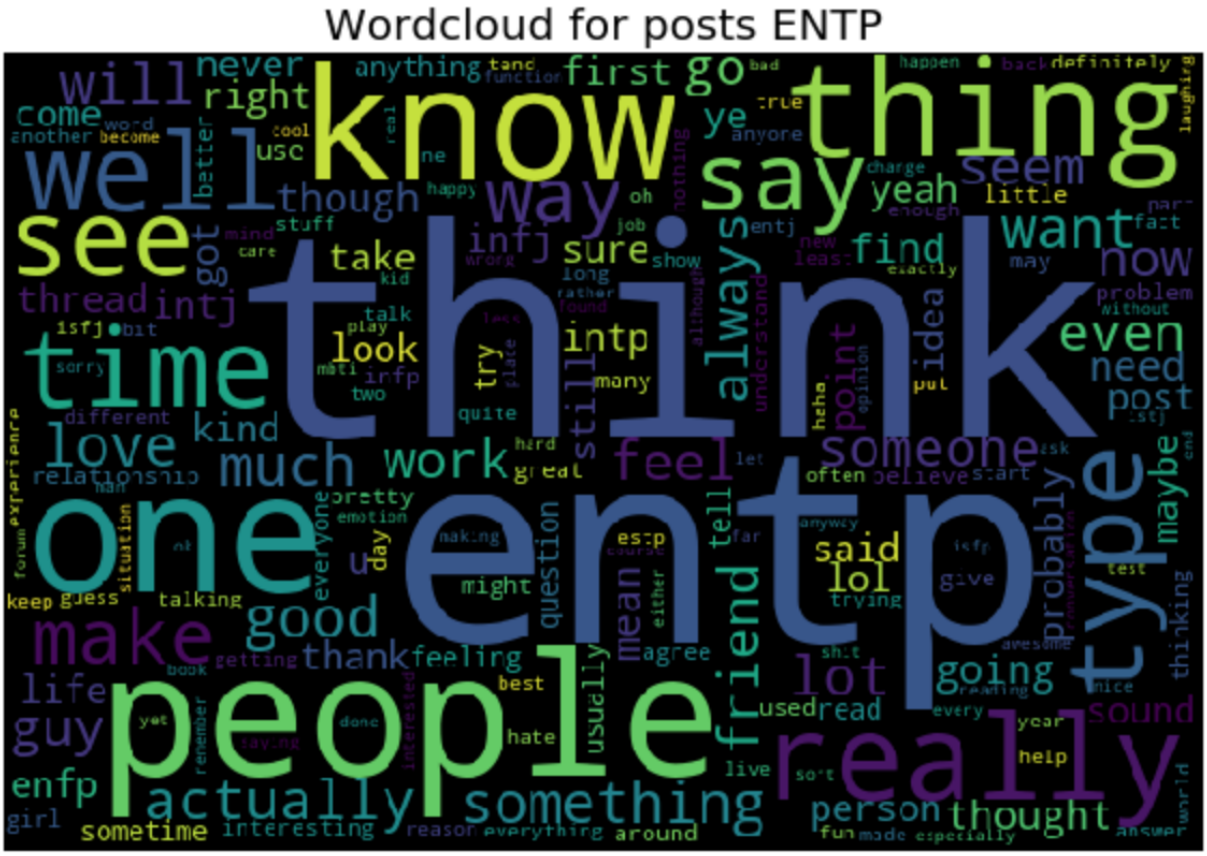
- Most Used Words and Word Cloud in each personality group
- INFJ



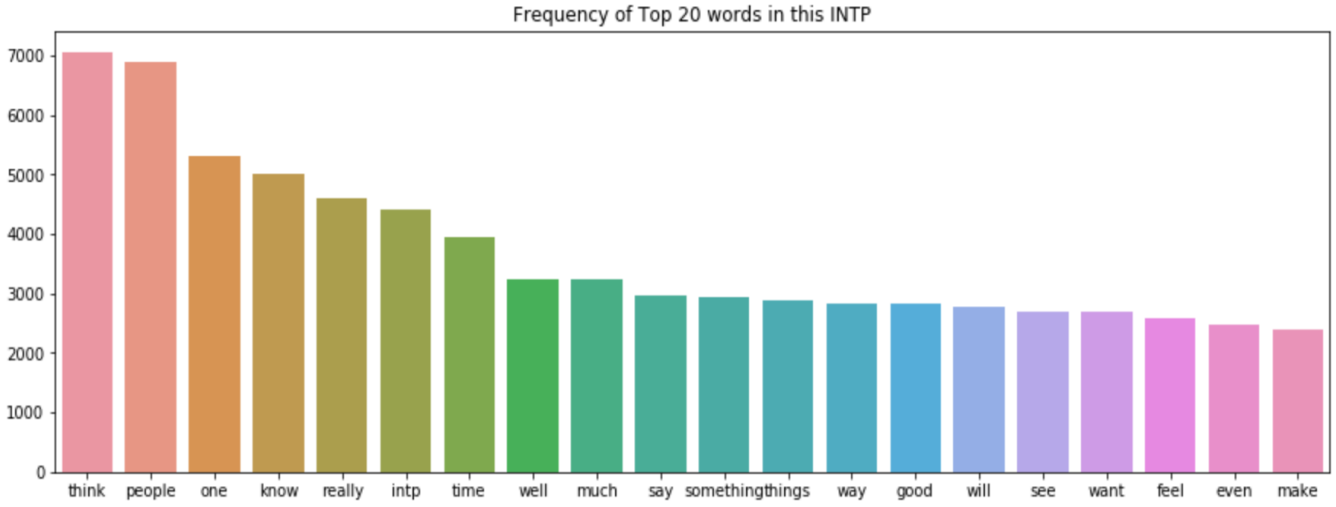


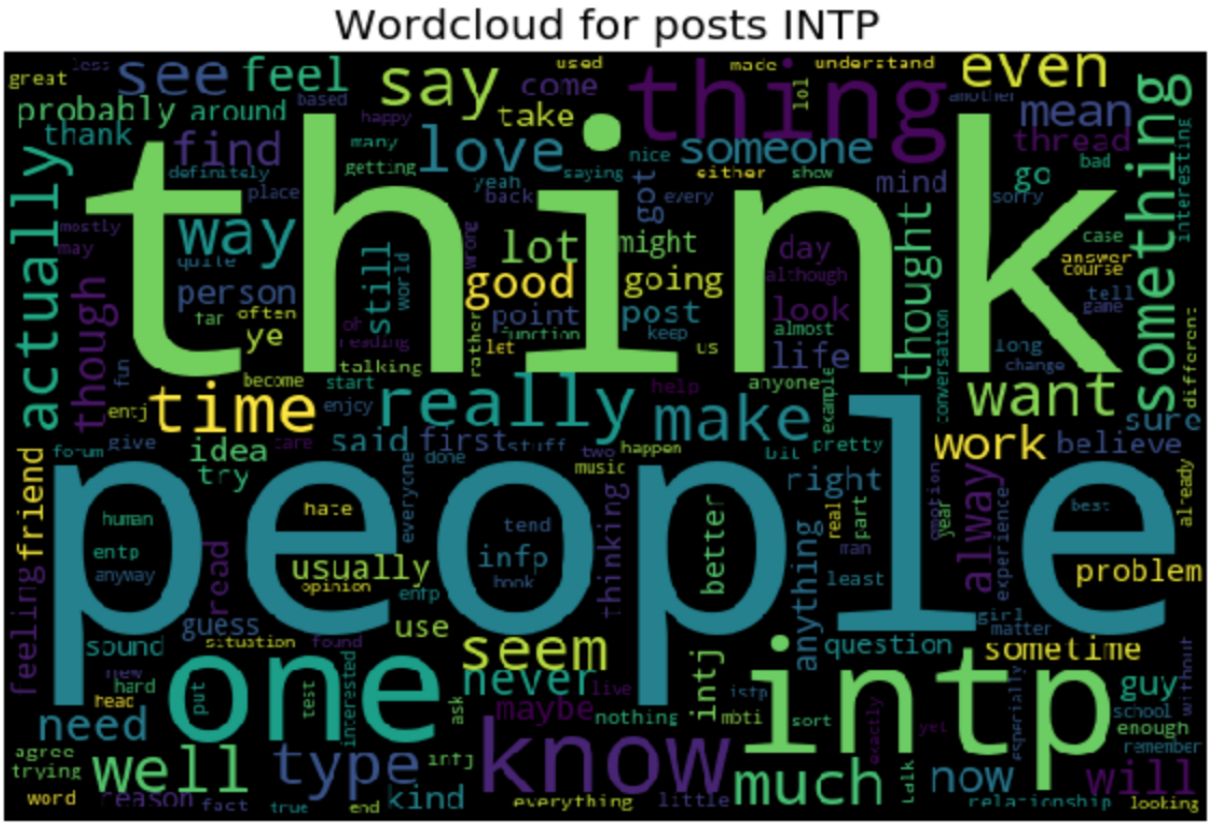
- ENTP


Frequency of Top 20 words in this ENTP
4000
3500
3000
2500
2000
1500
1000
500
0
think
people
one
know
entp
really
time
well
much
good
see
way
say
will
things
type
something
want
love
make

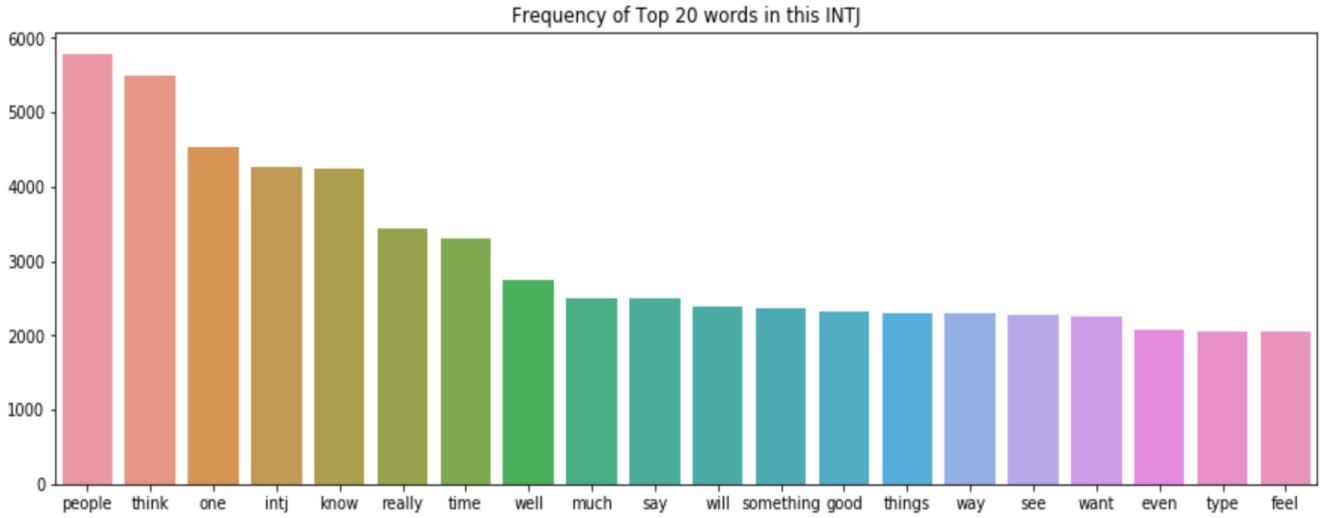

Wordcloud for posts ENTP
think
know
thing
entp
people
one
really
time
well
see
say
make
love
life
guy
enfp
actually
something
person
thought
lot
going
sound
type
maybe
probably
friend
feel
work
much
kind
look
take
thread
intj
intp
infj
sure
always
seem
yeah
want
find
even
need
post
someone
idea
point
still
tell
mean
question
good
thank
feeling
agree
usually
said
lol
give
trying
sometime
interesting
everything
around
thinking
sorry
different
relationship
experience
situation
talking
guess
keep
hate
best
yet
remember
getting
book
girl
reason
especially
definitely
anything
never
right
come
another
become
though
way
first
go
ye
anyore
usc
stuff
mind
care
kid
talk
play
two
many
quite
hard
great
let
often
believe
start
ask
hahaha
opinion
pull
nice
awesome
year
help
live
sort
world
made
fun
answer
laughing
happen
back
bad
true
cool
real
nothing
happy
oh
job
entj
new
least
exactly
may
fast
problem
without
long
rather
found
although
understand
wrong
show
isfj
bit
infp
enj
istj
esfp
esfj
intj
test
shit
used
read
every
thanks

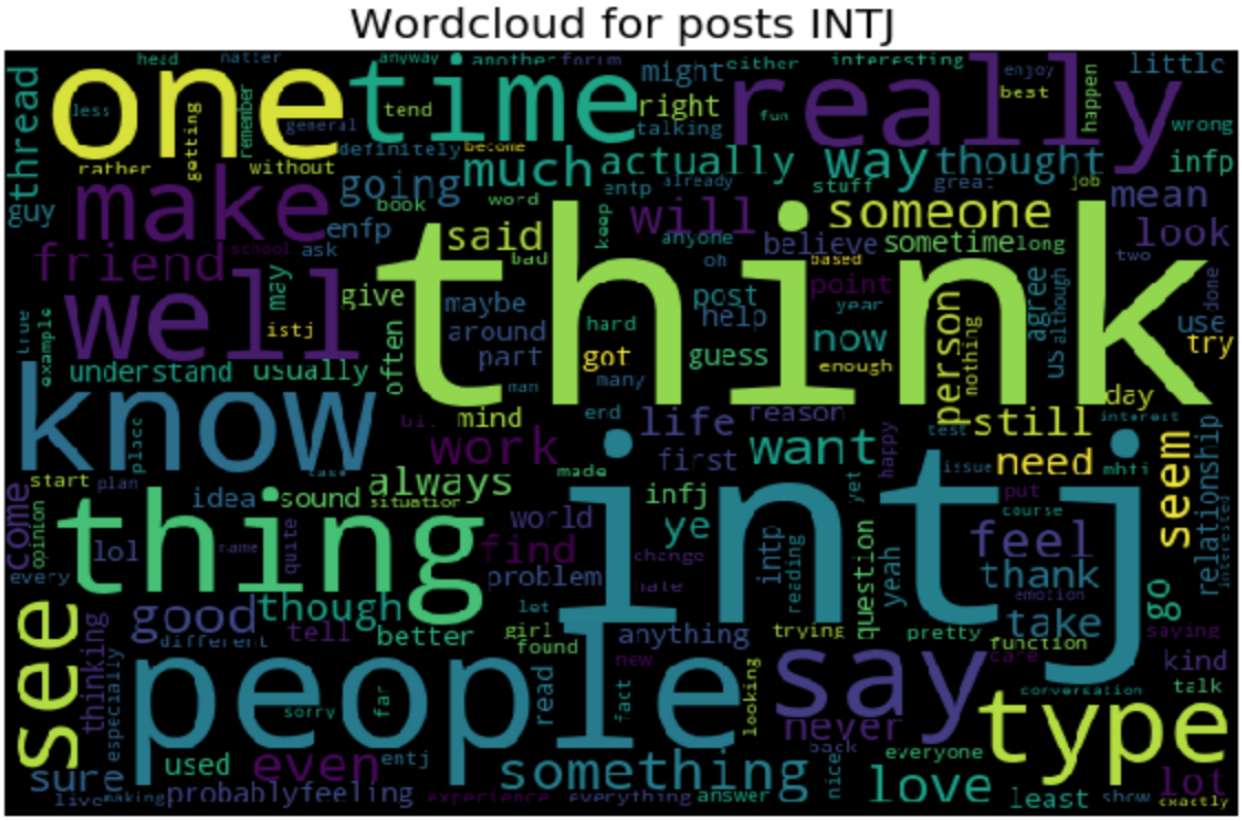
- INTP


Frequency of Top 20 words in this INTP
7000
6000
5000
4000
3000
2000
1000
0
think
people
one
know
really
intp
time
well
much
say
somethingthings
way
good
will
see
want
feel
even
make

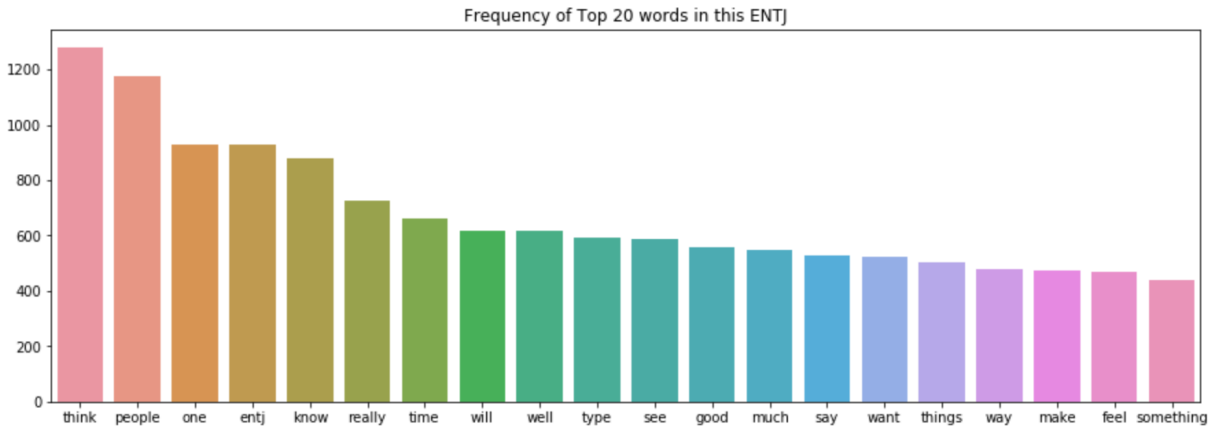

Wordcloud for posts INTP
think
people
intp
one
know
thing
something
really
time
actually
make
want
love
feel
see
say
even
mean
thread
probably
around
thank
find
way
lot
might
good
going
person
though
still
point
post
look
life
thought
mind
work
believe
sure
idea
try
said
first
friend
feeling
usually
seem
never
maybe
need
well
type
much
kind
now
will
guy
problem
sometime
question
anything
always
right
better
infp
intj
entp
hate
read
sound
guess
use
nothing
everything
little
relationship
looking
agree
trying
word
reason
fact
true
end
hard
human
anyway
opinion
situation
found
interested
test
put
ask
mbti
istp
sort
talk
yet
remember
enough
school
girl
matter
experience
already
best
interesting
different
conversation
answer
course
case
tell
game
long
change
sorry
bad
nice
saying
either
show
yeah
back
every
getting
many
happy
based
used
understand
come
take
someone
mind
go
day
although
almost
keep
us
let
rather
anyone
example
help
start
become
enjoy
talking
talk
give
forum
stuff
two
happen
music
bit
pretty
done
everyone
tend
thinking
part
real
least
year
ye
far
often
quite
may
mostly
place
definitely
back
yeah
world
more
reading
function
fun
entj

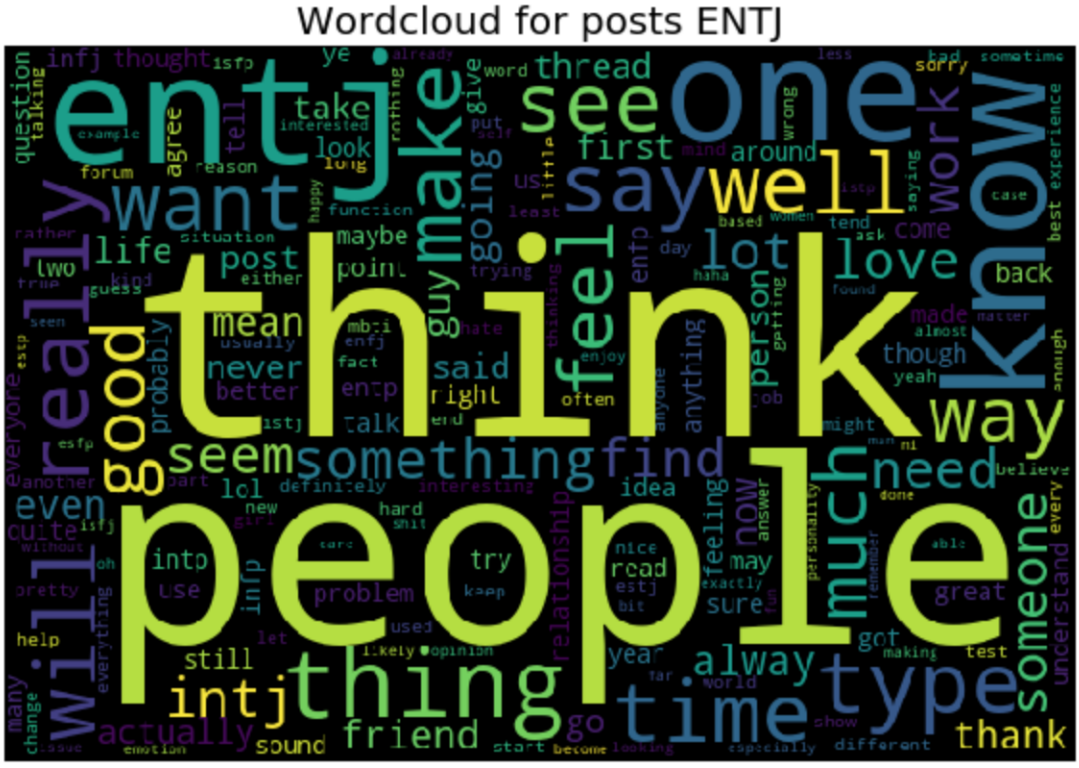
- INTJ


Frequency of Top 20 words in this INTJ
6000
5000
4000
3000
2000
1000
0
people
think
one
intj
know
really
time
well
much
say
will
something
good
things
way
see
want
even
type
feel


Wordcloud for posts INTJ
people
think
intj
thing
know
one
time
really
make
well
friend
see
say
type
something
love
sure
even
probablyfeeling
used
good
though
better
tell
thinking
sound
idea
always
work
life
want
need
still
seem
relationship
feel
thank
take
go
kind
talk
lot
least
never
question
yeah
find
problem
world
ye
infj
intp
infp
thought
actually
way
much
going
someone
said
will
believe
sometime
look
mean
thread
guy
enfp
istj
maybe
around
part
give
post
help
point
person
agree
use
try
day
reason
first
made
start
plan
come
opinion
every
lol
quite
different
tend
right
might
talking
without
rather
getting
remember
general
definitely
become
book
word
ask
school
bad
keep
anyone
oh
based
long
two
stuff
great
job
already
entp
hard
got
guess
now
enough
many
man
erd
mind
often
may
true
example
place
interesting
another
anyway
either
enjoy
best
happen
wrong
littlc
fun
test
issue
mbti
put
course
interest
interested
change
idle
reading
yet
happy
yeah
trying
pretty
function
care
emotion
saying
conversation
everyone
everything
answer
experience
looking
fact
read
found
girl
let
new
back
nice
sorry
far
entj
live
making
especially
exactly
show
done
although
nothing
anything
understand
usually
lot

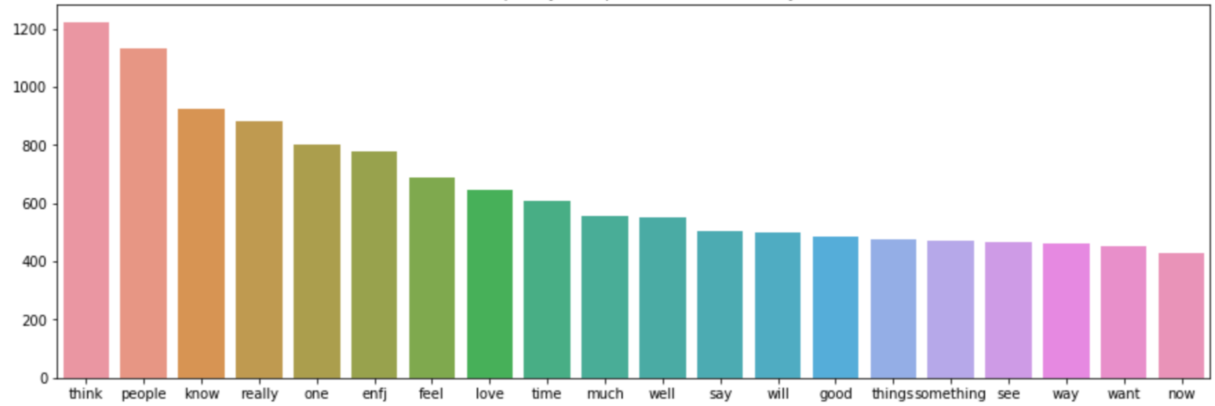
- INTJ


Frequency of Top 20 words in this ENTJ
1200
1000
800
600
400
200
0
think
people
one
entj
know
really
time
will
well
type
see
good
much
say
want
things
way
make
feel
something


Wordcloud for posts ENTJ
think
people
entj
one
know
really
will
want
make
see
feel
say
well
thing
time
type
good
something
find
much
need
way
love
lot
life
work
thank
intj
actually
friend
seem
mean
never
still
always
someone
relationship
problem
infp
intp
infj
thought
question
talking
agree
take
look
first
around
thread
post
maybe
point
guy
going
lwo
back
come
though
made
person
probably
better
said
right
talk
idea
feeling
sure
great
read
try
use
even
quite
lol
definitely
interesting
hard
new
part
another
everyone
everything
sound
start
become
looking
especially
different
sorry
sometime
best experience
understand
believe
annoy
yeah
almost
matter
found
haha
trying
thinking
anything
enjoy
often
anyone
getting
job
might
ni
done
able
every
remember
personality
answer
nice
estj
exactly
bit
fun
far
world
making
test
got
year
likely
opinion
used
let
keep
help
pretty
without
oh
shit
girl
change
many
issue
emotion
show
less
mind
wrong
saying
case
based
women
tend
ask
day
entp
enfj
fact
usually
seem
hate
mbti
istp
esfp
esfj
istj
kind
guess
true
rather
situation
function
happy
long
reason
forum
example
self
put
give
word
already
ye
isfp
interested
tell
nothing
little
us
least

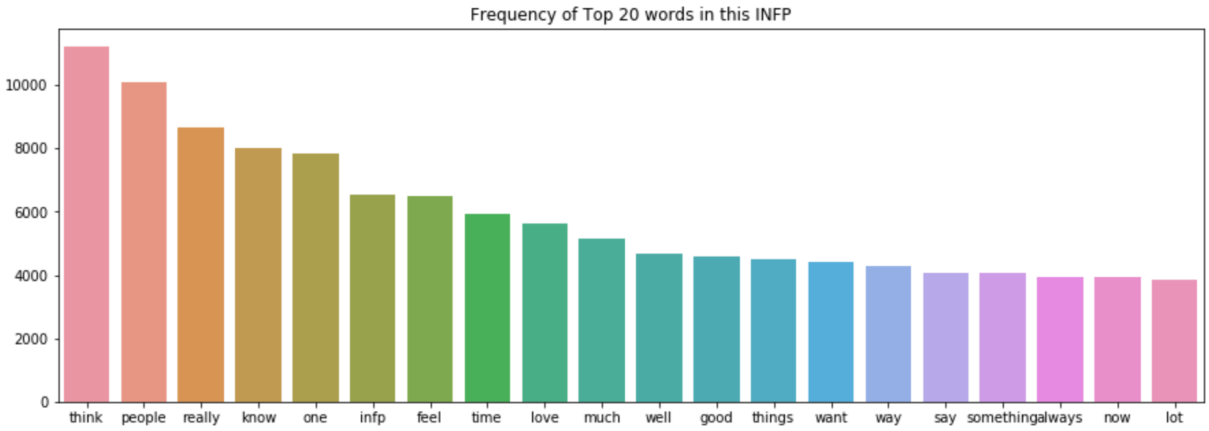
- **ENFJ**

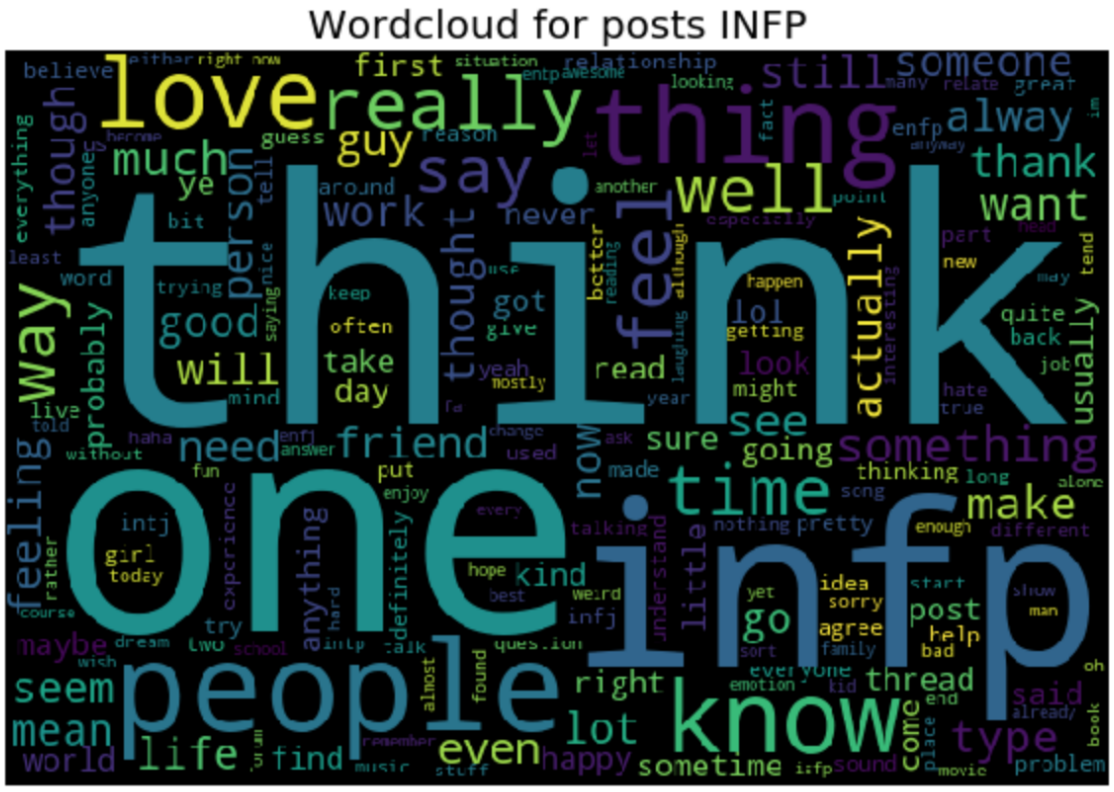
Frequency of Top 20 words in this ENFJ



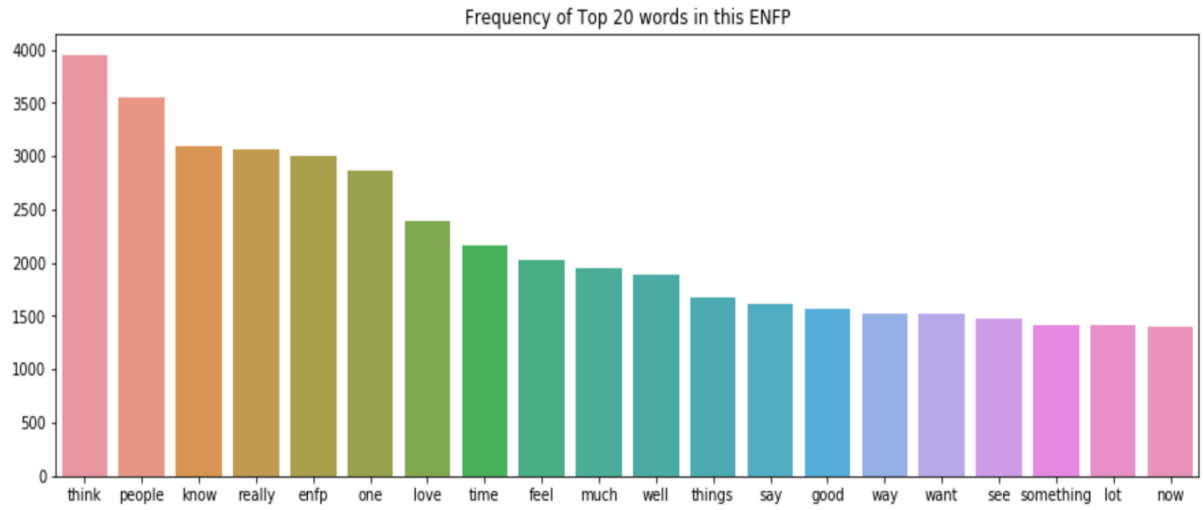
Wordcloud for posts ENFJ

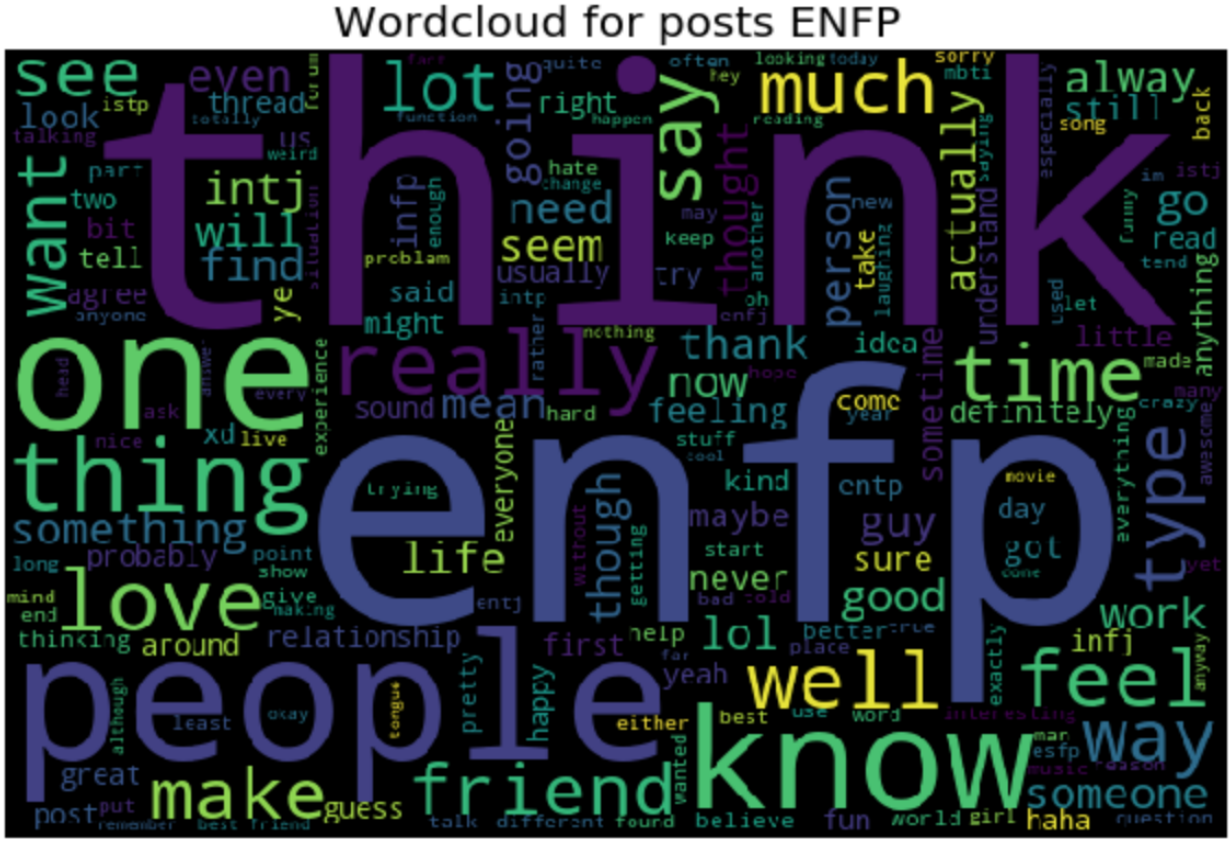
- INFP


Frequency of Top 20 words in this INFP
10000
8000
6000
4000
2000
0
think
people
really
know
one
infp
feel
time
love
much
well
good
things
want
way
say
somethingalways
now
lot

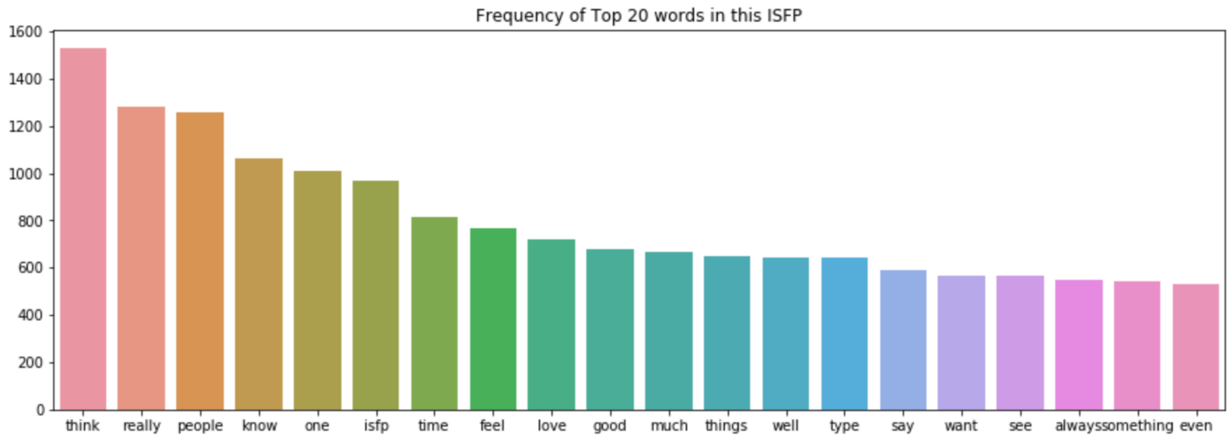

Wordcloud for posts INFP
think
one
infp
people
love
really
know
thing
feel
time
life
mean
seem
world
find
even
happy
sometime
lot
right
thread
said
type
make
something
actually
well
still
someone
always
thank
want
guy
say
much
work
never
thought
good
will
take
day
need
friend
see
going
sure
now
way
feeling
probably
person
though
believe
first
relationship
situation
awesome
looking
reason
guess
around
another
lol
look
might
read
got
give
yeah
mostly
thinking
song
make
kind
hope
weird
infj
go
post
help
bad
agree
sorry
idea
start
slow
man
come
place
movie
problem
sound
stuff
music
remember
found
almost
talking
understand
little
nothing
pretty
enough
different
alone
long
end
already
book
oh
emotion
kid
family
sort
yet
everyone
question
talk
dream
maybe
wish
school
two
try
hard
definitely
anything
expericnce
rather
course
girl
today
intj
enjoy
put
made
used
change
ask
answer
enfj
haha
without
fun
mind
live
told
probably
word
trying
saying
nice
keep
often
bit
tell
ye
person
anyone
everything
least
better
reading
laughing
although
interesting
happen
getting
yeah
year
hate
true
back
quite
usually
tend
may
new
part
head
job
enfp
awesome
looking
relate
great
many
fact
let
another
point
especially
reason
hermap
right now
neither
believe
first
situation
relationship
entp
awesome
looking
still
someone
always
thank
want
guy
say
much
work
never
thought
good
will
take
day
need
friend
see
going
sure
now
way
feeling
probably
person
though
love
really
know
thing
feel
time
life
mean
seem
world
find
even
happy
sometime
lot
right
thread
said
type
make
something
actually
well

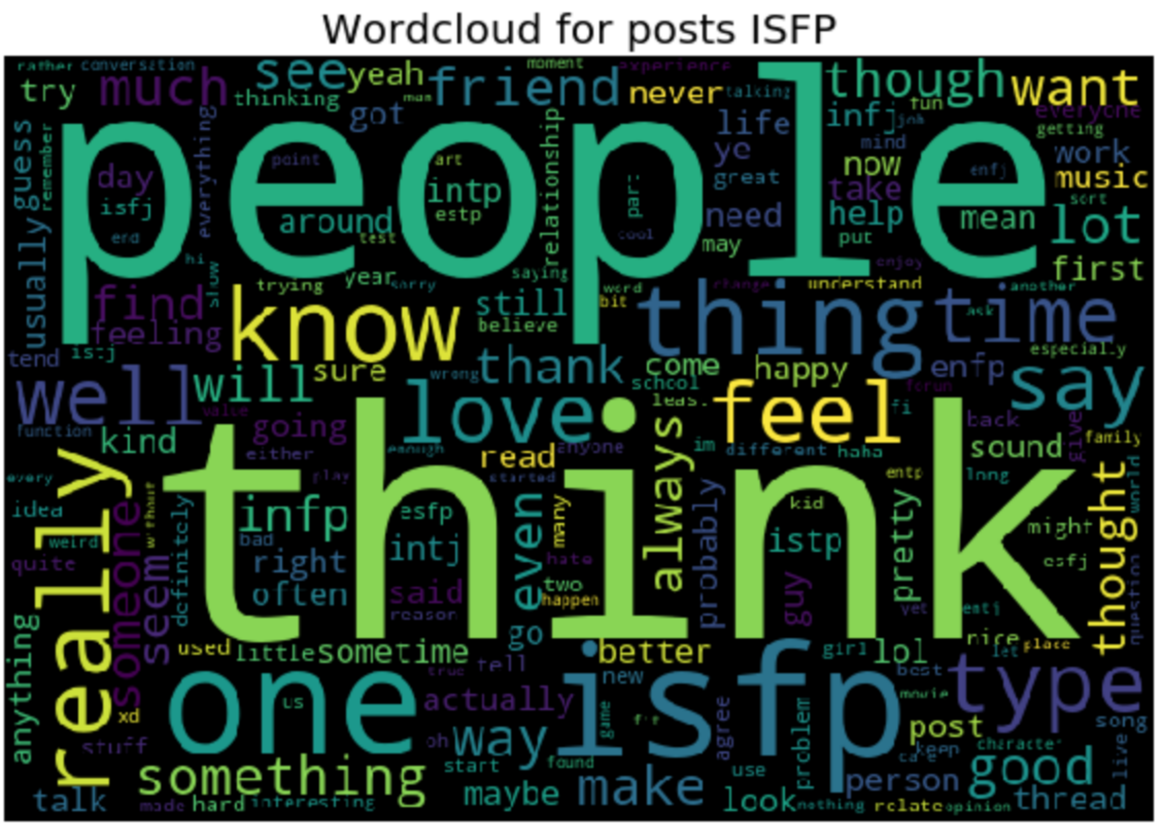
- ENFP


Frequency of Top 20 words in this ENFP
4000
3500
3000
2500
2000
1500
1000
500
0
think
people
know
really
enfp
one
love
time
feel
much
well
things
say
good
way
want
see
something
lot
now

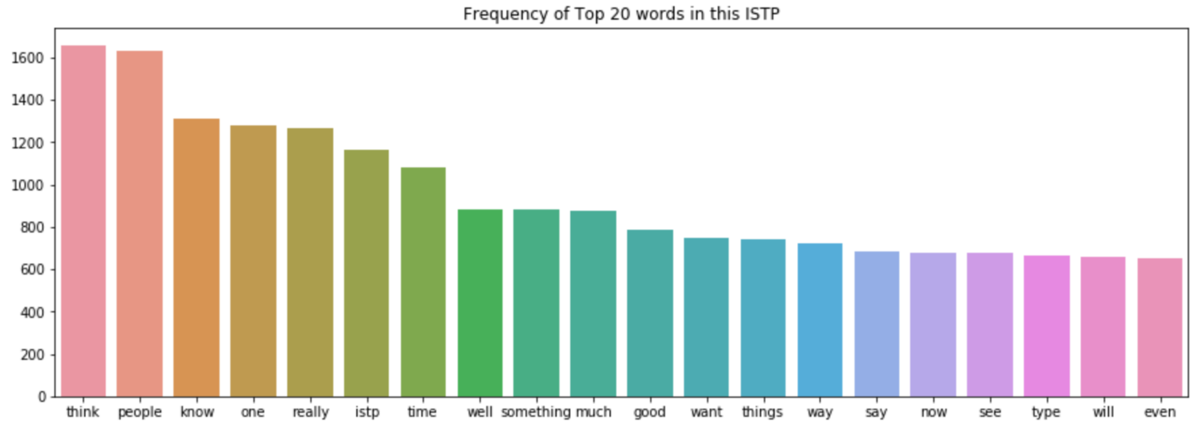

Wordcloud for posts ENFP
think
enfp
people
one
really
know
lot
much
see
feel
time
love
thing
make
friend
well
life
way
something
probably
relationship
around
look
want
intj
will
find
need
seem
usually
said
might
thank
now
feeling
sound
mean
everyone
though
maybe
kind
guy
sure
good
never
lol
work
infj
someone
haha
definitely
actually
understand
person
thought
say
go
read
anything
type
always
still
thread
even
right
going
hate
tell
two
bit
part
agree
sometime
idea
little
made
crazy
awesome
movie
day
got
yeah
first
happy
pretty
great
least
okay
post
guess
talk
different
found
believe
fun
world
girl
question
trying
stuff
cool
hope
hard
nice
live
xd
ask
experience
relationship
thinking
mind
end
long
point
show
give
making
entj
getting
without
start
bad
cold
help
place
better
true
exactly
anyway
interesting
music
reason
esfp
infp
problem
enough
change
keep
another
try
oh
intp
enfj
nothing
rather
laughing
take
new
may
song
saying
reading
function
happen
quite
often
hey
looking
today
sorry
mbti
especially
back
im
istj
funny
tend
used
let
anything
many
awesome
everything
comc
year
come
yet
although
tongue
either
best
use
word
wanted
remember
best friend
put
istp
totally
talking
weird
forum
situation
ye
agree
anyone
bit
two
part
us

- ISFP

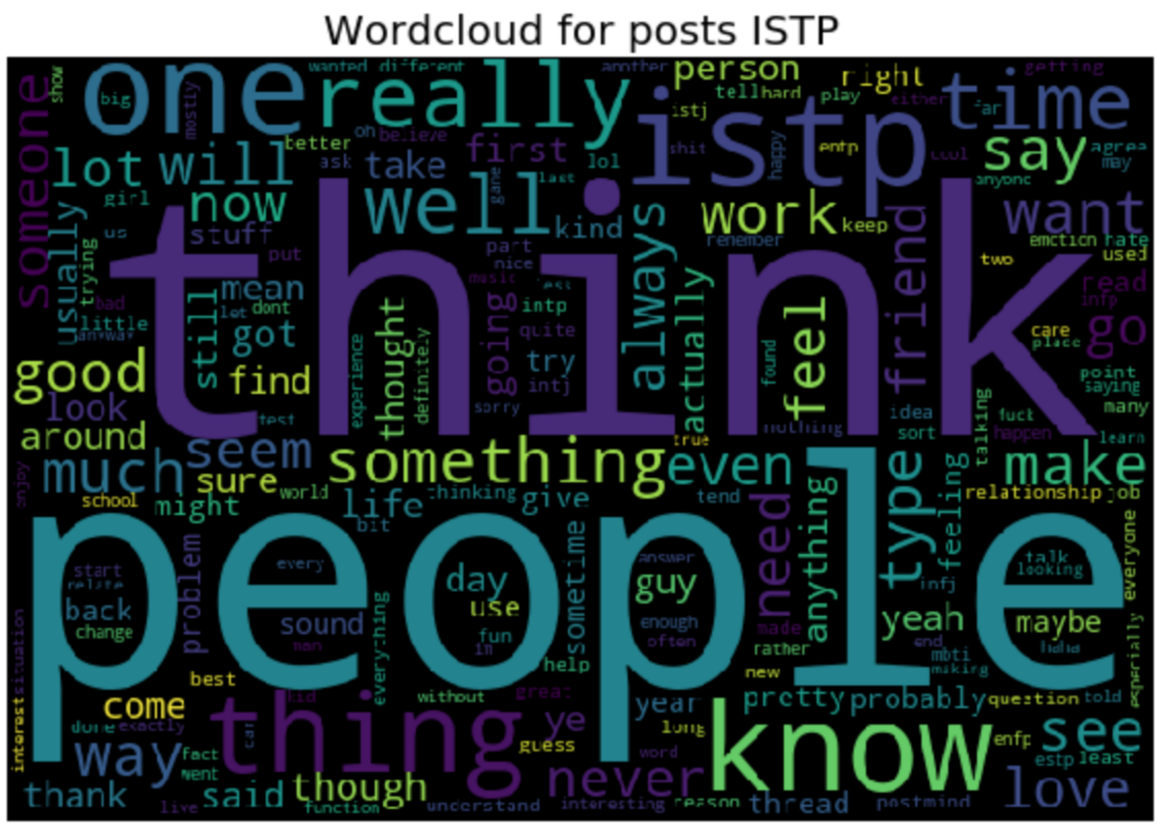

Frequency of Top 20 words in this ISFP
1600
1400
1200
1000
800
600
400
200
0
think
really
people
know
one
isfp
time
feel
love
good
much
things
well
type
say
want
see
always
something
even


Wordcloud for posts ISFP
people
think
know
thing
time
one
isfp
well
love
feel
say
type
really
though
want
friend
much
see
something
make
good
way
always
thank
even
infp
lot
will
someone
find

- ISTP

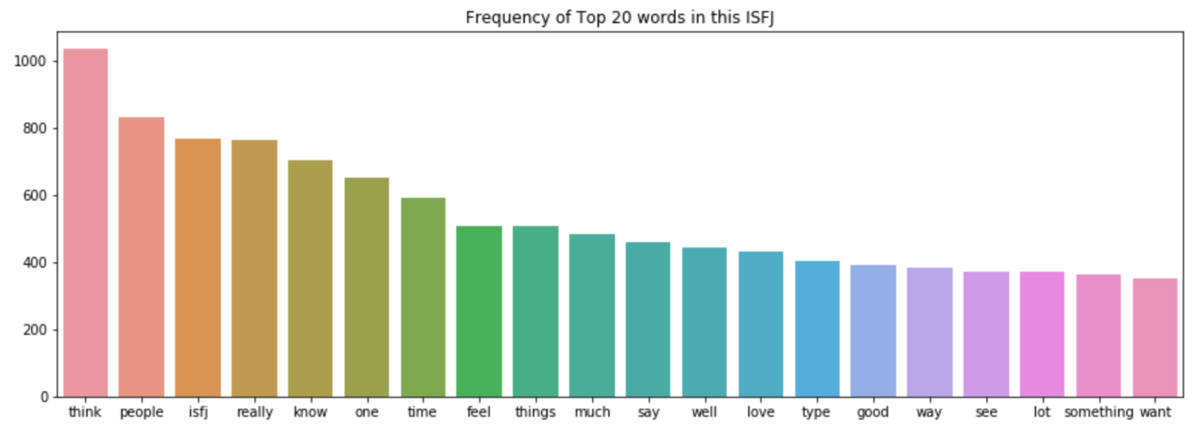


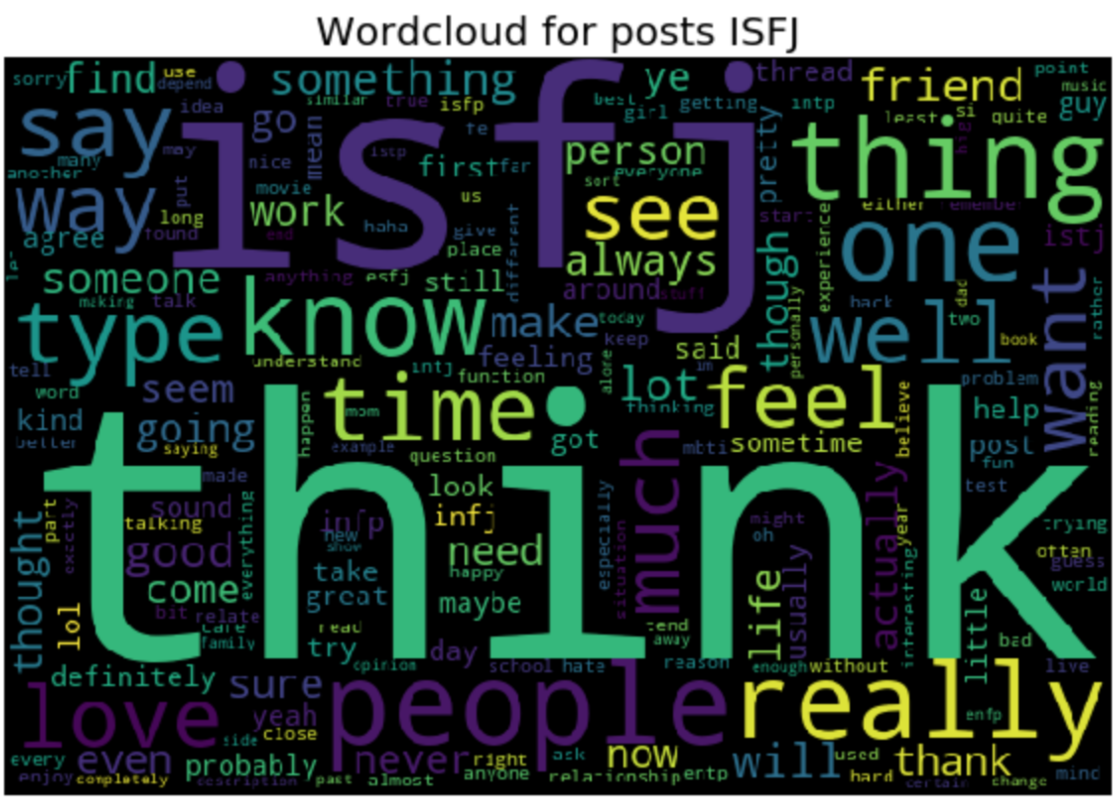
Frequency of Top 20 words in this ISTP
0
200
400
600
800
1000
1200
1400
1600
think
people
know
one
really
istp
time
well
something
much
good
want
things
way
say
now
see
type
will
even



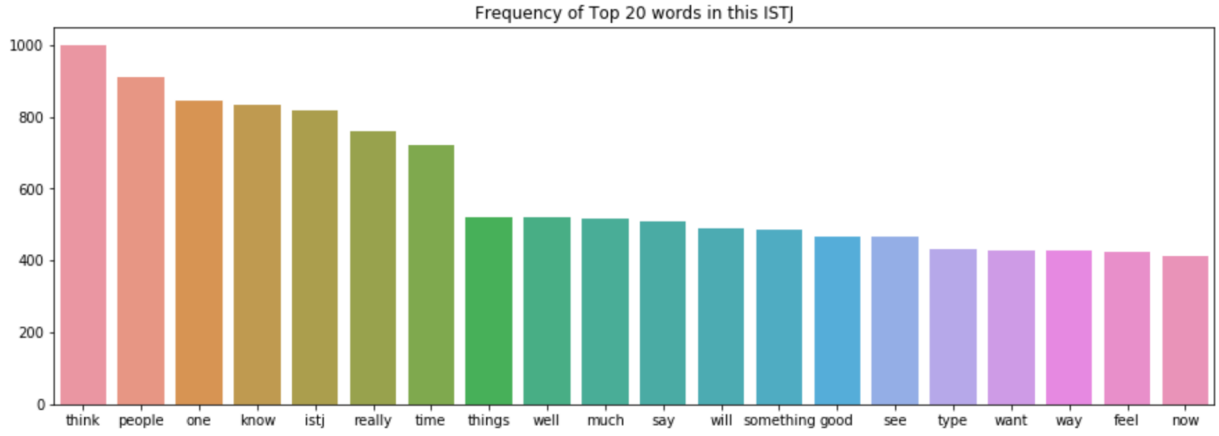
Wordcloud for posts ISTP
think
people
one
really
istp
time
know
thing
well
something
say
lot
will
now
work
want
good
much
seem
even
make
type
friend
feel
always
actually
someone
love
see
never
way
thank
said
though
come
yeah
maybe
guy
day
person
right

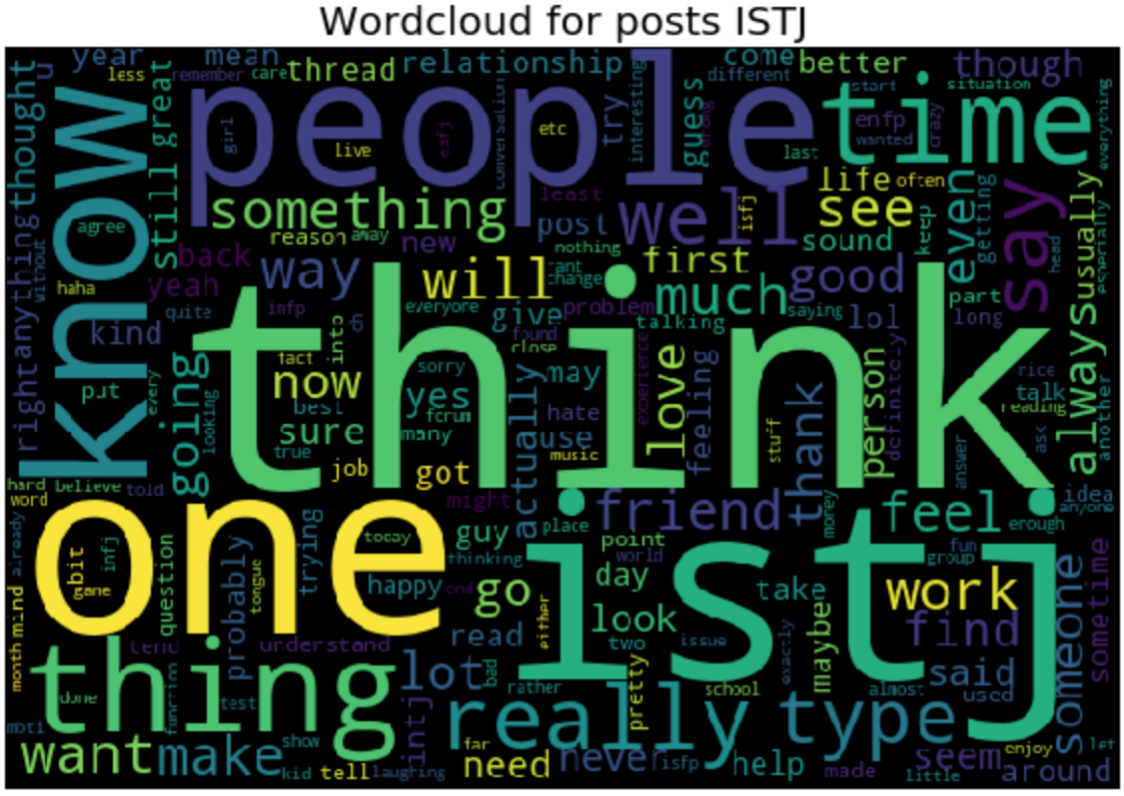
- ISFJ


Frequency of Top 20 words in this ISFJ
0
200
400
600
800
1000
think
people
isfj
really
know
one
time
feel
things
much
say
well
love
type
good
way
see
lot
something
want

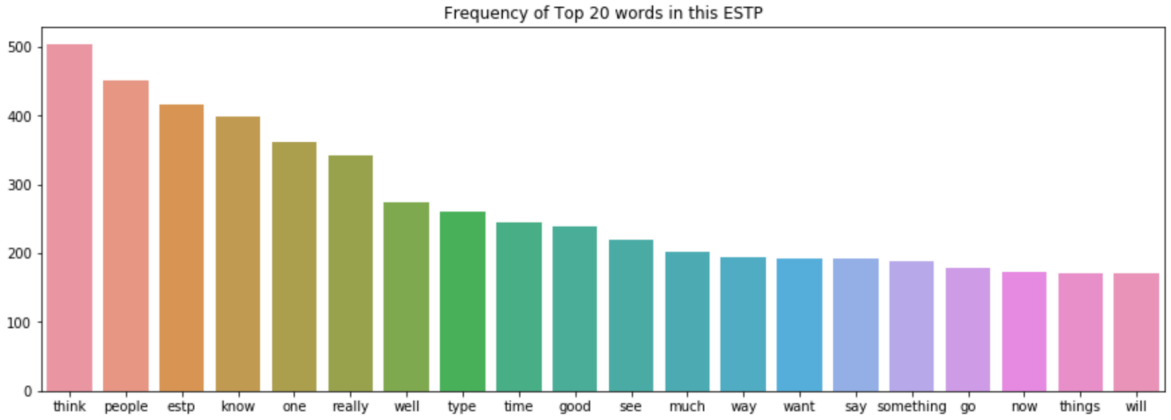

Wordcloud for posts ISFJ
think
isfj
people
really
know
type
time
feel
thing
one
well
want
say
way
see
love
much
something
friend
always
make
lot
find
person
someone
go
work
seem
going
good
kind
thought
sound
come
need
take
great
maybe
look
life
actually
usually
little
definitely
sure
even
never
probably
now
will
thank
yeah
lol
help
post
sometime
though
pretty
feeling
said
got
still
ye
thread
guy
agree
try
day

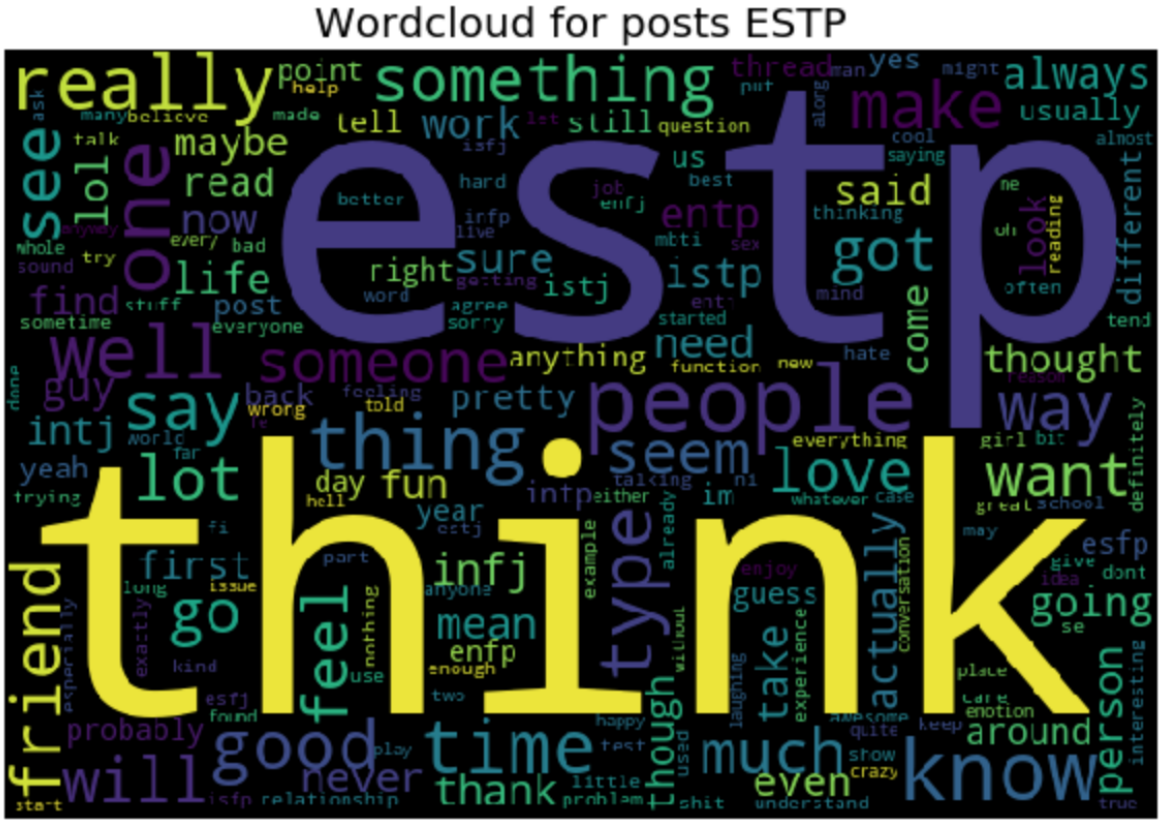
- ISTJ


Frequency of Top 20 words in this ISTJ
1000
800
600
400
200
0
think
people
one
know
istj
really
time
things
well
much
say
will
something
good
see
type
want
way
feel
now

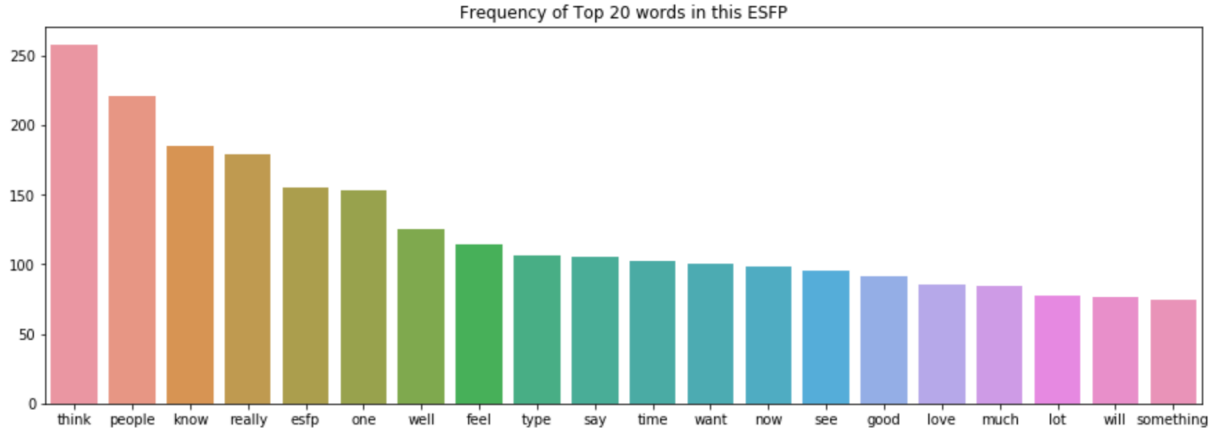

Wordcloud for posts ISTJ
think
people
one
know
istj
thing
really
time
well
something
will
much
say
way
good
make
want
need
friend
feel
work
find
look
go
love
thank
person
relationship
thread
mean
year
still
great
thought
anything
right
going
kind
back
yeah
haha
agree
reason
away
new
post
first
even
sound
life
see
always
usually
actually
probably
trying
question
lot
read
understand
guy
happy
day
take
maybe
said
someone
sometime
around
seem
help
never
tell
kid
show
laughing
pretty
school
almost
rather
bad
two
issue
either
exactly
enfp
isfj
intj
isfp
infp
infj
esfj
esfp
estj
estp
entj
entp
intp
mbti
omg
lol
etc
yes
now
sure
got
job
sorry
thank
talking
saying
everyone
problem
give
may
hate
cause
feeling
experience
definitely
music
stuff
point
place
world
thinking
today
might
believe
told
word
hard
already
month
mind
bit
game
tongue
fact
true
many
into
quite
part
long
talk
rice
idea
another
answer
ask
fun
group
enough
anyone
morning
different
start
situation
crazy
wanted
keep
getting
head
everything
especially
usually
interesting
guess
come
better
though
try
last
often
least
nothing
change
found
close
looking
every
put
best
forum
hope
hear
live
girl
remember
care
less
without
you
u
i
me
my
we
they
he
she
it
a
the
and
to
of
in
for
on
with
as
by
at
from
or
but
not
so
if
then
than
too
very
just
only
also
can
could
would
should
may
might
must
shall
will
do
does
did
have
has
had
be
am
is
are
was
were
been
being
get
got
make
made
go
went
gone
come
came
see
saw
seen
know
knew
known
think
thought
take
took
taken
give
gave
given
find
found
tell
told
ask
asked
work
worked
seem
seemed
feel
felt
try
tried
leave
left
call
called
need
needed
become
became
keep
kept
let
begin
began
begun
help
helped
show
showed
shown
hear
heard
play
played
run
ran
move
moved
live
lived
believe
believed
hold
held
bring
brought
happen
happened
write
wrote
written
provide
provided
sit
sat
stand
stood
lose
lost
pay
paid
meet
met
include
included
continue
continued
set
learn
learned
change
changed
lead
led
understand
understood
watch
watched
follow
followed
stop
stopped
create
created
speed
spoke
spoken
read
allow
allowed
add
added
spend
spent
grow
grew
grown
open
opened
walk
walked
win
won
offer
offered
remember
remembered
love
loved
consider
considered
appear
appeared
buy
bought
wait
waited
serve
served
die
died
send
sent
expect
expected
build
built
stay
stayed
fall
fell
fallen
cut
reach
reached
kill
killed
remain
remained
suggest
suggested
raise
raised
pass
passed
sell
sold
require
required
report
reported
decide
decided
pull
pulled

- ESTP


Frequency of Top 20 words in this ESTP
500
400
300
200
100
0
think
people
estp
know
one
really
well
type
time
good
see
much
way
want
say
something
go
now
things
will

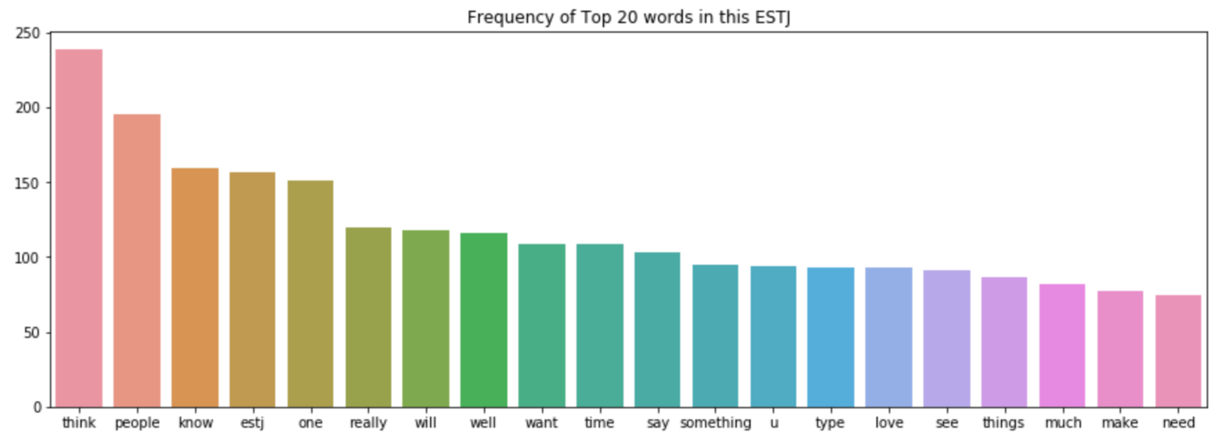

Wordcloud for posts ESTP
estp
think
really
something
people
make
always
usually
maybe
read
now
one
see
lol
life
find
well
someone
say
thing
lot
seem
love
want
way
thought
different
friend
feel
type
actually
good
time
much
know
person
will
go
first
infj
mean
entp
going
probably
never
thank
even
around
esfp
istp
entp
said
got
come
sure
right
istj
intj
yeah
fun
day
year
infp
intp
guess
take
though
probably
tell
work
still
question
point
thread
yes
might
talk
hard
better
thinking
look
reading
post
everyone
sometime
anything
need
pretty
back
wrong
told
feeling
girl
school
definitely
interesting
dont
give
idea
place
emotion
crazy
show
understand
relationship
problem
shit
true
start
issue
kind
enough
example
conversation
experience
laughing
enjoy
already
used
test
happy
little
two
part
anyone
nothing
use
exactly
esfj
found
quite
keep
awesome
laughing

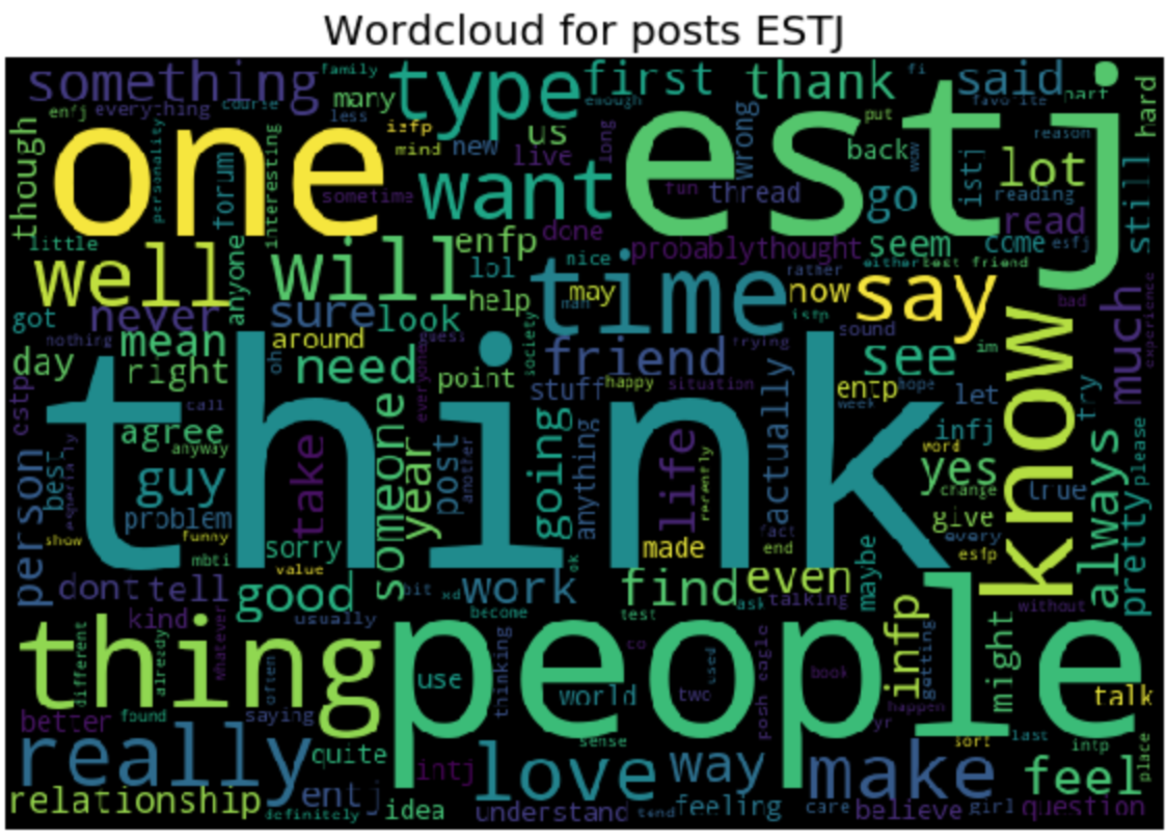
- ESFP


Frequency of Top 20 words in this ESFP
250
200
150
100
50
0
think
people
know
really
esfp
one
well
feel
type
say
time
want
now
see
good
love
much
lot
will
something

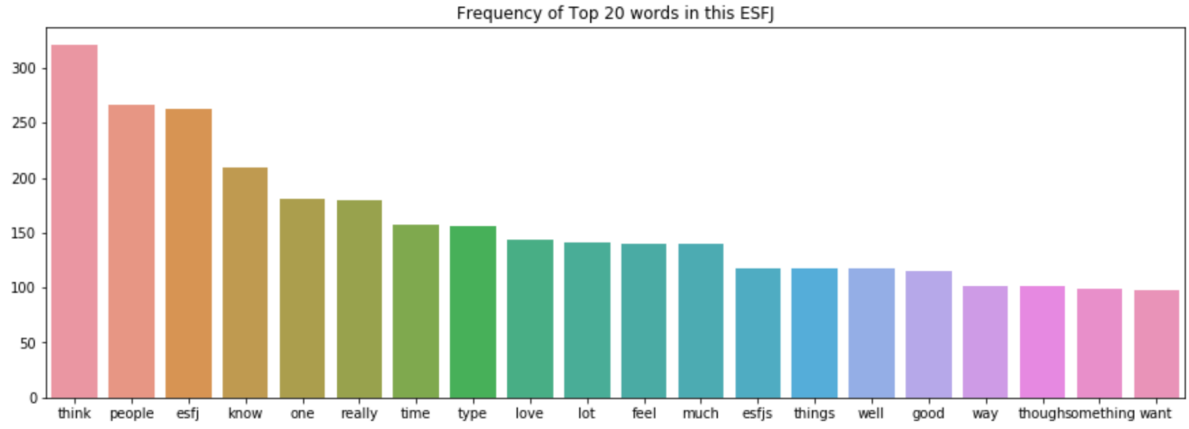

Wordcloud for posts ESFP
think
people
esfp
know
love
good
want
thought
kind
now
something
see
much
one
sound
post
need
enfp
lot
isfp
actually
day
got
mean
sure
thing
well
make
though
thread
person
type
anything
friend
life
estp
new
lol
take
probably
first
intj
always
really
will
going
feel
entj
never
said
year
yes
say
time
thank
entp
someone
around
find
maybe
right
guy
sorry
curious
test
reason
seem
relationship
infj
pretty
personality
using tapatalk
opinion
question
read
stuff
enough
different
idea
understand
probably
anyone
everyone
believe
problem
function
long
hate
stop
mind
let
back
fun
happy
point
another
start
talking
still
work
prefer
example
agree
laughing
bad
reason
looking
cute
wanted
crazy
getting
mbti
fact
social
away
made
told
true
guess
forum
others
consider
two
definitely
trying
might
hey
nice
bit
keep
hope
yeah
late
esfj
isfj
istp
estj
intp
infp
entj
isfp
enfp
intj
esfj
entp

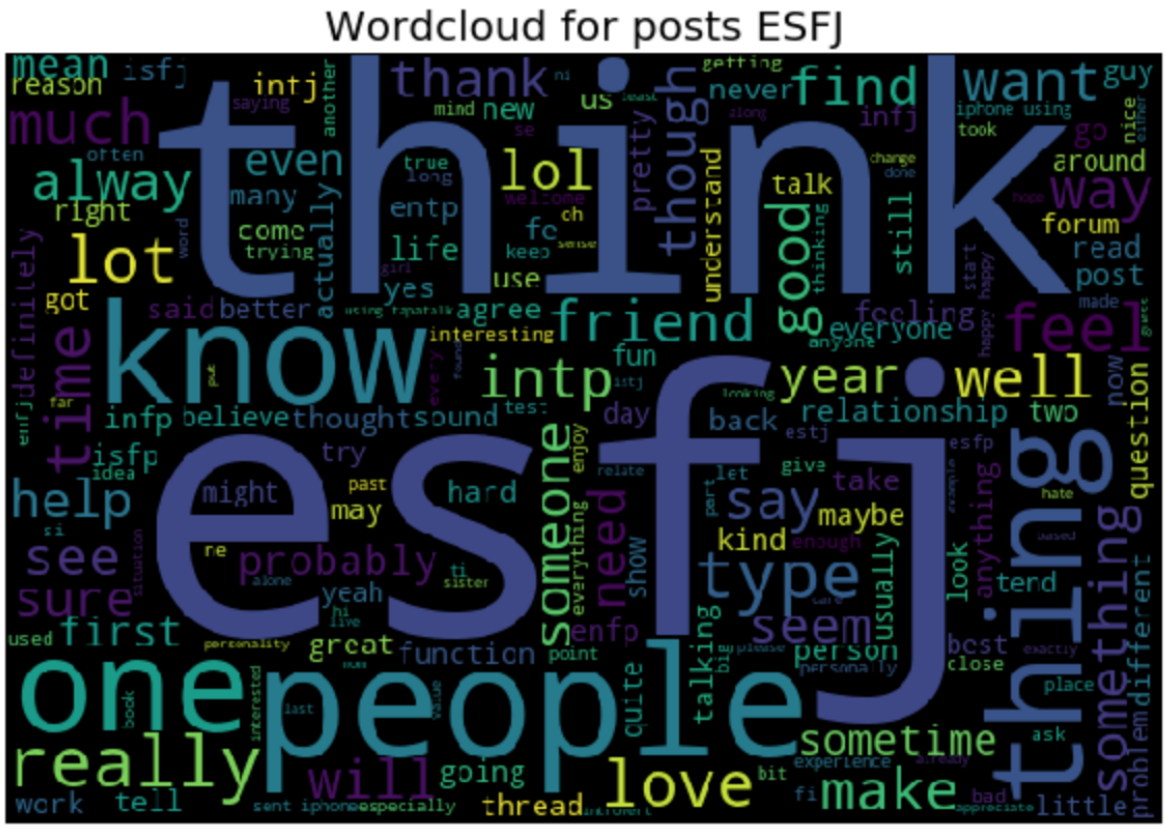
- ESTJ


Frequency of Top 20 words in this ESTJ
250
200
150
100
50
0
think
people
know
estj
one
really
will
well
want
time
say
something
u
type
love
see
things
much
make
need


Wordcloud for posts ESTJ
think
people
estj
one
know
thing
time
say
well
will
want
type
something
first
thank
said
really
make
love
find
work
good
feel
way
relationship
always
even
lot
friend
need
see
much
life
actually
someone
take
going
guy
mean
right
agree
person
sure
look
never
go
read
still
seem
come
probably
thought
yes
pretty
might
infp
infj
enfp
entp
entj
intj
istj
maybe
post
year
around
help
point
stuff
day
got
though
forum
anyone
interesting
wrong
thread
back
lol
done
nice
may
now
sorry
problem
dont
tell
kind
better
quite
idea
understand
feeling
believe
question
talk
made
give
let
stuff
us
live
new
mary
little
interesting

- ESFJ


Frequency of Top 20 words in this ESFJ
300
250
200
150
100
50
0
think
people
esfj
know
one
really
time
type
love
lot
feel
much
esfjs
things
well
good
way
thoughsomething
want


Wordcloud for posts ESFJ
esfj
think
people
know
thing
one
really
time
lot
much
always
help
sure
see
first
will
love
make
sometime
something
friend
feel
want
find
thank
lol
good
intp
year
well
relationship
probably
type
seem
need
someone
life
even
many
come
actually
though
understand
talk
still
forum
read
post
around
way
go
guy
never
mean
reason
isfj
intj
another
saying
mind
new
us
pretty
infj
iphone using
took
nice
often
true
long
welcome
entp
fc
keep
yes
use
agree
interesting
fun
everyone
feeling
made
happy
start
thinking
right
word
trying
girl
sense
ch
change
done
hope
got
said
better
believe
thought
sound
day
back
estj
two
question
now
isfp
try
past
might
may
hard
take
maybe
kind
enough
enjoy
relate
let
give
esfp
hate
anything
look
tend
best
close
place
ask
little
bad
appreciate
experience
bit
fi
thread
going
work
tell
sent iphone
especially
great
function
point
quite
talking
big
please
personality
exactly
different
problem
ni
se
ti
sister
alone
yeah
hi
live
re
situation
interested
last
value
book
put
far
enfj
definitely
idea
enfp
show
everything
usually
person
girl
part
example
saw
already
ago

Machine Learning Algorithms.

We are going to use **clean_posts** to predict personality type. First, we split data into train set and test set. The train set contains 75% of data points and the test set contains 25% of data points.

The X_train, X_test are sets of posts which are not the form that we can feed Machine Learning models. In order to apply Machine Learning models we need to have features expressed in numbers for X. We have to use **CountVectorizer to** convert the text documents to a vector of term/token counts.

We set min_df=0.1 and max_df=0.9. It means we only consider words that appear in more than 10% of the posts and less than 90% of the posts. It makes sense here, since we do not want to take into account words that only appear in a few posts neither words that appear in almost every post.

count_train and **count_test** are sparse matrices whose columns are selected words and rows are posts. Each entry in the matrix is the frequency that the word (column) appears in the post(row).
With min_df=0.1 and max_df=0.9, we are going to apply the models(Logistic Regression, K Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, Stochastic Gradient Descent and Naive Bayes) to the count_train and evaluate the accuracy of these models on count_test.

Models

Now we are ready to train a model and predict the required solution. Here,

y is value of target feature 'type_enc'

and

X is a sparse matrix whose columns are selected words and rows are posts. Each entry in the matrix is the frequency that the word (column) appears in the post(row) which we will use to fit a machine learning model and predict the value of y.

There are so many predictive modelling algorithms out there to choose from. We must understand the type of problem and solution requirement to narrow down to a select few models which we can evaluate. Our problem is a classification problem. We are also performing a category of machine learning which is called supervised learning as we are training our model with a labeled dataset. With these two criteria - Supervised Learning plus Classification, we can narrow down our choice of models to a few. These include:

- K_Nearest_Neighbor
- Logistic Regression
- Support Vector Machine
- Stochastic Gradient Descent
- Random Forest
- Decision Tree
- Naive Bayes

Metrics

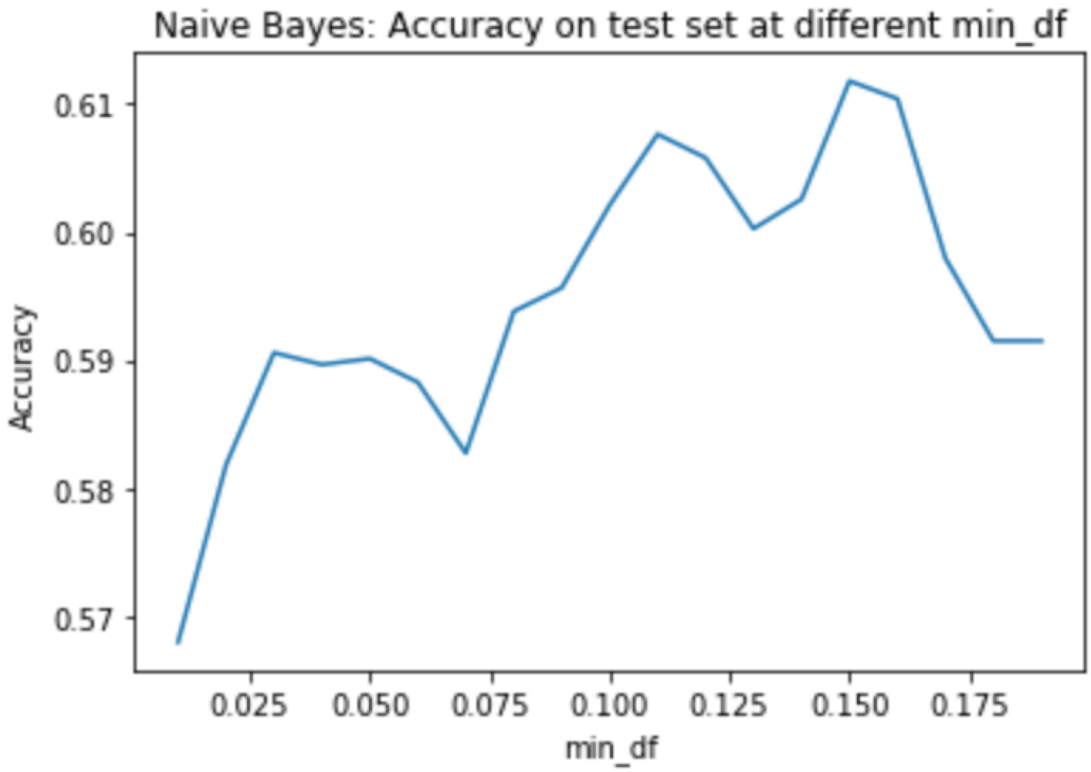
Beside the accuracy, we are going to use precision,recall, f1-score, AUC-ROC to evaluate our models.

	Models	Accuracy	Precision	Recall	F1-score	AUC-ROC
	Logistic Regression	0.54	0.54	0.54	0.54	0.91
	K-Nearest Neighbor	0.36	0.36	0.36	0.35	0.77
	Support Vector Machine	0.56	0.56	0.56	0.56	0.93
	Decision Tree	0.47	0.47	0.47	0.47	0.72
	Random Forest	0.59	0.60	0.59	0.54	0.91
	Stochastic Gradient Descent	0.53	0.56	0.53	0.53	0.80
	Naive Bayes	**0.60**	**0.62**	**0.60**	**0.61**	**0.92**

Based on the summary table, Naive Bayes is the best model.

More about Naive Bayes

Out of all models we applied, Naive Bayes turns out to be the best model. We would like to look deeper into Naive Bayes model. We vary the values of min_df and max_df and apply the best model we have(Naive Bayes).

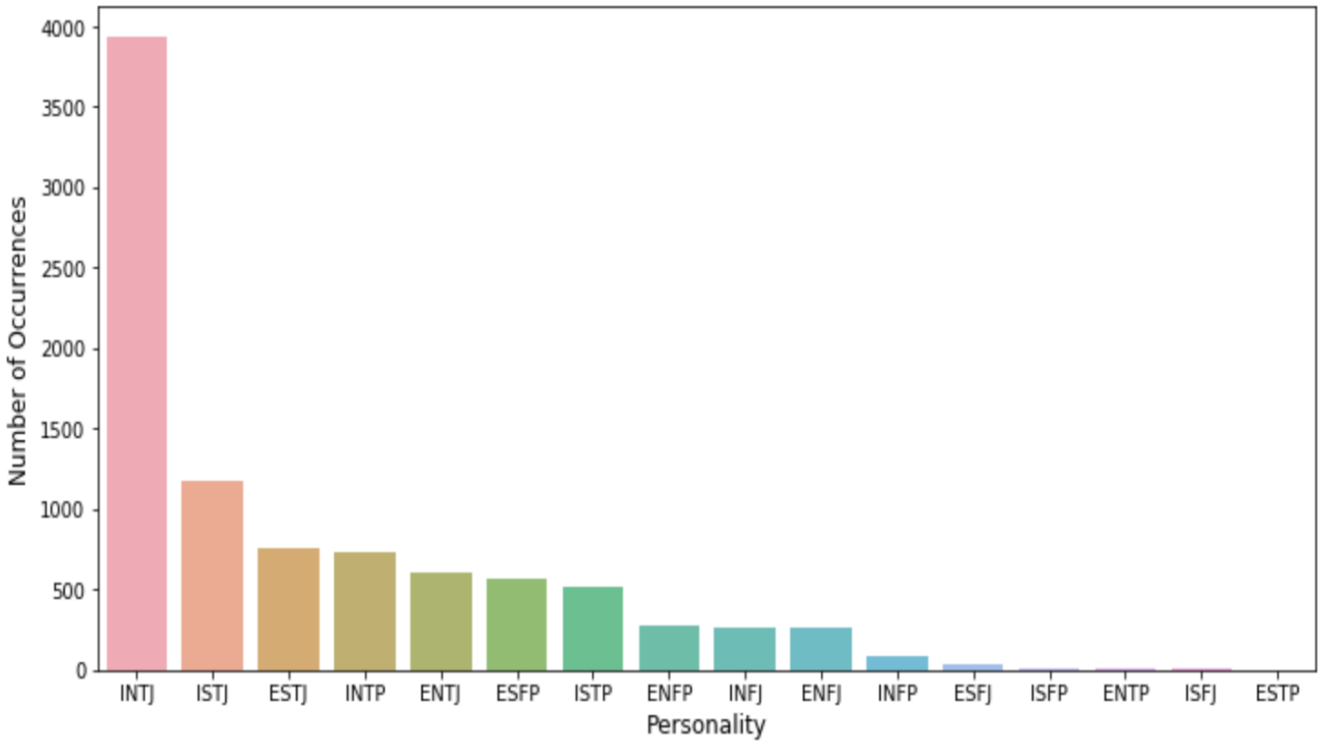


The highest accuracy is 0.612 when we only consider words that appear in more than 15% and less than 85% of the posts.

Predict Personality of Kaggle Users

We are going to apply our best model to find out Kaggle users personality based on their comments on Kaggle. The ForumMessages.csv is available on Kaggle at https://www.kaggle.com/kaggle/meta-kaggle?select=ForumMessages.csv

	PostUserId	Message	Message_length	clean_messages
1	368	Here are some papers that analyze Eurovision v...	32612	papers analyze eurovision voting patterns migh...
2	381	Hi Bdol. \r\n Please comment out the l...	28821	hi bdol please comment line dbclear error main...
3	387	From an economic perspective let's look at ...	526	economic perspective look demand supply declar...
6	393	[quote=Josette_BoozAllen;155818] \n\n ...	789	thomasseleck aws credits opportunity will anno...
7	412	Are there any criteria around model validit...	789	criteria around model validity nbsp e model ne...



We observe that the majority of kaggle users have INTJ personality type.

Predict Personality of Public Figures

There are two US political figures that seem to stand for the opposite point of views: President Donald Trump and previous president Barack Obama. We will collect their posts and use our best model to predict their personality.

	Name	Posts	posts_length	clean_posts
0	Barack Obama	Michelle and I have been spending a lot of tim...	766	michelle spending lot time together past month...
1	Donald Trump	No doubt many people told him his vision wasn...	1146	doubt many people told vision wasn t possible ...

Based on our best model, they are both predicted to have INFP(Introversion-Intuition-Feeling-Perception) personality type.

Conclusion

People use social media to share their feelings and opinions. It is true that social media posts can reveal many things about the owner including his personality. Social media posts become golden data if they are studied correctly. My project focuses on using machine learning algorithms in supervised machine learning to predict the personality of a person from the type of posts they put on social media.

Psychologists can use this project to study personality. This project can also help companies who wish to learn more about their customers from their social media posts to provide appropriate services.