Alon Itach: 301790515

Franziska Wehrmann: 777934738

Text Mining

FRIEDRICH NIETZSCHE - BEYOND GOOD AND EVIL

Frequency of tokens

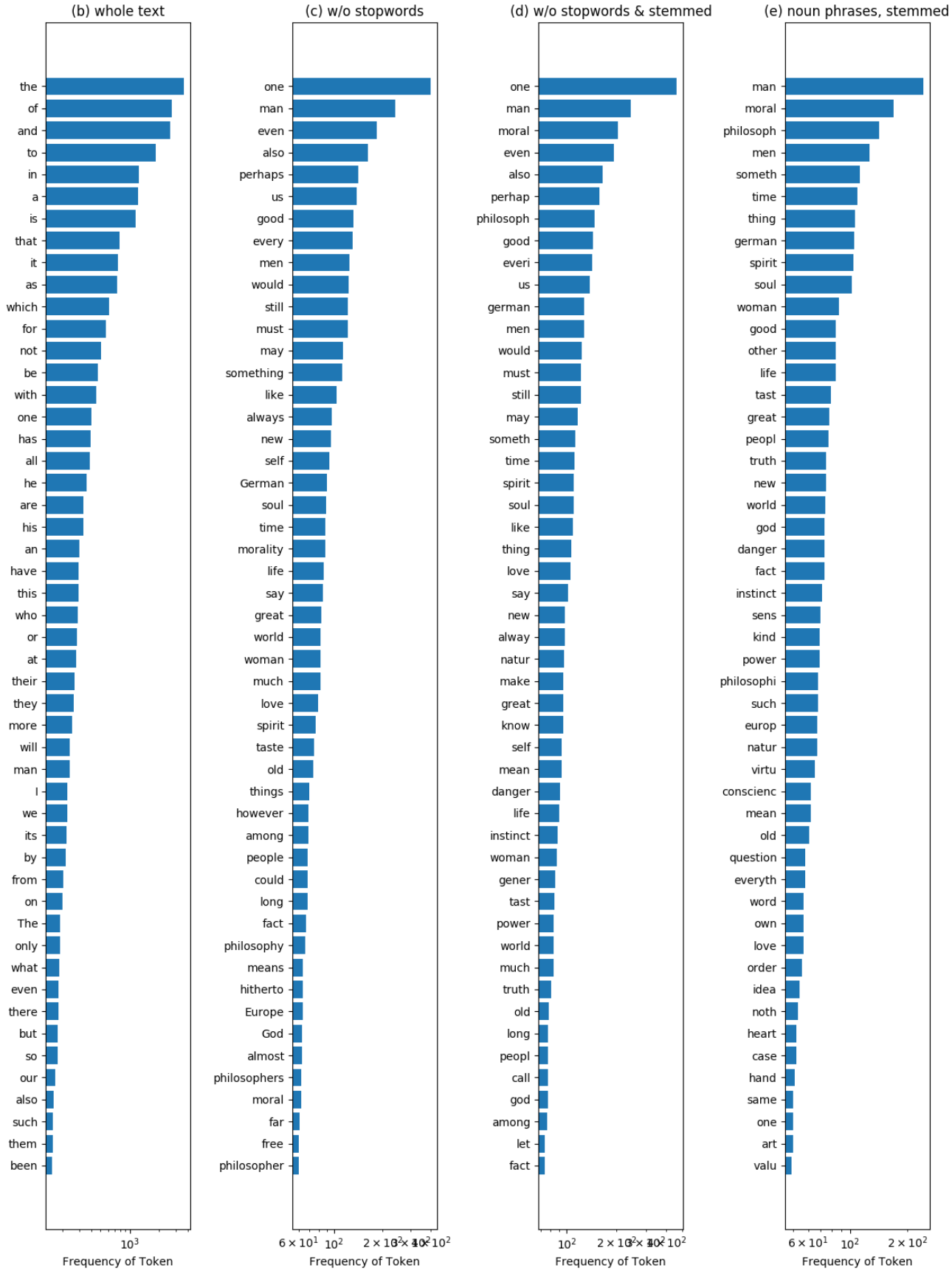
This plot shows the 50 most frequent tokens.

In **(b)** we tokenized the whole text, in **(c)** we removed stopwords from the text and tokenized again. The plot shows that among the 50 most frequent tokens in the book, only the words (one, man, even, also, perhaps) are not stopwords. In **(d)** we went further and also stemmed the words of the text. It is interesting to see how stems get more frequent, that describe the character of the book, for example

'moral' (3) ← {moral (21) , morality (48), +?}, 'philosoph' (7) ← {philosophy (42) , philosophers (49), +?}

The number after the token indicates the rank of the token. NOTE: From the plot we can only reconstruct two words each, that stem their token. It is very likely that also other words stem to the same token, for example philosophizing → philosoph.

Plot **(e)** shows the most frequent tokens of all the noun phrases in the book. From this analysis one can again conclude that the book had a philosophical character, since tokens like (moral, phiosoph, spirit, soul, instinct, sens, truth, virtu) would not be that frequent in another kind of book.


(b) whole text
the
of
and
to
in
a
is
that
it
as
which
for
not
be
with
one
has
all
he
are
his
an
have
this
who
or
at
their
they
more
will
man
I
we
its
by
from
on
The
only
what
even
there
but
so
our
also
such
them
been
10^3
Frequency of Token
(c) w/o stopwords
one
man
even
also
perhaps
us
good
every
men
would
still
must
may
something
like
always
new
self
German
soul
time
morality
life
say
great
world
woman
much
love
spirit
taste
old
things
however
among
people
could
long
fact
philosophy
means
hitherto
Europe
God
almost
philosophers
moral
far
free
philosopher
6×10^1
10^2
2×10^2
Frequency of Token
(d) w/o stopwords & stemmed
one
man
moral
even
also
perhap
philosoph
good
everi
us
german
men
would
must
still
may
someth
time
spirit
soul
like
thing
love
say
new
alway
natur
make
great
know
self
mean
danger
life
instinct
woman
gener
tast
power
world
much
truth
old
long
peopl
call
god
among
let
fact
10^2
2×10^2
Frequency of Token
(e) noun phrases, stemmed
man
moral
philosoph
men
someth
time
thing
german
spirit
soul
woman
good
other
life
tast
great
peopl
truth
new
world
god
danger
fact
instinct
sens
kind
power
philosophi
such
europ
natur
virtu
conscienc
mean
old
question
everyth
word
own
love
order
idea
noth
heart
case
hand
same
one
art
valu
6×10^1
10^2
2×10^2
Frequency of Token

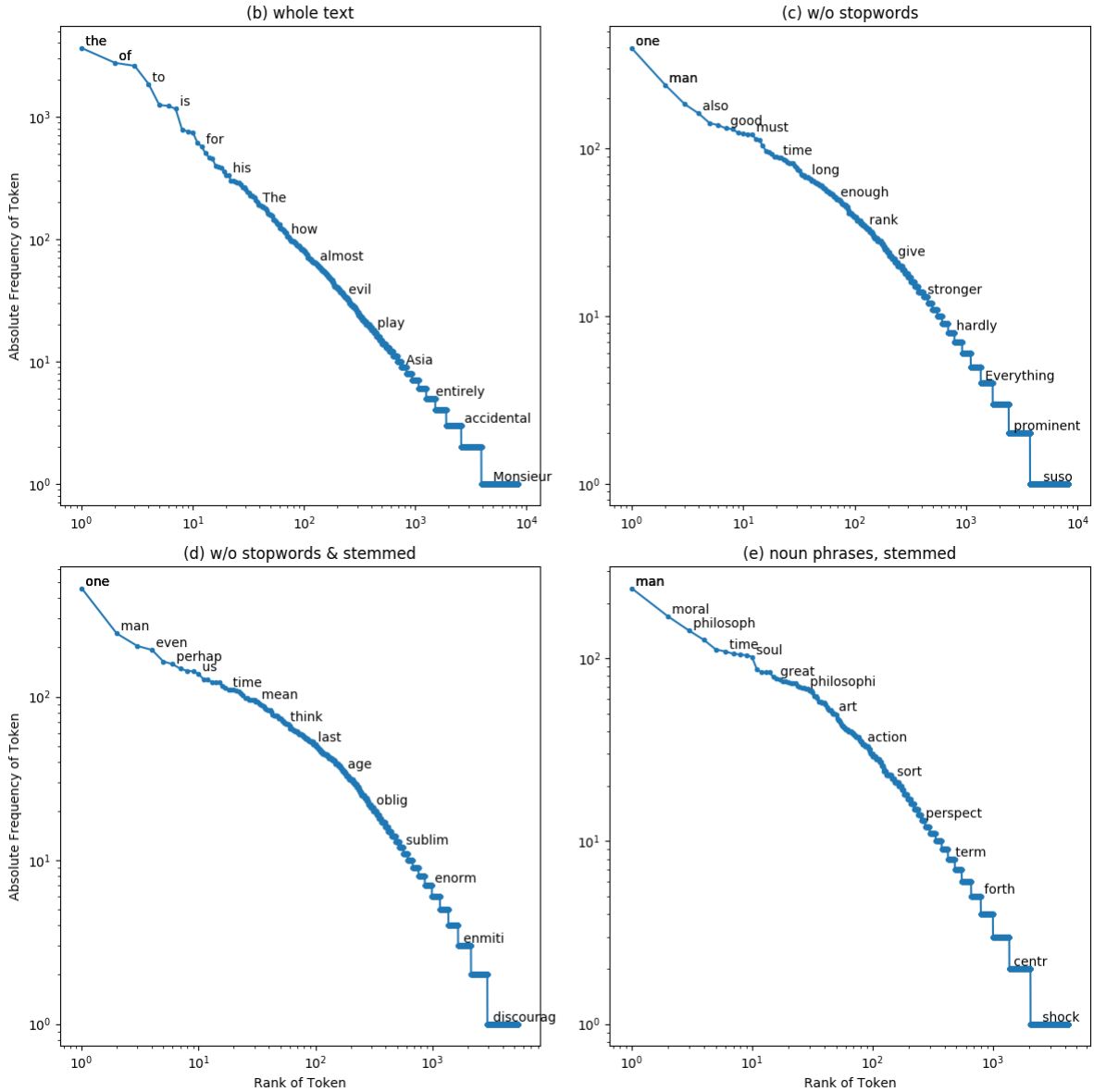
Zipf's plot

The word frequency of a document can be modeled by Zipf's Law, which states:

The *r*th most frequent word has a frequency *f*(*r*) that scales according to $f(r) \propto \frac{1}{r^\alpha}$. Where $\alpha \approx 1$ [source] This relation plotted in a log-log-graph results in a diagonal. Hence the distribution of words in a document follows Zipf's law, when its plot of log-frequency-of-token over log-rank-of-token is a diagonal line as well.

Indeed we see that Nietzsche's vocabulary follows Zipf's Law **(b)**.

When stopwords are removed from the text **(c)** and **(d)**, the diagonal changes in a slightly concave shape. Less frequent words are still used more frequently than Zipf's Law would suggest. As an extreme case imagine that every word would be used with the same frequency, then the plot would be a horizontal line at this frequency and not descent for higher ranks. So when in a text also higher ranked tokens are used "more frequently than expected", the diagonal is lifted up in the direction of the extreme case and we get a concave shape. Only analyzing the noun phrases in **(e)** includes the deletion of stopwords, since they are mainly words that can not be found in a noun phrase.



POS Tagging

In the book some words are written in all-capital letters to emphasize them. the `nltk.pos_tag` function always interprets most of them to be some kind of noun (mainly `NNP`), even though they are not.

After lowercasing such words, the function worked correctly.

```
sentence = 'The fundamental belief of metaphysicians is THE BELIEF IN ANTITHESES OF
VALUES.'

print(nltk.pos_tag(nltk.word_tokenize(sentence))
[('The', 'DT'), ('fundamental', 'JJ'), ('belief', 'NN'), ('of', 'IN'),
('metaphysicians', 'NNS'), ('is', 'VBZ'), ('THE', 'DT'), ('BELIEF', 'NNP'), ('IN',
'NNP'), ('ANTITHESES', 'NNP'), ('OF', 'NNP'), ('VALUES', 'NNP'), ('.', '.')]

# lowercase the words that were written in all capitals
sentence_lc = to_lowercase_words(sentence)

print(nltk.pos_tag(nltk.word_tokenize(sentence_lc)))
[('The', 'DT'), ('fundamental', 'JJ'), ('belief', 'NN'), ('of', 'IN'),
('metaphysicians', 'NNS'), ('is', 'VBZ'), ('the', 'DT'), ('belief', 'NN'), ('in',
'IN'), ('antitheses', 'NNS'), ('of', 'IN'), ('values', 'NNS'), ('.', '.')]
```

In the next sentence we evaluate the tag to 'such'. It got tagged as `JJ` (adjective, for example 'He is such a liar.'), which is also true in some cases, but here 'such' is actually used as a determiner like in 'Such men are dangerous.' so we would expect the pos tag `DT`.

```
sentence = 'and with such hypnotic rigidity to see Nature FALSELY, that is to say,
Stoically'

print(nltk.pos_tag(nltk.word_tokenize(sentence_lc))
[('and', 'CC'), ('with', 'IN'), ('such', 'JJ'), ('hypnotic', 'JJ'), ('rigidity', 'NN'),
('to', 'TO'), ('see', 'VB'), ('Nature', 'NNP'), ('falsely', 'RB'), (',', ','), ('that',
'DT'), ('is', 'VBZ'), ('to', 'TO'), ('say', 'VB'), (',', ','), ('Stoically', 'RB')]
```