Machine Learning

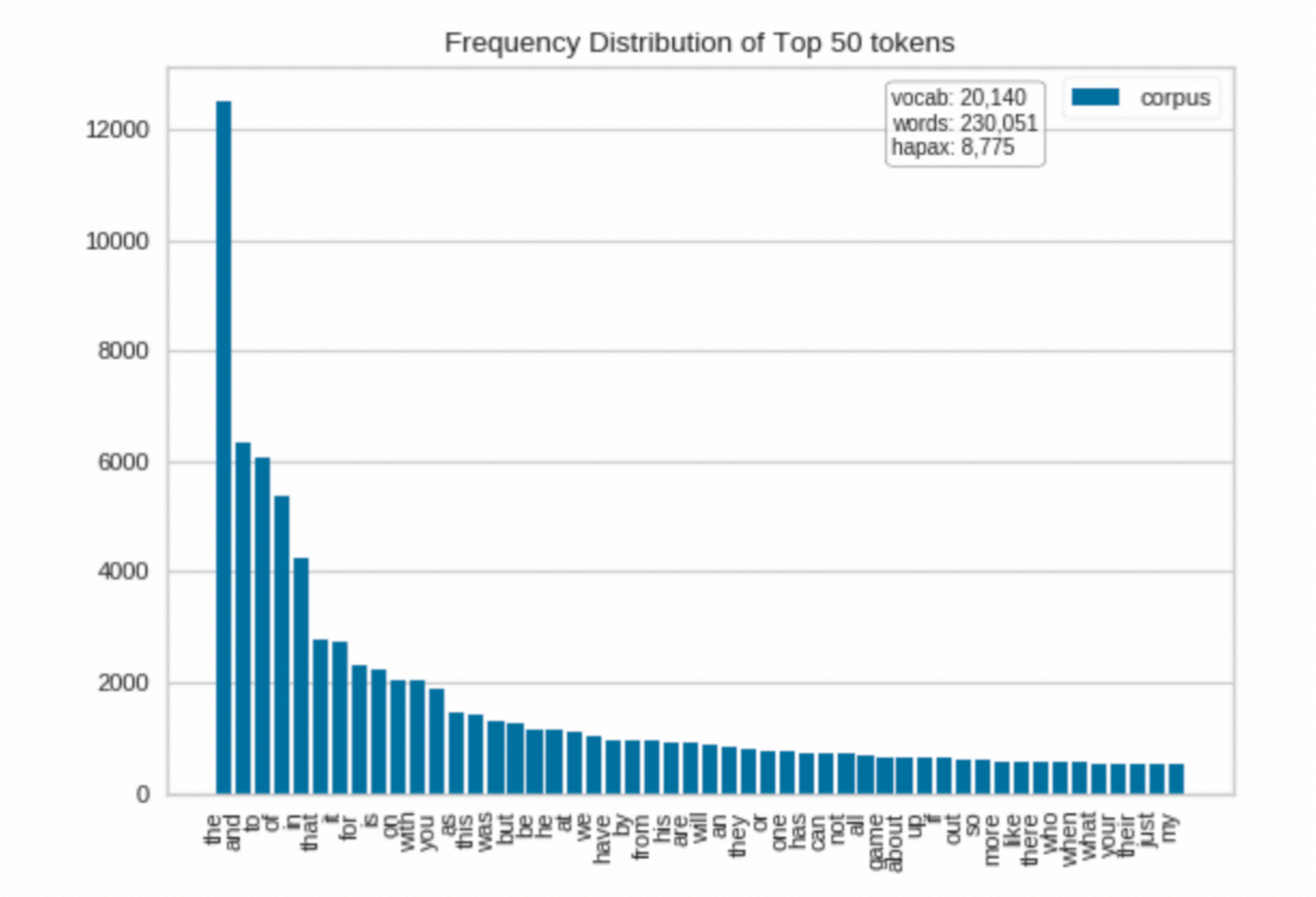
- At a high level, machine learning is set of techniques, or *machinery*, that is designed to automatically learn complex relationships between random variables. Typically there is some output that we are interested in predicting, and there are inputs that we hypothesize affect that output. The goal is to build a model that accurately predicts the output given the input.

- In the context of natural language there are several things that make it unique:
  - The **distribution of words is highly non uniform** in the wild and follows a power law with a very long tail [1].
  - Unlike images or tabular data, **language is highly compositional** and meaning is derived from how characters, words, and phrases are composed. There are countless ways of representing the same meaning using language; this makes it hard to automate the extraction of meaning from it.

[1] Zipf, 1949

Word frequency distribution

- Zipf’s Law: The frequency of the kth most frequent word is roughly proportional to 1/k


Frequency Distribution of Top 50 tokens
vocab: 20,140
words: 230,051
hapax: 8,775
corpus
the
and
to
of
in
that
it
for
is
on
with
you
as
this
was
but
be
he
at
we
have
by
from
his
are
will
an
they
or
one
has
can
not
all
game
about
up
if
out
so
more
like
there
who
when
what
your
their
just
my