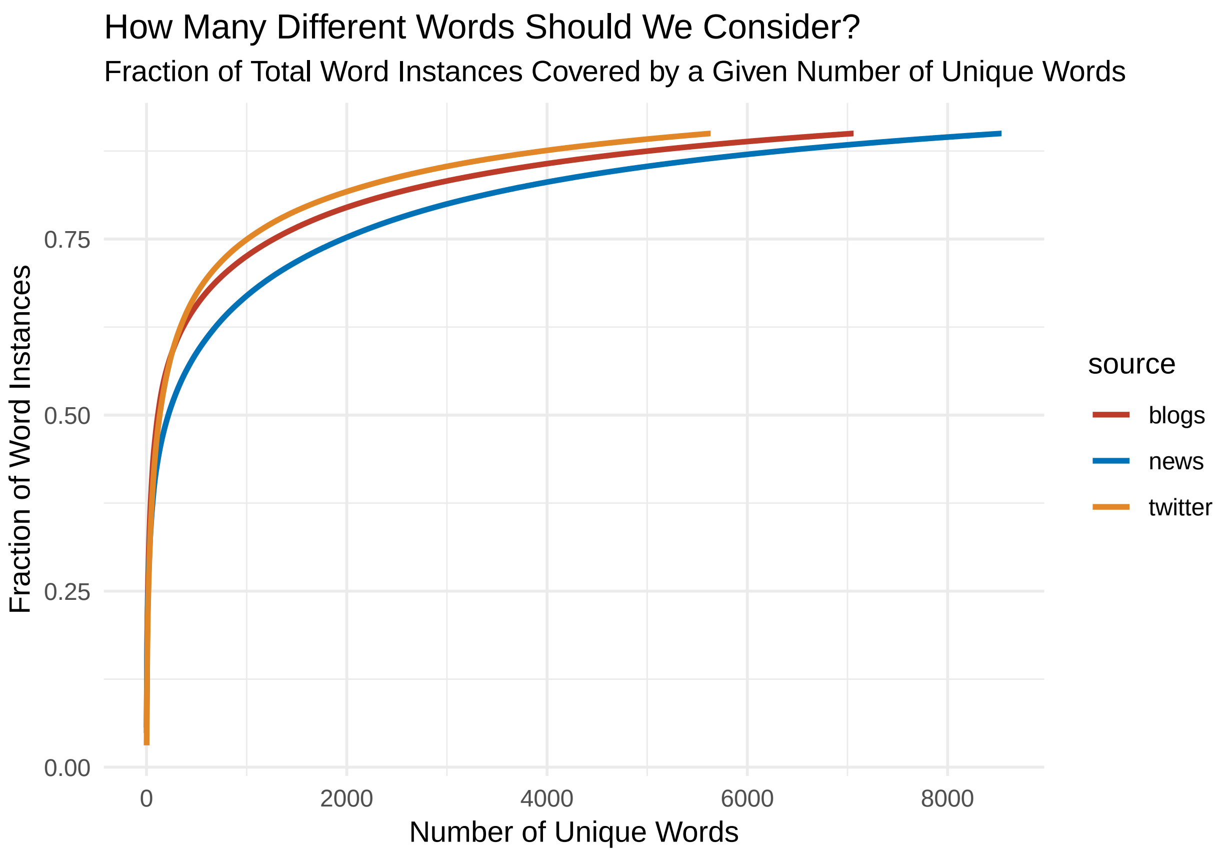
Exploratory Analysis

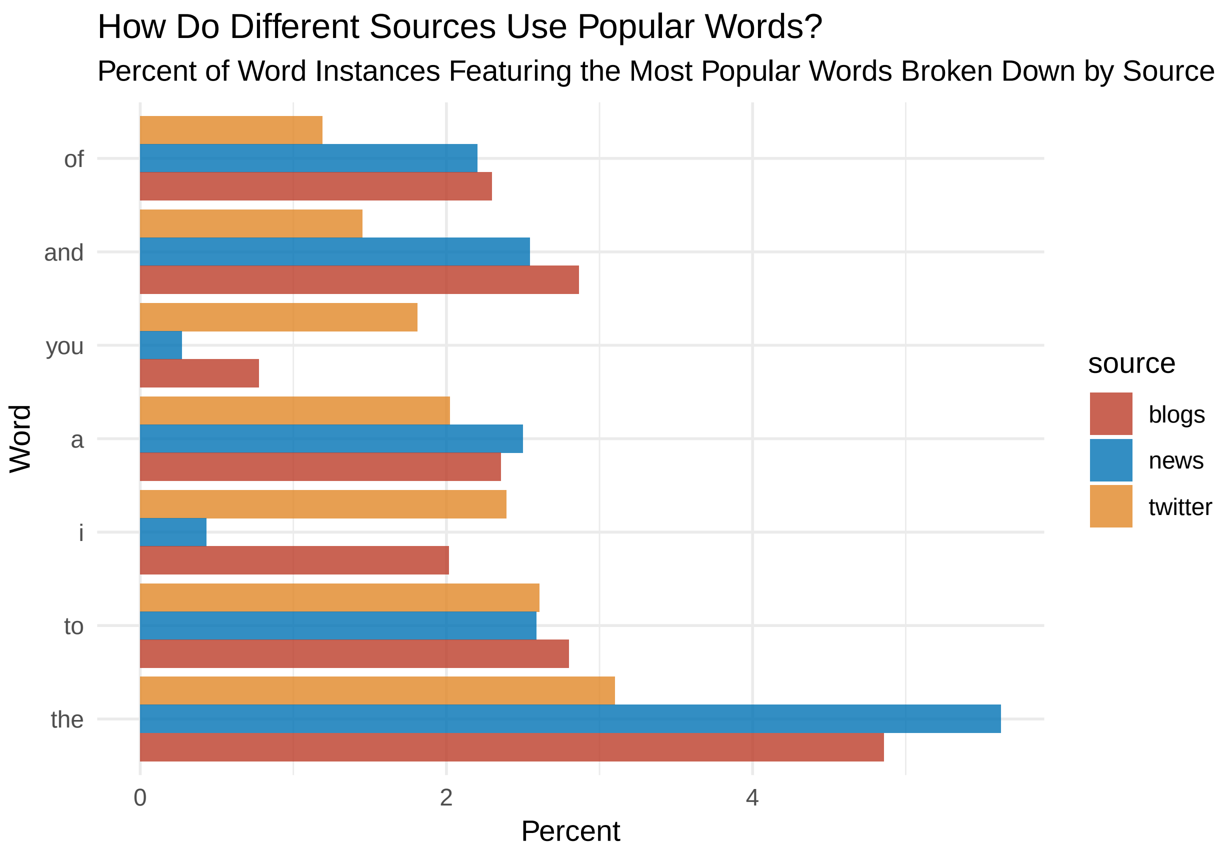
Matthew Halbasch

April 12, 2019

Next we want to break up each of these documents into two data frames - one which counts the words and their frequency in the sample, and another that instead counts pairs.

Source	Documents	Word Instances	Pairs	50% Coverage	90% Coverage
Twitter	2,360,148	30,218,125	27,858,438	132	5,634
News	77,259	2,693,898	2,616,865	218	8,539
Blogs	899,288	38,154,238	37,265,138	114	7,061




How Do Different Sources Use Popular Words?
Percent of Word Instances Featuring the Most Popular Words Broken Down by Source
Word
of
and
you
a
i
to
the
Percent
0
2
4
source
blogs
news
twitter