fraction of text common with 200 most frequent words in corpus