Word frequency data

For samples of the four different datasets, see https://www.wordframequency.info/samples.asp.

Note: "lemmas" on this page means that all of the different word forms are grouped together. For example, the freque
decides, decided, deciding} are all grouped together under the one entry {decide}. Word forms refer to each of the dis
{*decide, decides, decided, deciding*}. The "lemmatized" entries always distinguish by part of speech, however, so that
(*the deciding factor*) and *deciding* as a verb (*he really had a hard time deciding what to do*) will always be distinguished from each other and calculated separately.

1. lemmas_60k.txt: top 60,000 lemmas, Explanation of columns:

rank	1-62,000. Based on word frequency.
lemma	Again, the "dictionary / headword" entry. This is why *was* or *happier* or *shoes* would not be included here; they are word forms of the lemmas *be*, *happy*, and *shoe*.
PoS	Part of speech. This is the first letter of the codes from https://ucrel.lancs.ac.uk/claws7tags.html
freq	Total frequency
perMil	The "normalized" frequency -- per million words in COCA (992,960,152 words total)
%caps	The percent of all tokens that are capitalized, e.g. February, German, Trump, Adobe, Hummer, Marshall
Especially for nouns, the [%caps] column can be very useful to find entries that might actually be used as proper nouns the majority of the time, even though the CLAWS 7 tagger did not tag them as proper nouns (e.g. Trump, Springer, Savannah, Newt, etc. Most proper nouns (e.g. Minnesota, Alice) have been removed from the frequency list. But there are many "intermediate" cases like those just listed, and it is impossible to have one single rule for what percent of the time it needs to be used as a proper noun, in order to be removed from the list. But with the [%caps] column, you can find (and delete, if desired) some of the "marginal" words.	
%allC	The percent of all tokens that are completely capitalized, e.g. USB, DNA, CEO
Related to #1 are words that are completely capitalized, and are often acronyms, e.g. BC (before Christ), IT (Information technology), DNA, USB, ROI (Return on Investment). Which acronyms should we include, and which ones (e.g. state or local agencies, or highly technical scientific terms) should we omit? There is no clear answer on this, but the [%allC] (all caps) column can at least provide data on this.	
range	The number of the 485,179 texts in which the lemma occurs at least one time
disp	The Juilland "d" dispersion measure (0.00 to 1.00) shows how "evenly" a word is spread across the corpus.
For example, if the word occurs in 1000 texts, but only 1 or 2 times in 987 of those 100 texts (and many times in the other 13), the "range" figure simple shows "1000". The dispersion measure can see that even though all 1000 texts contain the word, it is not evenly spread across these texts. A word like "the" or "with" will have a dispersion value very close to 1.00, and a highly specialized word (which occurs in such a few texts) will be closer to 0.00.	
{blog, web, TVM, spok, fic, mag, news, acad}	The raw frequency in each of these eight genres.
{blogPM, webPM, . . .newsPM, acadPM}	The normalized frequency (per million words: PM) in each of these eight genres.

2. lemmas_60k_subgenres.txt: top 60,000 lemmas + sub-categories. This table is essentially a continuation of #1 above. But because there are so many columns (nearly 200 columns), we've created a separate file for this, for those who don't need this much detail. Explanation of columns:

rank, lemma, PoS	Same as for #1 above.
x101-104	The raw frequency in each of the 96 sub-categories of the corpus
p101-214	The normalized frequency (per million words) in each of the 96 sub-categories

3. lemmas_60k_words.txt: top 60,000 lemmas + words (more than 100,000 forms). Shows the frequency of each word form for each of the top 60,000 lemmas, where the word form occurs at least five times total. Explanation of columns:

lemRank, lemFreq	Same as [rank] and [freq] in #1 above
lemma, PoS	Same as the columns in #1 above

wordFreq	The frequency of the individual word forms, e.g. {decide, decides, decided, deciding}, {big, bigger, biggest}, or {shoe, shoes}. The word form must have a frequency of at least 5. For some lower frequency words, it is possible that not all of the word forms will be listed (but this can often be compensated for by using the data from #3 below)
word	The frequency of the individual word forms, e.g. {decide, decides, decided, deciding}, {big, bigger, biggest}, or {shoe, shoes}. The word form must have a frequency of at least 5. For some lower frequency words, it is possible that not all of the word forms will be listed (but this can often be compensated for by using the data from #3 below)

4. words_219k.txt: top ~220,000 word forms. This includes all word forms that occur at least 20 times in the corpus, in at least five different texts (so a strange name that occurs in just 1 or 2 of the 500,000 texts wouldn't be included).

rank	Same as [rank] in #1 above
word	Same as the columns in #1 above
freq	Same as [freq] in #1 above (but obviously for the individual word form, rather than lemma)
#texts	The number of the 485,179 texts in which the lemma occurs at least one time
%caps	Same as in #1 above.
As is mentioned there, this can often be be used to distinguish between proper nouns (capitalized) and common nouns (not capitalized), and this is particularly useful in a list like this, where there is no indication of part of speech.	
{blog, web, TVM, spok, fic, mag, news, acad}	Same as the columns in #1 above
{blogPM, webPM, . . .newsPM, acadPM}	Same as the columns in #1 above