

 [first20hours](#) / [google-10000-english](#) Public

This repo contains a list of the 10,000 most common English words in order of frequency, as determined by n-gram frequency analysis of the Google's Trillion Word Corpus.

 [View license](#)


 **3.8k** stars  **1.9k** forks  Branches  Tags  Activity

 Star

 Notifications

 **Code**  **Issues** 15  **Pull requests** 7  **Actions**  **Projects**  **Wiki**  **Security**  **Insights**

 **master** ▾

 **1 Branch**

 **0 Tags**














 


 Go to file


Go to file


Code

...

 worldlywisdom	Update README.md	3 years ago		
 20k.txt	Replace the last half of 20k.txt us...	8 years ago		
 LICENSE.md	Update LICENSE.md	3 years ago		
 README.md	Update README.md	3 years ago		
 google-10000-english-no...	Remove more NSFW words from...	5 years ago		
 google-10000-english-us...	Remove more NSFW words from...	5 years ago		
 google-10000-english-us...	Remove more NSFW words from...	5 years ago		
 google-10000-english-us...	Remove more swear words from ...	7 years ago		
 google-10000-english-us...	Remove more NSFW words from...	5 years ago		
 google-10000-english-us...	add alternative list with America...	10 years ago		
 google-10000-english.txt	Remove trailing \t characters	11 years ago		

 **README**

 License



Maintenance Level

Not Maintained

About This Repo

This repo contains a list of the 10,000 most common English words in order of frequency, as determined by [n-gram frequency analysis](#) of the [Google's Trillion Word Corpus](#).

According to the [Google Machine Translation Team](#):

Here at Google Research we have been using word n-gram models for a variety of R&D projects, such as statistical machine translation, speech recognition, spelling correction, entity detection, information extraction, and others. While such models have usually been estimated from training corpora containing at most a few billion words, we have been harnessing the vast power of Google's datacenters and distributed processing infrastructure to process larger and larger training corpora. We found that there's no data like more data, and scaled up the size of our data by one order of magnitude, and then another, and then one more - resulting in a training corpus of one trillion words from public Web pages.

We believe that the entire research community can benefit from access to such massive amounts of data. It will advance the state of the art, it will focus research in the promising direction of large-scale, data-driven approaches, and it will allow all research groups, no matter how large or small their computing resources, to play together. That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

This repo is derived from [Peter Norvig's](#) compilation of the [1/3 million most frequent English words](#). I limited this file to the 10,000 most common words, then removed the appended frequency counts by running this sed command in my text editor:

```
sed 's/[0-9]*//g'
```



Special thanks to [koseki](#) for [de-duplicating the list](#).

Swear-free lists

There are two additional lists which are identical to the original 10,000 word list, but with swear words removed. Swear words were removed based on these lists:

- [reimertz/curse-words](#)
- [MauriceButler/badwords](#)
- [LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words](#)

Word length lists

Three of the lists (all based on the US english list) are based on word length:

- **Short:** 1-4 characters
- **Medium:** 5-8 characters
- **Long:** 9+ characters

Each list retains the original list sorting (by frequency, decending).

Usage

This repo is useful as a corpus for typing training programs. According to analysis of the [Oxford English Corpus](#), the 7,000 most common English lemmas account for approximately 90% of usage, so a 10,000 word training corpus is more than sufficient for practical training applications.

To use this list as a training corpus in [Amphetype](#), paste the contents into the "Lesson Generator" tab with the following settings:

Make ****3**** copies of the list



Divide into sublists of size ****3****

Add to sources as ****google-10000-english****

In the "Sources" tab, you should see **google-10000-english** available for training. Set WPM at 10 more than your current average, set accuracy to 98%, and you're set to train.

Enjoy!

Releases

No releases published

Packages

No packages published

Contributors 12

