corpus of contemporary

AI generated article from Bing

Word frequency list based on a 15 billion character corpus: BCC (BLCU ...

The Beijing Language and Culture University created a balanced corpus of 15 billion characters. It’s based on news (□□□□ 1946-2018□□□□□□□□ 2000-2018), literature (books by 472 authors, including a significant portion of non-Chinese writers), non-fiction books, blog and weibo entries as well as...

Word frequency list based on a 15 billion character corpus: BCC (BLCU ...

I would read in the BCC corpus frequency list as a dictionary, then Having concatenated all the news/magazine articles as plain text, I would build a dictionary of all the words in the news/magazine articles up to 8 characters long, counting their number of occurrences with the help of the BCC frequency list (which tells us which combinations ...

Integrating BCC Corpus Data into Dictionary - Pleco Software Forums

The BCC corpus seems to have pretty loose licensing terms. Pleco already seems to be using frequency data to sort the search results. Adding them meaningfully to dictionary definitions would be even better, I believe. That is something which printed dictionaries can’t do.

Common Idioms; A Collection by Grade [HSK / old HSK / □□ / □□ / ...]

The Beijing Language and Culture University created a balanced corpus of 15 billion characters. It’s based on news (□□□□ 1946-2018□□□□□□□□ 2000-2018), literature (books by 472 authors, including a significant portion of non-Chinese writers), non-fiction books, blog and weibo entries as well as classical Chinese.

Integrating BCC Corpus Data into Dictionary

I guess in my case, I could go with per-corpus flashcard sets to keep the per-corpus tagging, and one user dictionary (without tags) with all the per-corpus ranking info included in one entry per term.

Bigrams sorted by frequency with pinyin & English?

The Beijing Language and Culture University created a balanced corpus of 15 billion characters. It’s based on news (□□□□ 1946-2018□□□□□□□□ 2000-2018), literature (books by 472 authors, including a significant portion of non-Chinese writers), non-fiction books, blog and weibo entries as well as...

Sentences flashcards generator (Python script) - Pleco Software Forums

The Beijing Language and Culture University created a balanced corpus of 15 billion characters. It’s based on news (□□□□ 1946-2018□□□□□□□□ 2000-2018), literature (books by 472 authors, including a significant portion of non-Chinese writers), non-fiction books, blog and weibo entries as well as...

HIT IR-New Lab (Extended)

The frequency of some words is not less than 3 (the statistical result of a small-scale corpus), and 14,706 rare words and non-common words can be eliminated.

Flashcards for TOCFL (2023), CCCC, TBCL - Pleco Software Forums

I've parsed out vocabulary from these taiwanese tests and converted to flashcards in pleco's format. Useful e.g. for seeing term levels, intended part of speech and sometimes definitions/examples. TOCFL vocab was updated some couple years ago and I haven't yet seen a processed version of the...

Media-related vocabulary gathering project - Pleco Software Forums

With a small corpus of 650 articles from People's Daily, downloaded using a Python script, I hope to start providing a more modern frequency list of media-related vocabulary. The frequency list has the following features: It uses all sections of the □□□□ / People's Daily newspaper, including the sports section.