COCA: The Corpus of Contemporary American English

https://www.english-corpora.org/coca/

概要

・約 10 億語
・1990 年から 2019 まで 30 年× 2,500 万語以上ずつ

■ 8 つのジャンル

spoken
　テレビやラジオ番組のスクリプト
fiction
　文芸雑誌や本
magazines
　各種雑誌
newspapers
　新聞
academic texts
　学術雑誌
TV and Movies subtitles (2020-)
　テレビや映画の字幕（口語的）
blogs (2020-)
　Google が blog と分類している Web ページ
web pages (2020-)
　各種 Web ページ

■ 5 つの調べ方

1. 上位 6 万語の頻度リスト
    1. 語形
    2. 品詞
    3. レンジ
    4. 意味
    5. 発音
2. 単語検索
    1. 共起語
    2. トピック
    3. クラスター
    4.Web サイト
    5. コンコーダンス
    6. 関連語
3. 英語文章と COCA との比較
4. フレーズ・文字列検索
    1. 品詞を指定可能
5. ランダム検索
    1.「Words of the Day」

利用

■全部使うにはライセンスの購入が必要

■オンラインでの Web 利用

■データダウンロード

サンプル：890 万語

https://www.corpusdata.org/coca/samples/coca-samples-text.zip
元データからランダムに、100 分の一

■ Word/lemma/PoS

■ Linear text (ダウンロード 20MB で、解凍後 73MB)

1. テキスト ID に続き本文の英文を改行なしで。（例：　@@4001441 Our purpose ....）
2. 縮約形は分割：can't は ca n't に
3. 句読点は、前後にスペース
4. ファイルは 8 種

・伏字は「 @ @ @ @ @ @ @ @ @ @ 」
・段落は <p>

■ 英単語以外の「ゴミ」の前処理が必要

匿名記号の削除
　@ @ @ @ @ @ @ @ @ @
段落記号の削除
　<p>
テキスト ID の削除
　@@4000241

サンプルを使っての分析例

KWIC 検索の利用例

・ユーザー登録が必要（無料）

■手順

1. ログイン画面の「SEARCH」タブ
2.List, Chart, Word, Browse + とあるが、その一番右の + をクリックすると、さらにメニューが広がる
3. そこから KWIC を選ぶ
4. 入力欄に、調べたい単語を入れる
5. 下の「Keyword in Context (KWIC)」ボタンを押す
6. 結果画面で、目的の単語をクリック