**NLP Task 2**

**Niyati Sahni**

**17BCE0506**

**COCA(**[**Corpus of Contemporary American English**](https://www.english-corpora.org/coca/x1.asp)**)**

The Corpus of Contemporary American English (COCA) is the only [large](https://www.english-corpora.org/coca/help/texts.asp), genre-balanced corpus of American English. COCA is probably the [most widely-used corpus of English](http://www.english-corpora.org/users.asp), and it is related to many other [corpora of English](http://www.english-corpora.org/) that we have created, which offer unparalleled insight into [variation in English](http://www.english-corpora.org/variation.asp).

The corpus contains more than [**one billion** words](https://www.english-corpora.org/coca/help/texts.asp) of text (25+ million words each year 1990-2019) from eight genres: spoken, fiction, popular magazines, newspapers, academic texts, and (with the [update in March 2020](https://www.english-corpora.org/coca/help/new2020.asp)): TV and Movies subtitles, blogs, and other web pages.

Click on any of the links in the search form to the left for context-sensitive help, and to see the range of queries that the corpus offers.

**There are three main ways to search the corpus:**

First, you can [browse a frequency list](https://www.english-corpora.org/coca/browse.asp) of the top 60,000 words in the corpus, including searches by word form, part of speech, ranges in the 60,000 word list, and even by pronunciation. This should be particularly useful for language learners and teachers.

Second, you can [search by individual word](javascript:choose('words')), and see collocates, topics, clusters, websites, concordance lines, and related words for each of these words. Note that some of these searches are unique to COCA and iWeb.

Third, you can [search for phrases and strings](javascript:choose('ilist')). And because the corpus is optimized for speed, searches for substrings ([\*ism](javascript:x('*ism','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0')), [un\*able](javascript:x('un*able','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0'))) and phrases are very fast, e.g.: [got VERB-ed](javascript:x('got%20_vvn','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0')), [BUY \* ADJ NOUN](javascript:x('BUY%20*%20ADJ%20NOUN','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0')), ["gorgeous" NOUN](javascript:x('=gorgeous%20NOUN','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0')) -- and even high frequency phrases like: [from ADJ to ADJ](javascript:x('from%20ADJ%20to%20ADJ','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0')), [phrasal verbs](javascript:x('VERB%20_rp','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0')), or [NOUN NOUN](javascript:x('NOUN%20NOUN','0','1','0','0','0','0','0','0','0','0','0','0','100','0','1','0','0','0','0','0')).

You might pay special attention to the [comparisons between genres and years](javascript:toggle_visibility('sections')) and [virtual corpora](javascript:toggle_visibility('VC')), which allow you to create personalized collections of texts related to a particular area of interest.













