This document sets out how to start to formulate queries to find useful returns in large scale textual datasets. Normally, we’d also ask what type of results you want out of the system - returning full articles, or paragraphs, or sentences, but for this initial phase we’ll just be providing a count of the terms, to give a sense of how many “hits” (frequency query) you will have to set up a work flow to deal with.

For running a **frequency** query we need to define 6 parameters (some of them are optional):

1. **Target word(s) [Optional]**: Those are the **list of terms** that must appear in the text of an article in order to “select it”. The main function of the target word(s) is:
   * To filter out all the articles which those words not appear
   * To set up the context of our search.

Very often, we only have **1 target word** in our queries, but we have in *defoe* the mechanism in place to work with more than target word. In that case, we must to specify which type of selection we want to do:

* + - Type 2: Select all the articles which have at least **one of the target words**

**“Disorder” OR “Mental Health”**

1. **Lexicon**: Those are the list of terms that we are going to calculate their frequency over time. Following the previous example, our lexicon is “*apple”* and “*bread”*. Note that defoe is able to work with terms that have more than a word (e.g. *Diego Velazquez*)

**Lexicon 1: Neutral**

**Lexicon 2: Explicit**

**Lexicon 3: Implicit**

1. **Period of Time [Optional]**: We could specify if we want to select articles that are from a specify period of time (e.g. 1615 to 1800). This parameter is optional, so if we do not specify it, the frequency will be calculated using all the articles available for our search. Again, this feature is very useful, if we want to **narrow the initial selection of articles** to a specific **period of time**.

**From 1950s**

1. **Hint count**: We need to specify how we want to “count” the frequency over time. We have two options:
   * Article count: The query counts as a “hint” every time that finds **an article with a particular term** from our lexicon in the previously selected articles (by the target words or/and time period). **So, if a term is repeated several times in an article, it will be counted just as ONE**. In this way, we are basically calculating the **“frequency of articles” over time**.

This is useful, when we are interested to see **how many articles** talk about a particular term, and how its frequency changes over the years.

1. **Treatment of words [Optional]:** Finally, we could also specify how we want to **treat the words from our target word(s) and lexicon lists, and selected articles**. These are the options:
   * **normalize**: words are normalized by removing all non-'a-z|A-Z' characters, and automatically transformed to lower case.
   * **lemmatize**: words are lemmatized using lexical knowledge bases to get the correct base forms of each word. Like stemming, lemmatization reduces inflectional forms to a common base form. As opposed to stemming, lemmatization does not simply chop off inflections. Instead it uses lexical knowledge bases to get the correct base forms of words. For example, in English, the verb 'to walk' may appear as 'walk', 'walked', 'walks' or 'walking'. **The base form, 'walk'**, that one might look up in a dictionary, is called the lemma for the word. So, “walk” will be the lemma of 'walk', 'walked', 'walks' and 'walking'.

Stem and lemmatize treatment also include normalization. **And if we do not specify which treatment we want, by default the query applies lemmatization**.

Those treatments allow us to include in our query variations of the specified terms (from the lexicon and target words lists).

1. **Grouping** (the frequency) **results**:
   * **By time**

We have used the frequency query, to calculate the **frequency over time** of the following Scotties Philosophers using the *Encyclopaedia Britannica* as our corpus: *Francis Hutcheson, David Hume, Adam Smith, Dugald Stewart and Thomas Reid*.

* Target Word(s): None
* Lexicon: <https://github.com/alan-turing-institute/defoe/blob/master/queries/sc_philosophers.txt>
* Period of Time: Not specified. All the Encyclopaedia Britannica.
* Hint count: Term count.
* Treatment: Normalize (<https://github.com/alan-turing-institute/defoe/blob/master/queries/sc_philosophers.yml> )
* Grouping results: By Time

Results can be checked here: <https://github.com/alan-turing-institute/defoe_visualization/blob/master/NLS/results_NLS/results_ks_philosophers_norm>

A close up of a logo

Description automatically generated

Figure 1: Notebook: https://github.com/alan-turing-institute/defoe\_visualization/blob/master/NLS/Visualization\_Frequency\_Taxonomy\_Ngrams.ipynb