A Using the Calculator

In this appendix we provide instructions on the use of the page calculator. The code is available on GitHub at `https://github.com/leifos/ifind` .There are a number of packages which need to be installed to use the calculator. These are as follows: nltk, json, beautiful soup, phantomjs, selenium, redis and PIL. To use the tool you should run the following from the command line

```
python generic_exp.py -u url -e engine -k key -d domain -c cutoff
-m maxqueries -s stopwordfile -ca cache
```

The arguments are as follows: url is the url of the page you want to score, e is the engine you want to use (which is one of bing, govuk, sitebing, twitter, wikipedia). If using sitebing or bing, you need to provide your own API key. If you are using sitebing and you wish to use a domain other than gla.ac.uk you will need to set d to be the domain. Cutoff is the number of pages within which we check to see if the target url is returned, this defaults to 50. The maximum queries value is the maximum number of queries which will be issued to the search engine, this defaults to 250. The stopword file is a file which contains typical stopwords which are ignored when generating queries. The stopword file provided on github is called 'stopwords.txt'. If you wish to use a cache, you must install redis and run a redis server before issuing the queries (see `http://redis.io/topics/quickstart`). Upon running the file with the arguments, the page will be fetched. A number of questions will then be asked regarding how you want to generate the queries. The questions are as follows:

- Do you want to use only a position based extractor? - Enter y if you want to limit your queries to divs with specified ids or to exclude divs with specified ids (use this option if you want to run it for the whole page without ranking queries).
  - If you enter y you will be asked if you want to include (enter i) or exclude (enter e) divs. You should then enter the ids of the divs you want to include or exclude. Exclude none for the whole page.
  - You will then be asked if you want to use a percentage of the content, if so enter an integer value for the percentage to be used. You will then be asked if you wish to use a specific number of words, if yes then enter the number of words to be used. You should select one or the other.
- Do you want to use only a rank based extractor? - Enter y if you want to use a rank based extractor, note you need a background file which you will be asked for. This defaults to background.txt which contains a term occurrences file from the gla.ac.uk website.
- Do you want to use a rank based extractor combined with a position extractor? Enter y if you want to rank a particular portion of a webpage e.g. a div. The questions asked as above for a position based extractor will be asked and a background file will be asked for.

Upon completion of answering these questions, the queries will be generated and a report will be displayed on screen which includes statistics such as the number

of queries generated, the number of queries which returned the page, and the cumulative and gravity based scores for the page.