User Manual Meme application

The following figure shows the program interface. The left sector of the image shows the many different configurations that are possible, such as the thresholds *alpha (e.g. 0.35)*, *ro* (e.g. 0.5) and *sigma (e.g. 0.5)*, as well as the outputs that can be produced and the filters which control the format of the output. The right part of the image corresponds to the application activity history log.

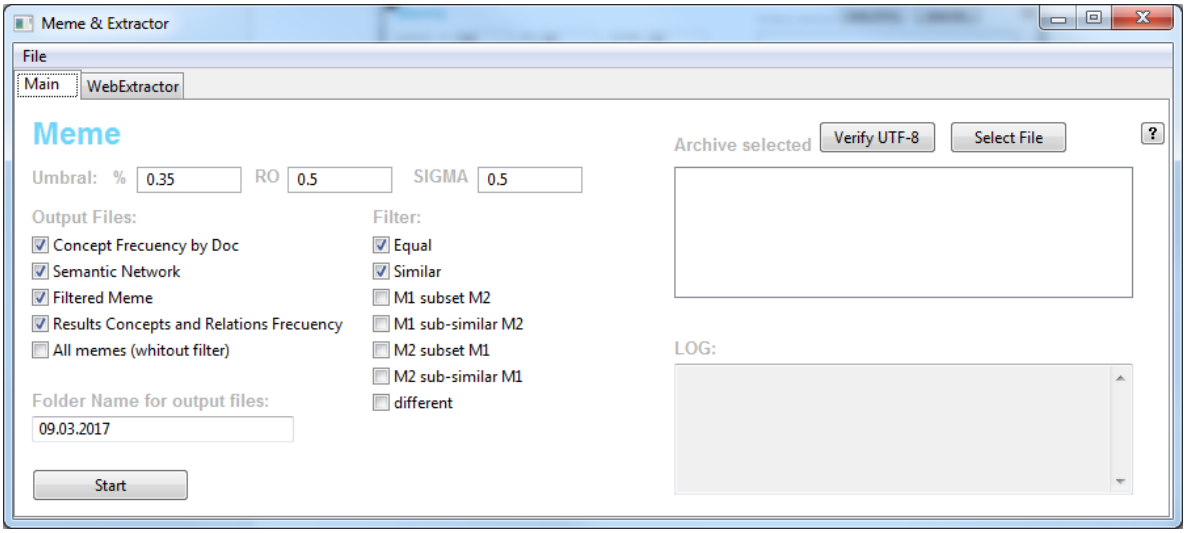
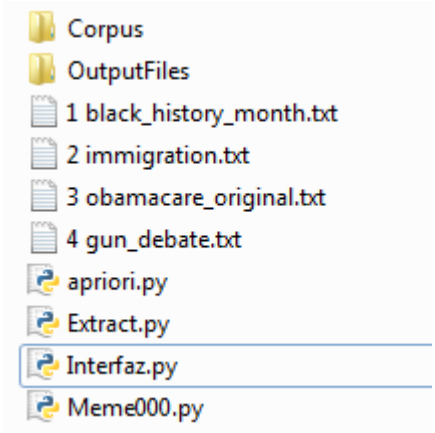


Figure 1. Meme application interface

The structure of the files is as follows, where the folder "Corpus" corresponds to the corpus used and the folder "OutputFiles" corresponds to the output files to be generated by the program. In addition, the text files that correspond to the input files to be processed are displayed.

To run the program, run the file "Interface.py".



Simple execution

An example sample to use the program is shown as follows:

- A UTF-8 text file must be entered to avoid character problems, using the Select File button.
- Configure the program parameters.
- Press the Start button to start generating semantic networks and memes.

Advanced configuration

For the Alpha field, it is possible to enter more than one value, for example "0.1 0.2 0.3 0.4" corresponding to 10, 20, 30 and 40 percent.

Extra

To access the interface information, there is a "?" Button, which gives the description of each button or field.

In addition, it is possible to change the name of the folder that will contain the output files (by default assigned as "OutputFiles")

Parameter assignments

Table 1
Best values found for each document corpus

CORPUS	N-gram_ size (*NS*)	Context_ window_ size (*CW*)	Concept_ freq_ thresh (α)	Concept_ sim_ thresh (ε)	Relation_ sim_ thresh (θ)	Meme_ sim_ thresh (τ)	Concept_ weight (ρ)	F-score
BHM	1	5	0.10	0.1	0.1	0.3	0.60	0.538
DU	1	5	0.05	0.1	0.1	0.3	0.70	0.733
GD	1	5	0.05	0.1	0.1	0.3	0.75	0.561
IM	1	5	0.05	0.1	0.1	0.3	0.70	0.564

In Table 1 we summarize the best parameters automatically found for each document corpus. Firstly, it can be seen that only two parameters vary depending on the corpus being processed: concept frequency threshold and concept weight. Corpus BHM has a slightly higher concept frequency threshold (because more concepts are identified for this corpus) and lower concept weight (as there are more concepts we can give more weight to relations). Likewise, GD can be seen to have a slightly higher concept weight (0.75) because there are less relations identified in this corpus. The results indicate that optimum processing would be obtained by adjusting the concept frequency threshold and/or concept weight for each document corpus.

Note that with reference to the screen image of Fig. 1, the first parameter (Umbral % 0.35) corresponds to the concept freq. threshold α of Table 1; the second parameter (RO 0.5) corresponds to the concept weight (ρ) of Table 1; the third parameter (SIGMA 0.5) corresponds to the relation weight which is not in Table 1 but is assigned as 1-ρ .

Please use the following reference citation in your documents and papers:

Héctor Beck-Fernandez, David F. Nettleton, Lorena Recalde, Diego Saez-Trumper, Alexis Barahona-Peñaranda, "A System for Extracting and Comparing Memes in Online Forums", Expert Systems with Applications, Volume 82, 1 October 2017, Pages 231–251.

Note that a preprint version of the paper which described the system in detail is included in the github project folder.