

Creative Computing II
Multimedia Information Retrieval Systems
26th January 2010

This lab sheet surveys existing Information Retrieval systems available for multimedia resources on the Web.

1. This part is about the Porter Stemmer.
 - (a) Find and download an implementation of the Porter Stemmer from the Internet. Note carefully the licencing terms under which the implementation is offered; if the terms under which you can use and modify the code are too restrictive, do not use it (and try to find another implementation).
 - (b) Compile and run the implementation. Try stemming various words, to understand better what the system does.

As discussed in lectures, a stemmer allows a certain amount of normalization of user input, converting multiple different forms of conceptually the ‘same’ word to a single canonical form. Of course, this can’t be done without some loss of information; sometimes, two completely different words end up having the same stem – but this can happen even without stemming: homographs (words with the same spelling but different meanings) are relatively common.
 - (c) Unless the implementation is already suitable, convert it so that it is usable from within *Processing*.
2. This part is about surveying, classifying and categorizing existing systems for multimedia information retrieval on the Web. Your task is to identify resources which allow users to retrieve multimedia content and metadata.
 - (a) Try to find systems for retrieval of
 - text;
*Google and other web search engines are probably the obvious answer here; also the find functionality usually accessible with **Ctrl-F** in many ordinary computer applications.*
 - images;
Google images, but also flickr tags.
 - symbolic music;
Musipedia, which allows content-based search through various input methods, including the Parsons Code
 - videos;
YouTube
 - musical audio;
Social radio sites such as last.fm and spotify; content services such as Shazam; metadata aggregators such as Apple’s ‘Genius’; also music retailers such as Amazon which allow listening to clips.

- films;
imdb is probably the most comprehensive example. Also services such as Netflix and Lovefilm, some of which do social playlisting.
 - television programmes;
imdb contains significant information about television programmes; content these days is available from the BBC's iplayer and similar services on other channels. The lookup system from Video+ identifiers to broadcast times is another information retrieval system, used by smart boxes.
 - web pages;
Google, Yahoo! (soon to be powered by Microsoft's Bing), AllTheWeb, Duck-DuckGo, ... the list is endless.
 - scientific papers;
CiteSeer, Google Scholar, arXiv.org, various institutional e-prints repositories.
 - web pages which no longer exist;
Google's cache will allow you to look up web pages by url, but only preserves the most recently crawled copy. By contrast, archive.org maintains a cache of all the crawled copies it has acquired, and displays a list of when it has detected a page change.
 - published books.
- (b) Categorize them as to whether they allow query by
- content;
 - metadata;
 - name;
 - reference;
 - other identifying information.
- (c) For each system, determine
- whether the item retrieved is the multimedia content itself or metadata about the content.
 - whether items that the system retrieves are 'exact' matches or over some scale of 'similarity'
 - whether there exists an 'API' giving programmatic access to the retrieval system (rather than a simple web form), and if so whether that API has any functionality that is not available to users of the web form.
- (d) Which of the systems that you have enumerated in part 2a do you think work well, and which do not? Comment and explain. How would you improve on the ones which do not work well?

Other resources:

- van Rijsbergen, C. J., S. E. Robertson and M. F. Porter, *New models in probabilistic information retrieval* (1980).
- Porter, M. F. *An algorithm for suffix stripping*. Program **14**(3) 130–137