UCLA Communication Studies Archive

Edge Search Engine

Regular Expressions Guide

Updated 2012-08-12

The search engine relies on the BRICS library, developed by Anders Møller at Aarhus University, which "implements real, unrestricted regular operations" (see details).

This page is an effort to develop a set of examples to illustrate the power of this regex library. Please add to the list.

For testing purposes, it may be useful to run a search against a single day for a single network, or even a single show.

Examples

```
/[A-Za-z]{10}/                          -- words at least ten letters long
/[0-9A-Za-z]{10}/                       -- words 10 alphanumeric characters long
/[A-Za-z]{10,12}\ /                     -- 10 to 12 letters followed by a space
/[A-Za-z]{10,12}(\ |\,|\.)/             -- 10-12 followed by a space, comma, or period
/[A-Za-z-]{10,12}S/                     -- including hyphens, and end in S
/[A-Za-z-]{12,}(\!|\?|\.|\,|\:|\;|\ )/  -- 12+ and then space or punctuation
/G[A-Za-z-]{10,12}\ /                    -- G followed by 10-12 letters and then a space
/\ G[A-Za-z-]{10,12}\ /                  -- same, but G is the first letter
/\ A\ G[A-Za-z-]{10,12}\ /               -- same, preceded by an indefinite article
/\ A\ G[A-Za-z-]{10,}\ /                 -- same, but no maximum number of letters
/IS\ NOW\ [A-Za-z-]{2,}ING\ /   -- "is now *ing"
```

Kinds of searches to avoid

/\ A\ G[A-Za-z-]{,12}\ / -- a capital indefinite article, followed by a capital G, followed up up to 12, but no minimum number of letters (dangerous -- is not completing after half an hour)

/[0-9A-Za-z_]{10,12}/ -- equivalent of [[:alnum:]] -- alphanumeric and underscore. We don't use underscore in captions, so [0-9A-Z] is enough

/[0-9A-Za-z]{1,}/ -- I think this is equivalent of [[:alnum:]]\+ -- meaning "one or more times of an alphanumeric character"

On a single day for one network, it found 45883 hits. When I tried to view them, I got:

```
Fatal error: Allowed memory size of 134217728 bytes exhausted (tried to allocate 835 bytes) in
/info-data/www-data/tna/edge/classes/TnaSearchDispatcher.php on line 329
```

/[[:alnum:]]{10,12} -- not finishing in twenty minutes on a single day of CNN files; in contrast, the version above took one second.

Questions

1. What characters need to be backslashed?

Summary

* searches must be bounded by /
* [0-9A-Za-z]{n,m} works fast, with variants {,m} (dangerous) and {n,}
* space, comma, period, underscore must be escaped
* hyphen can be included with [A-Z-] -- place it last