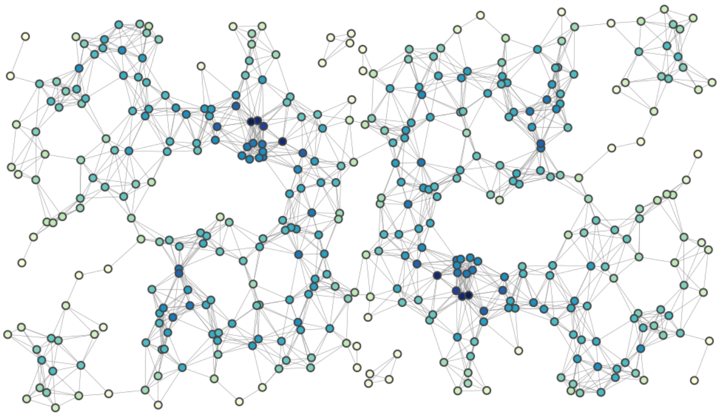

Old Dominion University

CS 432 Web Science

Assignment Three

Derek Goddeau

Professor

Michael L. Nelson

February 23, 2017

1 Download the HTML for the 1000 URIs

To download the HTML for the URIs the bash script `get_html.sh` is used. The `fetch()` function does all the work, downloading the HTML using `wget` and creating a `SHA-1` hash of the URI to store locally.

```bash
fetch() {
    while read uri; do
        local hash=$(echo -n "$uri" | sha1sum | cut -d ' ' -f 1)
        local hash+=".html"
        wget -O data/raw_html/"$hash" "$uri"
        if [[ "$?" != 0 ]]; then
            echo >&2 '[*] Error downloading file'
            FAILURES=$(expr FAILURES + 1)
        else
            echo >&2 '[*] Success'
        fi
    done < "$FILE"
}
```

2 Calculate TFIDF

```python
def get_num_mementos(link):

    url = 'http://memgator.cs.odu.edu/timemap/json/http://' + link

    try:

        mementos = requests.get(url).json()

    except ValueError as e:

        print("No memento for URL: {}".format(link))

        return 0

    num_mementos = len(mementos['mementos']['list'])

    return num_mementos


link_mementos = []

for link in final_links:

    link_mementos.append((link, get_num_mementos(link)))
```

```r
   %%R -i data


       library(plotly);


       p <- plot_ly(x = data, type = "histogram")


       embed_notebook(p)

       htmlwidgets::saveWidget(as.widget(p), "histogram.html")
```

3 Rank by PageRank