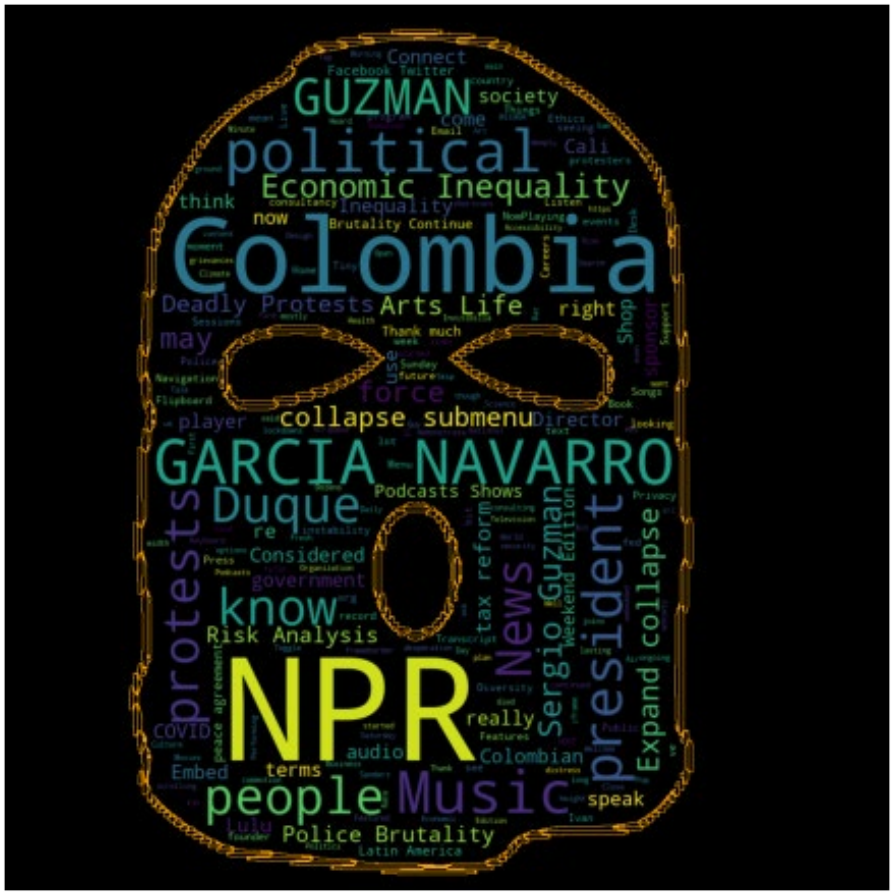
2021

ART384 Creative Coding

Exercise 3

Python – create a web scraping wordcloud generating robot



Wordcloud based on a NPR report: Deadly Protests Against Economic Inequality And Police Brutality Continue In Colombia. May 9, 20217:53 AM ET. Overlay with ski mask.

The third assignment asks you to combine multiple different code modules developed in class to a new program . You are asked to combine a module that collects URLs from a top-level domain, and then to collect the text content from these sites and create a word cloud display from them as shown above.

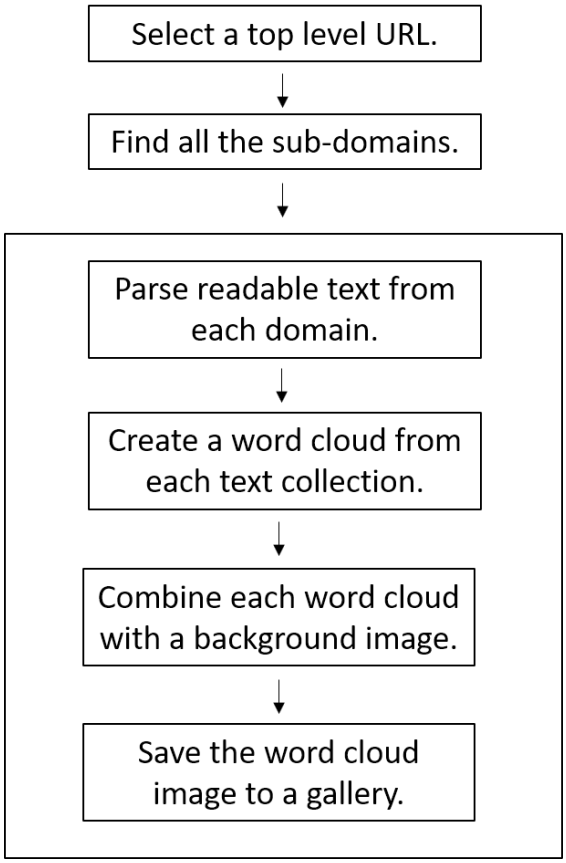


Diagram of the task

The working modules you will have at your disposal include:

1. Code to find subdomains of a given URL. [Beautiful Soup library]
2. Code to parse a website into readable text. [NLTK library]
3. Code to create a wordcloud from a collection of words. [word cloud library]

Here are snippets from each of the modules listed above:

1

```
#create a data soup from a selected URL
URL = "https://www.nytimes.com/"
page = requests.get(URL)
soup = BeautifulSoup(page.content, "html.parser" )

#append all the URLs into a list (updated_list)
alist = []
updated_list = []
term = 'https'

for tag in soup.find_all('a', href=True):
    text = tag['href']
    if(text.find(term) != -1):
        alist.append(text)
[updated_list.append(x) for x in alist if x not in updated_list]
```

2

```
#collect the text fom a website (here: specialpage.html)
 url = 'specialpage.html'
 html = urllib.request.urlopen(url).read()
 webtext = text_from_html(html)
```

3

```
#generate the wordcloud
wordcloud = WordCloud(width=xdim, height=ydim, mask=wave_mask, random_state=80,  contour_width
=1, contour_color='orange').generate(" ".join(token_text))

#convert to image and save
image = wordcloud.to_image()
image.save(datapath+'webtext.jpg', 'JPEG')
```

Each of these modules will be available in a python script. You have to combine the elements and create one single program from them. The solution could look something like this:

```
#put all the functions into a helper file
from assignment3_helper import *

#select a top level url
top_url = 'washingtonpost'
mask_image = 'someimage.png'
sub_urls = get_unique_urls_from_url(top_url)

for each u in sub_urls:
        token_text = collect_readable_text_from_url (u)
        create_wordcloud(token_text, mask_image)
```

We will begin the process together in class.

DELIVERABLE

A python program – a robot - that combines all the code elements required to parse text and create a word cloud. The program should take as an input one URL and create a wordcloud image for each sub domain it finds in the input URL. Save the images to your drive
{Extra credit: create a gallery of images from all the results}

POINTS 15 points. See syllabus. **Extra Credit** 5 points.

DEADLINE Tuesday, November 16st, noon

FORMAT and DELIVERY

- Demonstrate the code in class.
- Place copies of the python code (and any other files you used) onto UBBox, and send the instructor a link to the folder to before the deadline.