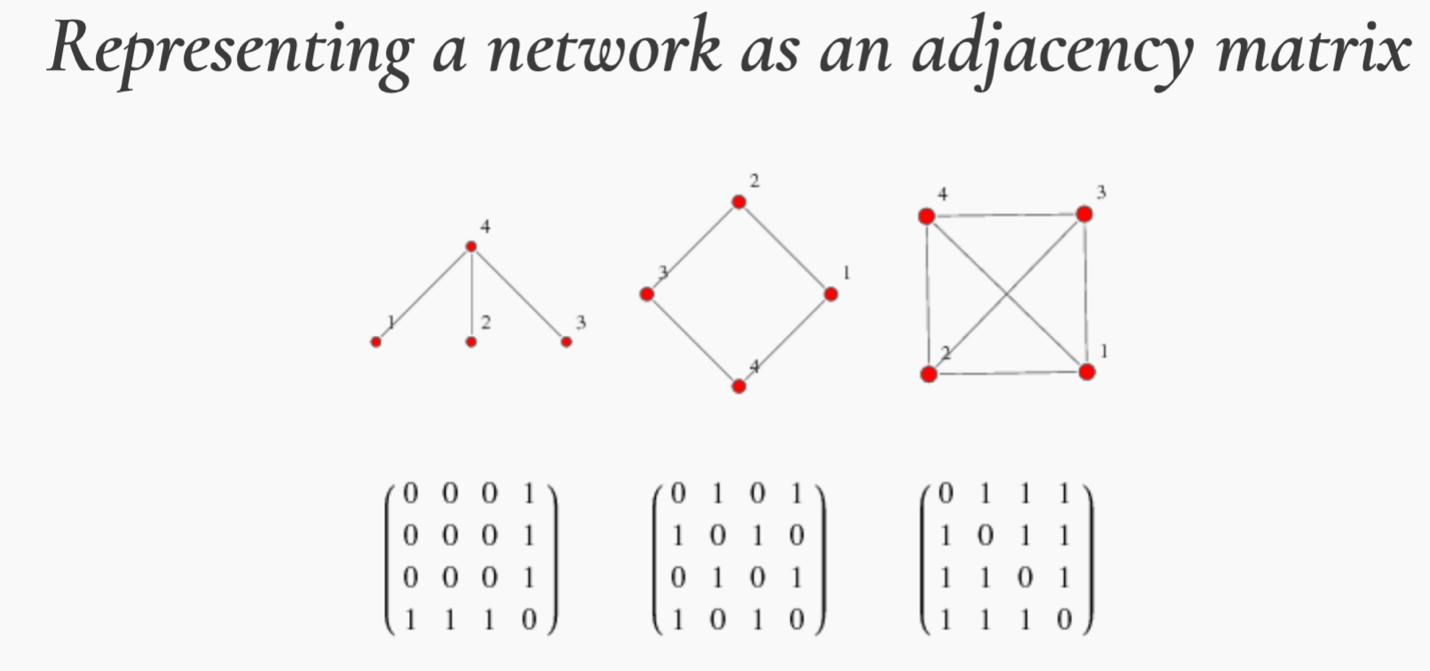
BIA660A midterm questions

1. the facility of translating between different representations of network, like the graphical distribution, the picture of it and adjacency matrices



1. In one sentence, describe the difference of a measure of node centrality of your choosing and pagerank algorithm

<https://en.wikipedia.org/wiki/Network_science#Network_properties>

Centrality indices produce rankings which seek to identify the most important nodes in a network model. Different centrality indices encode different contexts for the word "importance." The betweenness centrality, for example, **considers a node highly important if it form bridges between many other nodes**. The eigenvalue centrality, in contrast, **considers a node highly important if many other highly important nodes link to it**.

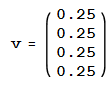
1. PageRank algorithm, damping factor

**Step 1 – Represent the network as an adjacency matrix**

Assume a network is composed of n pages. Each page represents a node in graph. If there’s link between 2 nodes, the element in the matrix will be 1; otherwise, the element will be 0. Then we can draw an adjacency matrix for the network.

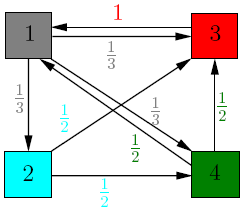
**Step 2 - Augment the matrix by making a few assumptions about the web surfer**

The most important assumption states that the probability of each page being visited by a web surfer is equal. Thus we can have a first rank vector with each element equal 1/n, where n is the number of pages or nodes. In the example below, n = 4.



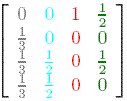
**Step 3 - Convert the network’s adjacency matrix into a stochastic matrix**

Let’s value the rank of each node as 1. The rank of a page is divided among its forward links evenly to contribute to the ranks of the pages they point to.



For example, node 1 has 3 outgoing links, then the weight to each link is 1/3. Same to node 2,3 and 4.

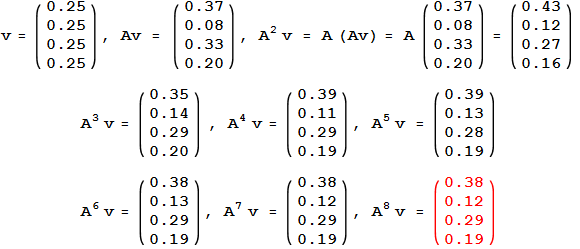
Let A(i,j) be the stochastic matrix transformed from the graph above. Each element in the matrix represents the rank of a page evenly contributing to other pages. If there is no link between to pages, the element will be zero. Then we can have the stochastic matrix as follows.

A(i,j) =

We can notice that in each column, the sum of elements equals 1.

**Step 4 - Exponentiate matrix**

Firstly, we update the rank of each page by adding to the current value the importance of the incoming links. This is the same as multiplying the matrix *A* with *v* . Then iterate the process, and we can notice that the sequences of iterates *v*, *Av*, ..., *A*k*v* tends to the equilibrium value *v*\*. We call this the PageRank vector of our web graph. Numeric computations give:



1. network properties and particularly characteristic path length

define characteristic path length, draw a network that its path length is 1, draw another that its path length is larger than 2.

<https://en.wikipedia.org/wiki/Network_science#Network_properties>

size-how many nodes

density-links / possible links

The average shortest path length is calculated by finding the shortest path between all pairs of nodes, and taking the average over all paths of the length thereof (the length being the number intermediate edges contained in the path, i.e., the distance du,v, between the two vertices u,v within the graph). This shows us, on average, the number of steps it takes to get from one member of the network to another.

1. history of the web

describe the relationship between Tim, Cern and www.

See the materials of the first class

1. response status

https://cds.cern.ch/record/369245/files/dd-89-001.pdf

1. regular expression, choose at least 3.

网上有很多，随便找个看一下，wikipedia就行

. 代表任意字符

\*代表0或任意

+代表1或任意

/d代表数字

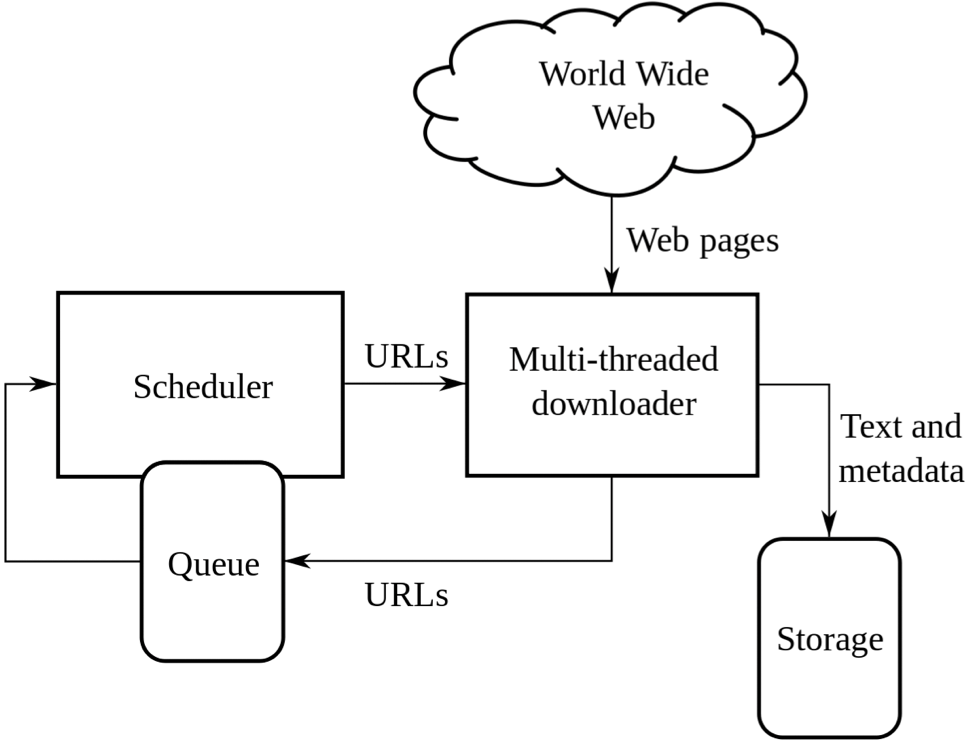
/w代表字母

/s代表空格

/n代表换行

？代表非贪婪

1. architecture of web crawlers



1. url deduplication

url deduplication is to deduplicate the same urls; otherwise, the crawler will go to the same page again and again. For example, the following 3 urls all link to the page Earth.

<https://en.wikipedia.org/wiki/Earth?foo=bar>

<https://en.wikipedia.org/wiki/Earth>?

https://en.wikipedia.org/wiki/Earth

这段代码只能deduplicate完全一致的url，不能deduplicate上面这个情况。

from urllib.parse import urlparse #urlparse用来清洗url

from urllib.parse import urljoin #urljoin用来将url的两个部分合并到一起

﻿for link in links:

u = link.get('href')

if not is\_absolute(u):

u = urljoin(url, u)

u\_parse = urlparse(u) *# Change the type of u into urlparse for the if test below.*

if "www.stevens.edu" in u\_parse.netloc: *# Avoid external URLs like "www.facebook.com/...www.stevens.edu...".*

if u not in url\_list: # Deduplicate

url\_list.append(u)

q.put(u)

1. web crawler’s revisit policy

In one sentence, define revisit policy

Explain what is refreshness and age

a re-visit policy states when to check for changes to the pages

Web crawlers use revisiting policies to determine the cost associated with an outdated resource. The goal is to minimize this cost. This is important because resources in the Web are continually created, updated or deleted; all within the time it takes a web crawler to finish its crawl through the Web. It is undesirable for the search engine to return an outdated copy of the resource. The cost to revisit the page are based on freshness and age, where freshness focuses on whether or not the local copy is the current copy of the resource and age focuses on how long ago the local copy was updated.

1. the levels of a linguistic analysis how those may or may not met onto different nlp tools

at what level of a linguistic analysis is POS, give reasons.

1. tf-idf

why idf

<https://blog.csdn.net/zrc199021/article/details/53728499>

tf = term frequency = the number of times the term X occurs in a document / the total number of terms in a document

词频 = 一篇文本中X出现次数 / 文本词语总数

idf = inverse document frequency = log ( the number of all documents / (the number of the documents that have the term X in it + 1) )

逆文本频率 = 文本库总数 除以 （包含X的文本数+1），再去对数

+1是为了防止包含X的文本数=0

1. vector space models of words and the data that is used to train them

What kind of data can be used to train vector space model

vector space model can calculate the similarity of 2 words.

For example, cat and dog, the similarity is close to 1 (夹角cosα接近 0)

Where do vector space models come from?

2 approaches:

CBOW (continuous bag of words)： use t-2, t-1, t1, t2 to predict t

Skip-gram: use t to predict t-2, t-1, t1, t2

1. politeness policy and in particular robots.txt files

write a robots.txt, wait 2 seconds, avoid visiting /Secrets

• Waiting (How long should it span between each crawl?)  
• Crawler identification via User-agent field (Who can crawl the web?)

• robots.txt (该文件包括了以上两个问题，并还规定了Which part of the web can be crawled?)

www.stevens.edu/robots.txt

1. sentiment analysis

how to use a few words to train a model for sentiment analysis

1. analyzing the success or failure of sentiment analysis algorithms

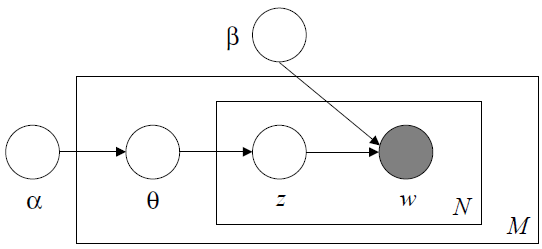
high recall low precision

1. topic modeling and in particular latent dirichlet allocation

precisely explain the relationship of topics and words

Topic modelling is the process of identifying topics in a set of documents. This can be useful for search engines, customer service automation, and any other instance where knowing the topics of documents is important. There are multiple methods of going about doing this. One of them is Latent Dirichlet Allocation.

LDA is form of unsupervised learning that views documents as bags of words (ie order does not matter). LDA works by first making a key assumption: the way a document was generated was by picking a set of topics and then for each topic picking a set of words.



1. browser automation tool for web scraping

when are browser automation tools useful

﻿from selenium import webdriver

﻿driver = webdriver.Chrome(executable\_path='C:/Users/…/chromedriver.exe')

1. something about networks

small world