Web Mining – Lab 5

HITS Algorithm

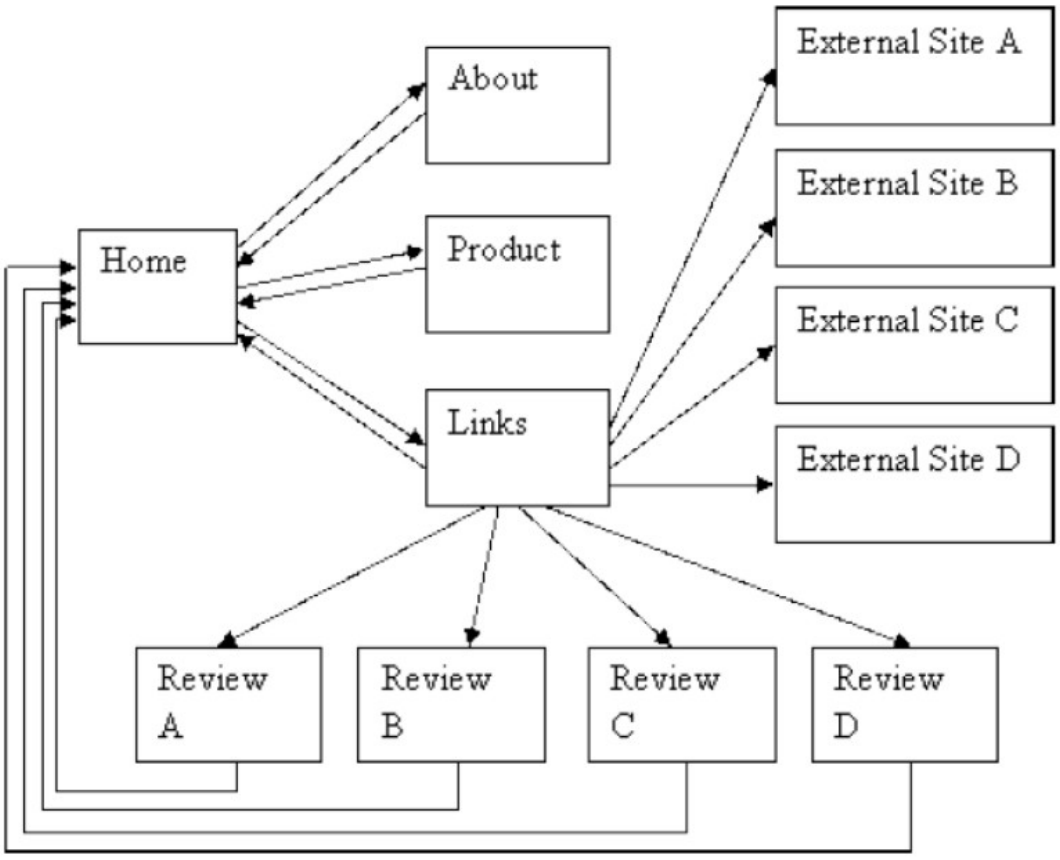
By Abhijeet Ambadekar (16BCE1156)

Lab task for the web mining lab on 5 th September 2018

[a] Use the following linkage data. Assume each site consists of only one text page.
a. Site A(outlinks to B,C,D)
b. Site B(outlinks to A,C,D)
c. Site C(outlinks to D)
d. Site D(outlinks to C,E)
e. Site E(outlinks to B,C,D)

Find the normalized hub score and authority score of all these pages using HITS algorithm.

[b] use your above program to find the final normalized hub score and authority score for the following web graph.



Program Code:

```
from pprint import pprint

def maptodictmat(nodemap):
	n = len(nodemap.keys())
	sitemat = {k:[0 for j in range(n)] for k in sorted(nodemap.keys())}
	for k1 in sorted(nodemap.keys()):
		for j, k2 in enumerate(sorted(nodemap.keys())):
			if k2 in nodemap[k1]:
				sitemat[k1][j] = 1

	return sitemat

class HITSRanker():
	def __init__(self, num_iter):
		self.num_iter = num_iter

	def fit(self, nodemap):
		self.nodemap = nodemap
		self.idx = maptodictmat(self.nodemap)
		self.idxmat = [self.idx[k] for k in sorted(self.idx.keys())]

		self.authorityscore = {k:1 for k in self.idx.keys()}
		self.hubscore = {k:1 for k in self.idx.keys()}


	def normalize_vec(self, vec):
		temp_sum = sum(vec.values())
		return {k:vec[k]/temp_sum for k in sorted(vec.keys())}


	def calculateA(self):
		for curr_page_index, curr_page in enumerate(sorted(self.idx.keys())):
			for page_index, page in enumerate(sorted(self.idx.keys())):
				if self.idxmat[curr_page_index][page_index]:
					self.authorityscore[curr_page] += self.hubscore[page]

		self.authorityscore = self.normalize_vec(self.authorityscore)

	def calculateH(self):
		for curr_page_index, curr_page in enumerate(sorted(self.idx.keys())):
			for page_index, page in enumerate(sorted(self.idx.keys())):
				if self.idxmat[page_index][curr_page_index]:
					self.hubscore[curr_page] += self.authorityscore[page]

		self.hubscore = self.normalize_vec(self.hubscore)

	def calculate_HA_score(self):
		for _ in range(self.num_iter):
			self.calculateA()
			self.calculateH()

	def print_details(self):

		print("The hub score of all the nodes are:")
		pprint(self.hubscore)
```

```
            print("The authority score of all the nodes are:")
            pprint(self.authorityscore)

            print("Sum of hub score in the end", sum(self.hubscore.values()))
            print("Sum of authority score in the end",
sum(self.authorityscore.values()))


if __name__ == '__main__':

      dummy_nodemap = {
            "A" : ["B", "C", "D"],
            "B" : ["A", "C", "D"],
            "C" : ["D"],
            "D" : ["C", "E"],
            "E" : ["B", "C", "D"]
      }

      hs = HITSRanker(num_iter=50)
      hs.fit(dummy_nodemap)
      hs.calculate_HA_score()
      hs.print_details()


      dummy_nodemap2 = {
            "Home" : ["About", "Product", "Links"],
            "About" : ["Home"],
            "Product" : ["Home"],
            "Links" : ["Home", "Review A", "Review B", "Review C", "Review D", "External
Site A", "External Site B", "External Site C", "External Site D"],
            "Review A" : ["Home"],
            "Review B" : ["Home"],
            "Review C" : ["Home"],
            "Review D" : ["Home"],
            "External A" : [],
            "External B" : [],
            "External C" : [],
            "External D" : [],

      }

      hs2 = HITSRanker(num_iter=50)
      hs2.fit(dummy_nodemap2)
      hs2.calculate_HA_score()
      hs2.print_details()
```

Output:

Part (a)

```
/media/anonymous/Work/Vit/Semester 5/WM/Lab/L5 - HITS Algorithm > python3 WM_HITS.py
The hub score of all the nodes are:
{'A': 0.08801146414810879,
 'B': 0.20288637217617891,
 'C': 0.3352927215619037,
 'D': 0.32941455687619264,
 'E': 0.044394885237615964}
The authority score of all the nodes are:
{'A': 0.2713767354683471,
 'B': 0.23544473261676774,
 'C': 0.10303838323104299,
 'D': 0.1187634132154951,
 'E': 0.2713767354683471}
Sum of hub score in the end 1.0
Sum of authority score in the end 1.0
```

Part (b)

```
/media/anonymous/Work/Vit/Semester 5/WM/Lab/L5 - HITS Algorithm > python3 WM_HITS.py
The hub score of all the nodes are:
{'About': 2.123779751683406e-11,
 'External A': 1.9335994375673552e-25,
 'External B': 1.9335994375673552e-25,
 'External C': 1.9335994375673552e-25,
 'External D': 1.9335994375673552e-25,
 'Home': 0.4999999999424931,
 'Links': 2.123779751683406e-11,
 'Product': 2.123779751683406e-11,
 'Review A': 0.12499999999844835,
 'Review B': 0.12499999999844835,
 'Review C': 0.12499999999844835,
 'Review D': 0.12499999999844835}
The authority score of all the nodes are:
{'About': 0.12499999999326634,
 'External A': 1.1571826104989418e-35,
 'External B': 1.1571826104989418e-35,
 'External C': 1.1571826104989418e-35,
 'External D': 1.1571826104989418e-35,
 'Home': 2.98417567981101e-11,
 'Links': 0.25000000001056044,
 'Product': 0.12499999999326634,
 'Review A': 0.12499999999326634,
 'Review B': 0.12499999999326634,
 'Review C': 0.12499999999326634,
 'Review D': 0.12499999999326634}
Sum of hub score in the end 0.9999999999999999
Sum of authority score in the end 1.0000000000000004
```