Web Mining Programming HW2

Main Tasks

- Task1: Implement PageRank algorithm on given two web-graph dataset
- Task2: Implement LexRank , a summarization technique which inspired from PageRank.

PAGERANK

- $P_0(i) = 1$
- $P_k(i) = (1\text{-}d) + d\sum_{(j,i)\in E}\frac{P(j)}{O_j}$
- Until $|p_k - p_{k\text{-}1}| < \varepsilon$ (Euclidean distance)
- In this assignment, use d = 0.85 in your final program
- You need to avoid adding edges explicitly to every node for those nodes with zero out degree.

DATASET

Data	Max. Memory	Nodes	Links
CS stanford.edu (Oct. 2004)	12M	50184	287844
stanford.edu (Aug. 2003)	54M	350004	918323

Note that the given datasets are very sparse, you can't implement the algorithm storing them in dense matrix

Input graph example

- #maxnode 6
- 1:3 2 4 6
- 2:3 4 5 6
- 3:1 4
- 4:2 1 6
- 5:2 4 6
- Each line is node_id:out_degree node_a, node_b ..
- The third line means (node 3)->(node 4)

Output Example

- 1:1.03486
- 2:0.695335
- 3:0.402125
- 4:1.48879
- 5:0.599137
- 6:1.77971
- Each line is node_number:pagerank_score

Evaluation

- We release our answer for data “stanford.edu (Aug. 2003)”
- We provide a program to evaluate the L1-norm and Spearman’s rho taking two pagerank vector as input.

- $ ./eval-pagerank answer.pagerank student.pagerank
- L1 norm: 0.332361
- Spearman's rho: -0.561746

Requirement(PageRank)

- Program Format :
- ./compute-pagerank graph_path
- Implement the PageRank algorithm on CS stanford.edu (Oct. 2004) and stanford.edu (Aug. 2003) and upload results
- Damping factor = 0.85
- ε = 1e-6

LexRank

- A graph-based method to summarize texts using PageRank idea.
- Nodes : each sentence is a node (a page in PageRank)
- Links : There are a link between two sentences if their similarity is large than threshold T.
- After constructing the LexRank graph, you can run PageRank algorithm on it, and then sorting each sentence by their score.

Sentence Similarity

- Similarity between two sentences Sim(S1,S2):
- We use tf-idf vector space model to construct sentence vector and use the cosine value between two sentence vector as the similarity.
- You need to count each term's frequency in the sentence.
- We would provide idf value of each term.

LexRank Example

1 Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.

2 Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.

3 Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.

4 Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.

5 The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.

6 Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, ``will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."

7 Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).

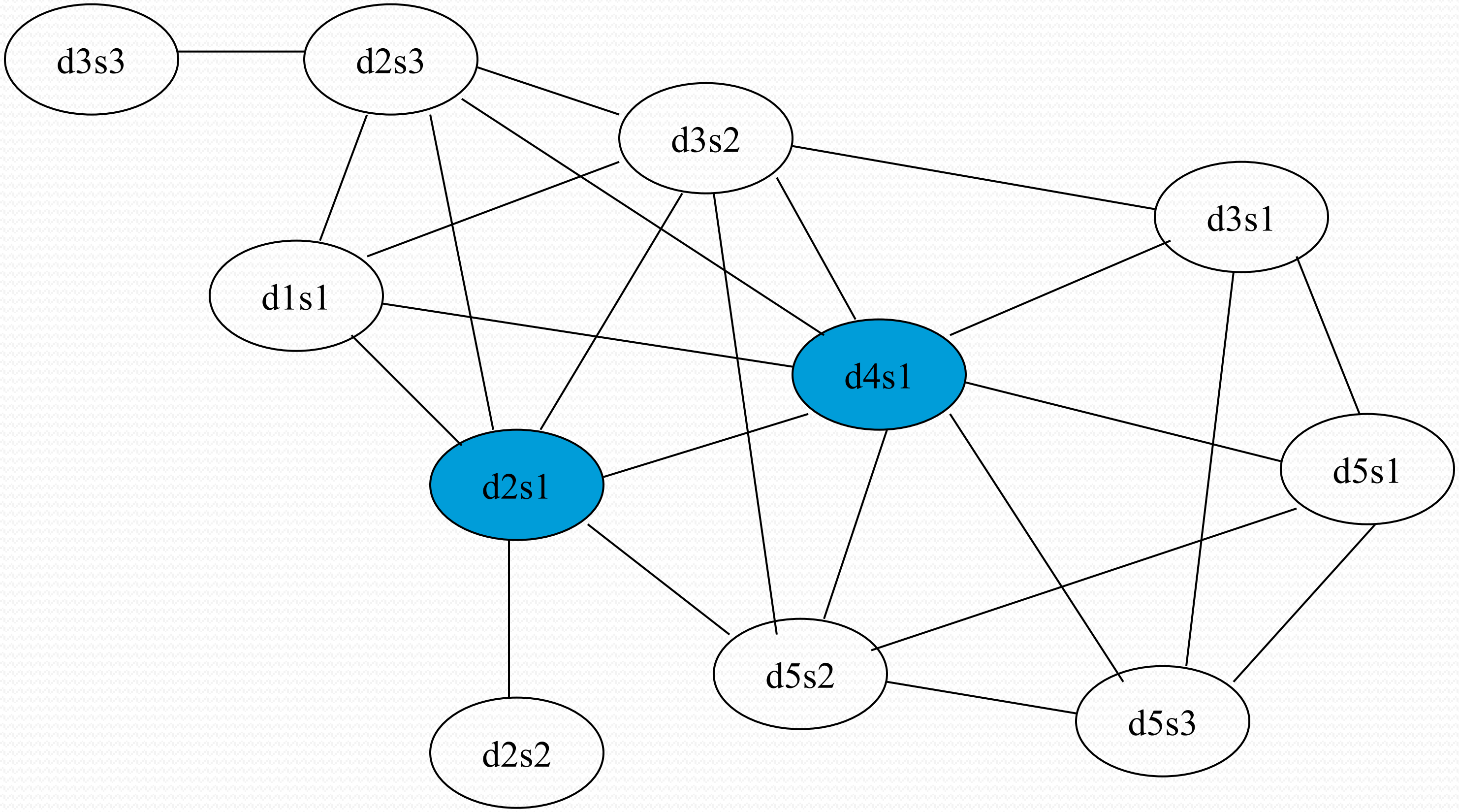
8 The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.

9 British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq ``did not end" and that Britain is still ``ready, prepared, and able to strike Iraq."

10 In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq ``will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.

11 A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

LexRank Example



Data

- We provide two sentence file(sample.sen, news.sen).
- The first is the small example in the previous page
- The second is sentences collected from two sport news(120 sentences)

Input Format

- Sentence file (*.sen) : each line is a sentence candidate.
- Just use space to separate each term
- IDF file (idf) : each line is “term” “idf-value”
- Note that if you can’t find a term in our idf file, let its idf value as : log(40252.0)

Output Format

- Sort each sentence by the LexRank score
- Output each sentencce with its sentence number and its score in the order decreasing score.
- 4 2.354
- 3 1.652
- 1 0.679
- … etc

Requirement(LexRank)

- Program Format:
- ./compute-lexrank sentence_path
- Run your LexRank program on news.sen and upload the result
- Threshold = 0.1
- We would provide answer for sample.sen

Report

- PageRank –
  - How do you implement the algorithm on sparse graph
  - What do you find in this task
- LexRank –
  - How do you implement LexRank
  - How do you think your output summary?
  - Any other things you try (What is the resulting graph and summary setting threshold = 0.2, 0.3 .. ?)

Upload Format

- Zip all your files with your student id as file name to ceiba
- R00922XXX.zip
- +---R00922XXX
- +---REPORT.pdf
- +--- R00922xxx.pagerank1 // CS stanford.edu (Oct. 2004)
- +--- R00922xxx.pagerank2 // stanford.edu (Aug. 2003)
- +--- R00922xxx.lexrank
- +---compilePR.sh //script file to compile PageRank
- +---compileLR.sh //script file to compile LexRank
- +---executePR.sh // script file to execute PageRank
- +---executeLR.sh // script file to execute LexRank
- All the other files and source code

Scoring

- 15% of 學期成績
- PageRank – 5%
- LexRank – 5%
- Report – 5%
- Deadline – 5/23 21:00

Contact

- Ptt2: WebMining
- E-mail:
- yishiang.Tzeng@gmail.com
- gn01812345@gmail.com
- pishen02@gmail.com
- Lab 302網路資訊檢索與探勘

Q&A