

open.michigan

Unless otherwise noted, the content of this course material is licensed under a Creative Commons Attribution 3.0 License.
<http://creativecommons.org/licenses/by/3.0/>

Copyright 2008, Lada Adamic

You assume all responsibility for use and potential liability associated with any use of the material. Material contains copyrighted content, used in accordance with U.S. law. Copyright holders of content included in this material should contact open.michigan@umich.edu with any questions, corrections, or clarifications regarding the use of content. The Regents of the University of Michigan do not license the use of third party content posted to this site unless such a license is specifically granted in connection with particular content objects. Users of content are responsible for their compliance with applicable law. Mention of specific products in this recording solely represents the opinion of the speaker and does not represent an endorsement by the University of Michigan. For more information about how to cite these materials visit <http://michigan.educommons.net/about/terms-of-use>.

 UNIVERSITY OF MICHIGAN



PROBLEM SET 7 – Week 10

1. LexRank - summarizing text

Select a piece of text (10-20 sentences) that you would like to summarize and paste it in the appropriate box of the LexRank demo

<http://tangra.si.umich.edu/clair/lexrank/>.

If you are unable to paste text into the text box (where currently there is some text about an Iraqi official), try a different browser. I've tested this on the DIAD PCs and it works there...

- There are two parameters you can vary: (1) the cosine similarity threshold determines how similar two sentences have to be in order to share an edge. (2) the salience threshold determines how high a sentence's PageRank has to be in order for that sentence to be included in the summary. Vary the cosine similarity threshold and record the most salient sentence. Does the most salient sentence change as you vary the threshold? Accordingly, report on a cosine similarity threshold that gave you the best result (if applicable).
- Compare the 1 sentence summary to the 2 or 3-sentence summary. In your opinion, how much do the 2nd and 3rd sentences add (in terms of adding more information). Would you have chosen them, or a different sentence? Relate your answer to the structure of the lexical similarity graph.

2. PageRank

Construct a small directed network (about 10 nodes) in GDF or .net format and load it into GUESS. Construct it such that you have at least one node that will have low indegree but high PageRank

- Compute the PageRank of each node by typing

```
g.nodes.pagerank
```

- Color by PageRank

```
colorize(pagerank,green,yellow)
```

- Compute the indegree

```
g.nodes.indegree
```

- Size the nodes by indegree

```
resizeLinear(indegree,minsize,maxsize) // (you are choosing minsize and  
maxsize)
```

Turn in an image of your network (**). Point out a node that has high PageRank but low indegree. Explain qualitatively how this came about.

an aside: You can also use the GUESS toolbar pageranktoolW.py, if you'd like to see how the algorithm converges...