NLP

Introduction to NLP

Summarization Techniques 3/3

Conroy and O'Leary (2001)

- Using Hidden Markov Models
- Takes into account the local dependencies between sentences
- Features
  - Position, number of terms, similarity to document terms
- HMM alternates between summary and non-summary states

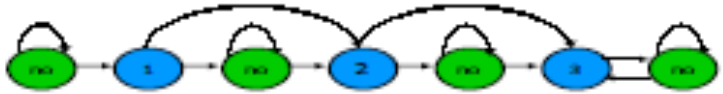


Figure 1: Summary Extraction Markov Model to Extract 2 Lead Sentences and Additional Supporting Sentences

Osborne (2002)

- Don't assume feature independence
- Use maxent (log-linear) models
- Better than Naïve Bayes
- Features
  - Sentence length
  - Sentence position
  - Inside introduction
  - Inside conclusion

Lexrank (Erkan and Radev 2004)

- Single and multi-document summarization
- Lexical Centrality
  - Represent text as graph
  - Graph centrality
  - Graph clustering
  - Random walks

Lexrank

1 (d1s1) 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 (d2s1) 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 (d2s2) 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 (d2s3) 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 (d3s1) 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 (d3s2) 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 (d3s3) Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).

8 (d4s1) 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 (d5s1) 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 (d5s2) 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 (d5s3) 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.

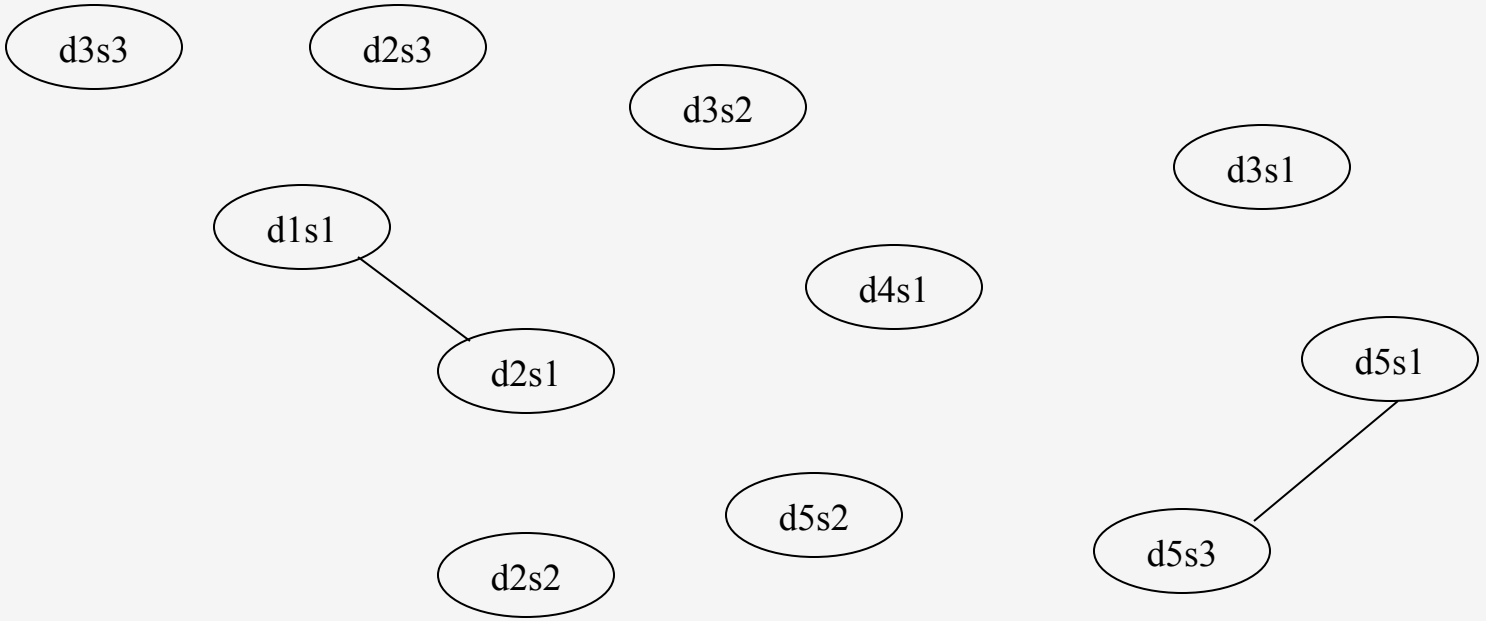
Lexrank

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

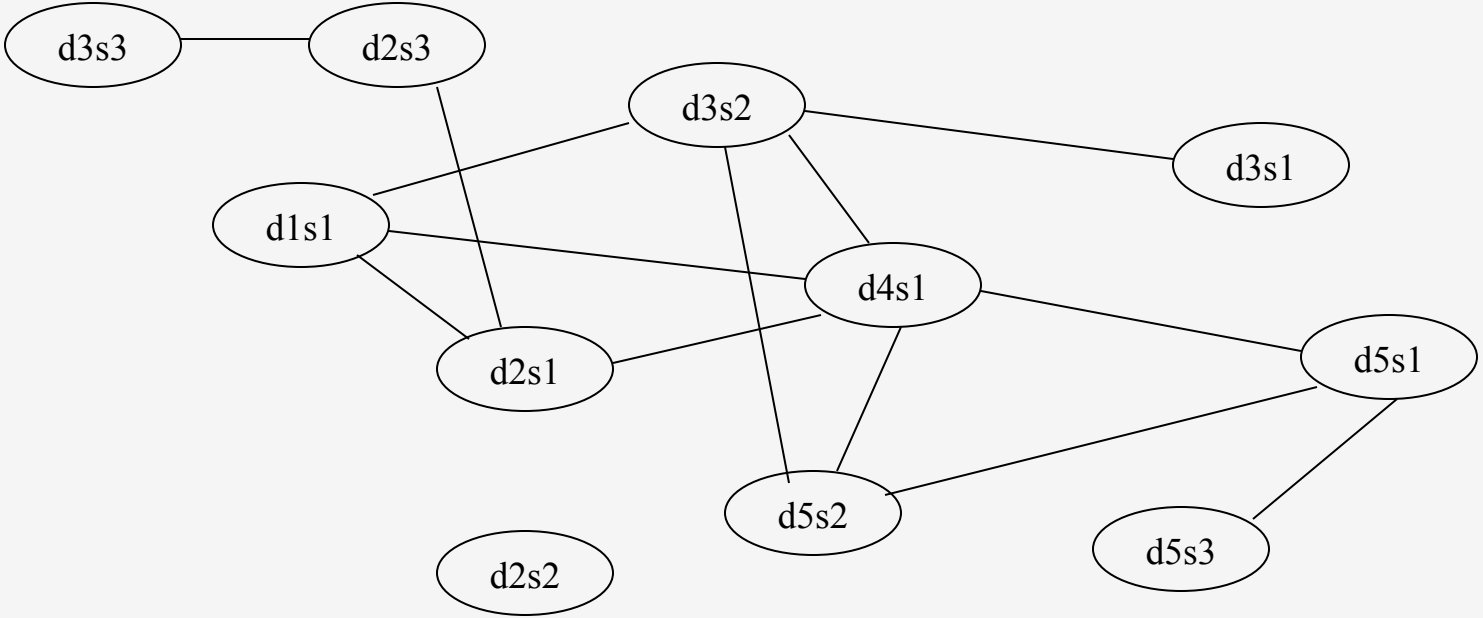
Lexrank

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

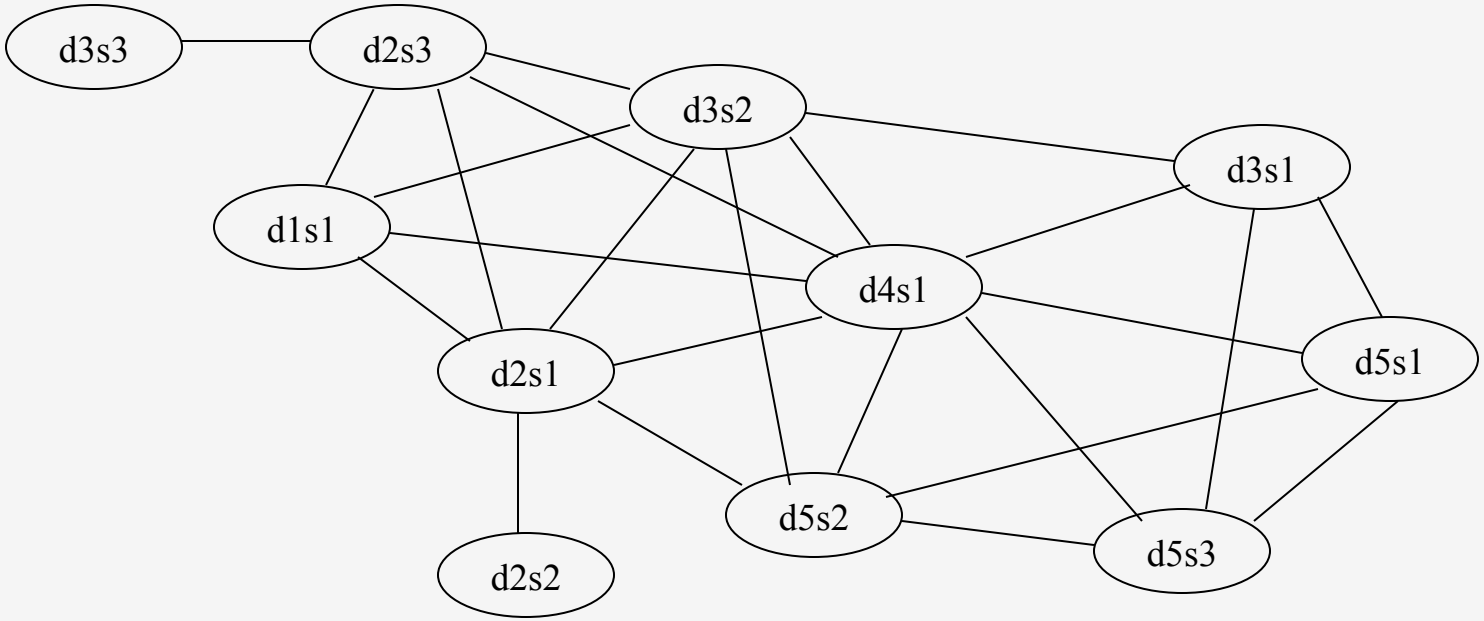
Cosine Centrality (t=0.3)



Cosine Centrality (t=0.2)

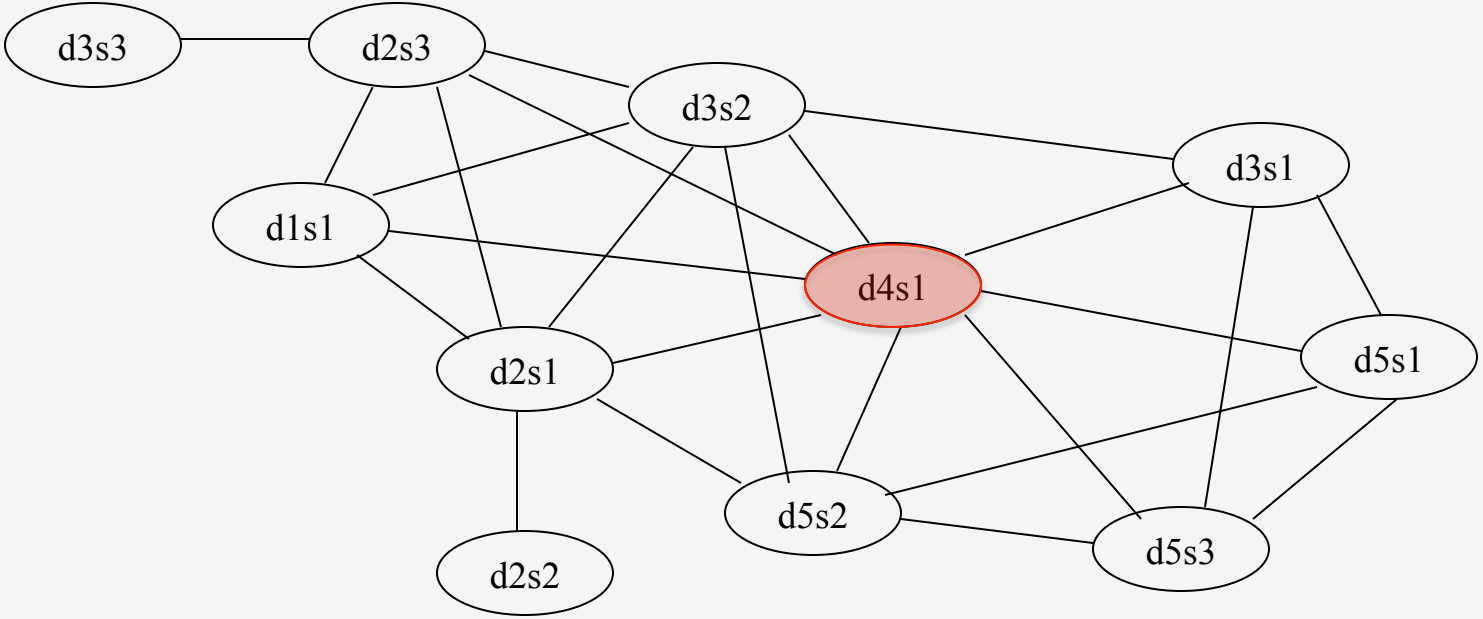


Cosine Centrality (t=0.1)



Sentences vote for the most central sentence!

Cosine Centrality (t=0.1)



Sentences vote for the most central sentence!

Lexrank (Advanced Material)

- Square connectivity matrix
- Directed vs. undirected
- An eigenvalue for a square matrix A is a scalar λ such that there exists a vector $x \neq 0$ such that $Ax = \lambda x$
- The normalized eigenvector associated with the largest λ is called the principal eigenvector of A
- A matrix is called a stochastic matrix when the sum of entries in each row sum to 1 and none is negative. All stochastic matrices have a principal eigenvector

Lexrank (Advanced Material)

- The connectivity matrix used in PageRank [Page & al. 1998] is irreducible [Langville & Meyer 2003]
- An iterative method (power method) can be used to compute the principal eigenvector
- That eigenvector corresponds to the stationary value of the Markov stochastic process described by the connectivity matrix
- This is also equivalent to performing a random walk on the matrix

Lexrank (Advanced Material)

- The stationary value of the Markov stochastic matrix can be computed using an iterative power method:

$$p = E^T p$$

$$(I - E^T)p = 0$$

- PageRank adds an extra twist to deal with dead-end pages. With a probability $1-\varepsilon$, a random starting point is chosen. This has a natural interpretation in the case of Web page ranking

$$p(v) = \frac{1-\varepsilon}{n} + \varepsilon \sum_{u \in pr[v]} \frac{p(v)}{|su[u]|}$$

su = successor nodes
pr = predecessor nodes

- Eigenvector centrality: the paths in the random walk are weighted by the centrality of the nodes that the path connects

Gong and Liu (2001)

- Using Latent Semantic Analysis (LSA)
- Single and multi-document
- Not using WordNet
- Each document is represented as a word by sentence matrix (row=word, column=sentence)
- TF*IDF weights in the matrix
- SVD: $A = USV^T$
- The rows of V^T are independent topics
- Select sentences that cover these independent topics

NLP