CADIP 2000

Document Ontology Extractor

Sadanand Srivastava,
James Gil de Lamadrid,
Yuriy Karakashyan

Department of Computer Science

Bowie State University

I. INTRODUCTION

It is very important to build document ontology to analyze a large collection of documents: to see what topics are discussed (or just mentioned) in this collection and how to retrieve related documents.

Contemporary statistical methods of information retrieval such as Boolean, Extended Boolean, Vector Space and others can be compared [Greengrass, 1997].

By classical Boolean approach one finds documents containing the exact words occurring in a query. It does not use the term weights, other than 0 (term is absent) and 1 (term is present). Hence, it can extract the documents not related to the topic of query only because some words from query were mentioned in the document (maybe, in another context) and will never extract documents closely related to the query but described in synonym terms.

Extended Boolean approach makes use of the weights assigned to the terms in each document in a range from 0 to 1 but still does not allow to extract related documents not containing the exact query terms.

The vector space method describes each document as a set of terms. This set defines the document space such that each distinct term represents one dimension in that space [Salton, 1983; 1989]. Each document in document space is defined by the weights of the terms that represent it. Different ways of normalization for these weights are applied [Lee, 1995], including document-length normalization method [Singhal, 1996]. The drawback of this approach is a big number of meaningful terms which are not independent even if the words undergo pre-processing, e.g., the elimination of the noise words (*and, why, but*), stemming - to extract the root of each word and some other procedures.

Another vector space approach called Latent Semantic Indexing (LSI) [Derwester, 1990] attempts to catch term to term statistical references by replacing the document space with lower-dimensional concept space. If document D_1 uses the term t_A and document D_2 uses the equivalent term t_B, LSI will effectively recognize this equivalence if t_A and t_B co-occur frequently in the similar contexts in other documents [Berry, 1995]. LSI accomplishes this by using a method of matrix decomposition called Singular Value Decomposition (SVD). The effectiveness of SVD as compared to other techniques is described in [Harman, 1995].

LSI has some disadvantages such as requirement to store substantially large set of values, requirement to store real numbers instead of integers and computationally expensive solution for the large collections [Hull, 1994]. But, on the other hand, this solution needs to be constructed only once for entire collection.

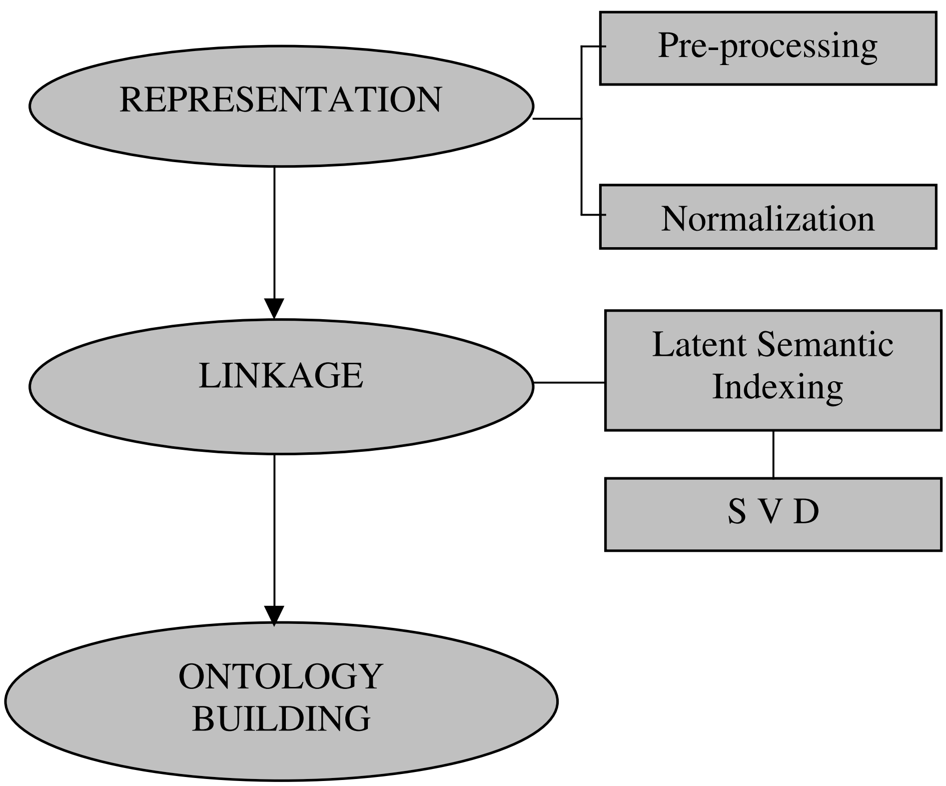
One more popular approach is n-grams method which is similar to LSI method. In that approach instead of words the dimensions of the document space can be represented by n-grams , strings of n consecutive characters extracted from the text without regard to word length and often without regard to word boundaries. Analysis of the collection of documents is based on SVD technique [Nicholas, 1998].

II. GOALS

The goal of Document Ontology (DO) research is to construct a system capable of reading a standard text file document, performing semantic analysis on the document and generating a useful ontology. This process should be automated as much as possible. Methods used for extracting ontological information might include statistical methods and user feedback. (Algorithms developed in SVDPACKC [Berry, 1993] calculate SVD solutions with C program. However, our project requires the codes to be written in the Java language).

To accomplish this goal we should perform the following tasks:

- Represent textual document as a set of meaningful terms. This is a form of keyword indexing. Then we can assign a numeric weight to each term, representing the usefulness of the given term as a descriptor of the given document (Fig.1).



- Associate and link terms in the document into a meaningful ontology. Building ontology is a complicated task, and to automate this process requires that several difficult problems should be solved. These problems include the deciding the criteria used to link terms, and types of term links, discerning significant terms from idiom structures and providing meaningful labels for abstract concepts. Integrate the components of the system into a robust and easily used tool.

- Test the system by running it on input documents.
- Debug and tune the system

Using a news article of *Washington Post* as an input textual file (Appendix 1) we tested the effectiveness of pre-processing procedures.

III. PRE-PROCESSING

Our basic program counts the frequencies of the terms in textual document. To extract meaningful terms from a text some pre-processing should be performed while reading the input document.

1. Numbers and punctuation marks standing alone should be cut; for example, 55 , 720 , 90 , 4 , -

2. It is necessary to cut punctuation marks at the end of the words; e.g., the program should not distinguish the words school and school, theaters and theaters , from each other.

3. The Program shouldn t be sensitive to the case and all the words should be examined as low-case words; e.g., if the text contains words school and School , the frequency of the word school in the output of the program should be two.

4. It is important to exclude stop words — articles, preposition, unions etc. — which cannot be used as descriptors of any document (words like and , the , when , though , for etc.). To achieve that goal, we prepare a stop list of such noise words (now approximately 250 words) in additional file. The program construct s a hash-table based on that list so that the current token in input file is compared with the content of this hash-table.

5. One more aspect of pre-processing is correct counting of the words in specific grammatical forms (irregular verbs, nouns of Latin origin etc.) We prepared one more auxiliary file which produced another hash-table to replace derivative forms with original ones; e.g.

broke	-	break
broken	-	break
phenomena	-	phenomenon

   By using this procedure for the text containing words break and broken the program gives a frequency of two for the word break

6. And finally the last procedure of the pre-processing stage — stemming . The objective is to eliminate the variation that arises from the occurrence of different grammatical forms of the same word (e.g., president - presidential , toured - tour , worked - workers etc.). This was done by using the array of words endings (more than 20 of

the most common endings); current token is checked for the presence of specified endings - if it has one, then this ending is cut and variations of the root with other endings are checked.

The output of these pre-processing procedures for the test document is shown in Table 1.

IV. NORMALIZATION

Our next task was to calculate the normalized weight of each word by using obtained term frequencies. We used one of the traditional methods and calculated the statistical weight of each term as a relation of word frequency to the total number of meaningful words in the document:

$$\omega_i = freq_i / (\Sigma\, freq_k), \ \ k = 1,2\ n$$

It is necessary to take into account the document length normalization component to avoid different weight for the terms with the same frequency occurring in documents of different length. The document length normalization component is:

$$\{1/ (\Sigma\, \omega_k^{2})\}^{1/2}, \qquad k = 1,2\ n$$

Normalized output for the words occurring in the tested document more than twice are shown in Table 2.

V. TERM-DOCUMENT MATRIX

Now we can represent a textual document as a set of meaningful normalized terms. These procedures allow us to create term-document matrix which describes the occurrence of meaningful terms in each document of the collection. To create such a matrix we wrote the class keeping track on all the meaningful terms occurred in the documents, and each time when the next document is read, new terms are added into matrix as rows, their frequencies are added as a new column. While reading the program checks whether these new terms are grammatical forms of some previous terms (stemming).

Appendix 2 contains four articles from *Washington Post* (March 18, 2000) used as input documents. The resulting output matrix A shown in the Table 3.

VI. LATENT SEMANTIC INDEXING (LSI)

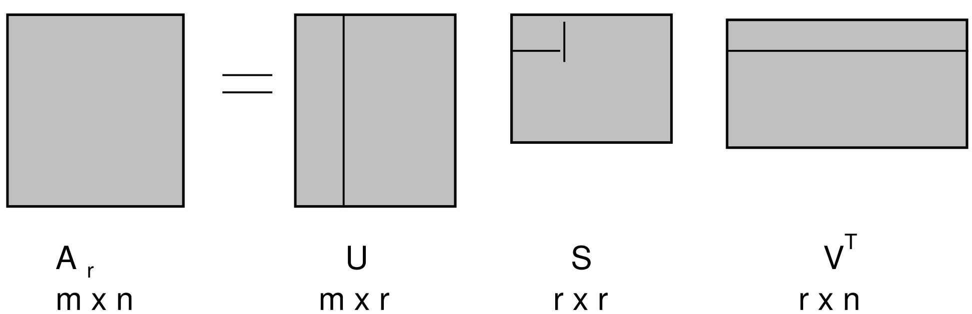
Using terms as the descriptors of a document has certain drawbacks. Its major limitation is that it assumes that the terms are independent. Hence, the relationships among terms are ignored, e.g., the fact that some terms are likely to co-occur in documents about a given topic.

LSI is a statistical method linking terms into useful semantic structure. By using this method each document is represented not by terms but by concepts that are truly statistically independent in a way that terms are not. Hence, LSI concepts are information rich [Hull, 1994]. LSI attempts to capture some of these semantic term dependencies using a pure statistical and automatic method, i.e., without syntactic or semantic natural language analysis and without manual human intervention.
LSI is based on singular value decomposition (SVD) technique.

SVD computes three matrices **U, S** and **V** such that the **m** x **n** term-document matrix **A**, of rank **r**, can be expressed as their product:

$$\boldsymbol{A = U * S * V^T}$$

where, **U** is an **m** x **m** orthogonal matrix; **S** is the **m** x **n** diagonal matrix in which the singular values of **A** are listed in descending order; and **V** is an **n** x **n** orthogonal matrix. The number of non-zero singular values is the same as the rank of the original matrix. The rank of **A** is **r**, so



Matrix **U** represents each term used in a collection (total number of terms is **m** which is the number of rows in **U**) with **r** concepts (number of columns in **U**). **V** is a document matrix; each column of **V** is one of the **r-**derived concepts, each row represents one document. Matrix **S** is the diagonal matrix formed by the singular values of **A** in descending order (singular values of **A** are the non-negative square roots the eigenvalues of $\mathbf{AA^T}$ or $\mathbf{A^TA}$).

To carry out necessary computations we developed different classes for matrix and polynomial analysis in Java.

VII. CONCLUSION

Now we have the following results:

(a) pre-processed each document of the collection:
- constructed the set of meaningful terms and set of corresponding frequencies of their occurrence takes place while reading each document of the collection;
- excluded stop words, (different grammatical forms are transformed into one);
- normalized each meaningful term using statistical weight and document-length

(b) created a term-document matrix for the collection based on (a):
- in the case of a collection of **n** documents with **m** meaningful terms we can construct output as a term-document matrix **A** with **m** rows and **n** columns; each row in this matrix represents a term, and each column represents a document;
- the special feature of this matrix **A** construction is a necessity to use stemming twice: while preparing a table for each document and when creating the final matrix by comparing terms from different documents.

(c) decomposed the matrix (b) into triples :

$$\mathbf{A} = \sum_{j=1}^{r} \mathbf{S}_{j,j}\, \mathbf{u}_j\, \mathbf{v}_j^T ,$$

where $\mathbf{S}_{j,j}$ are singular values of **A**, $\mathbf{u}_j$ and $\mathbf{v}_j^T$ are **j**-th columns of **U** and **V**, respectively, **r** is a rank of **S.**

VIII. OUR WEBPAGES

http://www.sem.bowiestate.edu/projects/doe/
http://students.sem.bowiestate.edu/~ykarakas/CADIP2000/cadip2000.html

IX. FUTURE WORK

The final and the most complicated part of our research - to build ontology for the collection of electronic documents.
To accomplish it, we need to perform the following tasks:
- build an ontology from existing decomposition of term-document matrix;
- minimize user s feedback;
- design algorithms for updating ontologies without re-computation.

X. REFERENCES

1. Berry,M.W., Do,T., O'Brien, Krishna,V., Varadhan,S., SVDPACKC: Version 1.0 User's Guide, Tech.Report CS-93-194, University of Tennessee, Knoxville, TN, October 1993.
2. Berry,M.W., Dumais,S.T., O'Brien,G.W. Using Linear Algebra for Intelligent Information Retrieval, SIAM Review, Vol.37, No.4, pp.573-595, December 1995;
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman,R., Indexing by Latent Semantic Analysis, Journal of the American Society For Information Science, 41(6), pp.391-407, 1990;
4. Greengrass E., Information Retrieval: An Overview, R521, 1997;
5. Harman,D. Overview of the Fourth Text REtrieval Conference (TREC-4), National Institute of Standards and Technology, 1995;
6. Hull,D. Improving Text Retrieval for the Routine Problem Using Latent Semantic Indexing, Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.282-291, 1994;
7. Lee,J.H. Combining Multiple Evidence from Different Properties of Weighting Schemes, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.180-188, 1995;
8. Nicholas,C., Dahlberg,R. Spotting Topics with the Singular Value Decomposition, from Principles of Digital Document Processing, St.Malo, Fiara, March 1998;
9. Salton,G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, MA, 1989;
10. Salton,G., McGill,M.J. Introduction to Modern Information Retrieval, McGraw Hill Publishing Company, New York, 1983;
11. Sighal,A., Buckley,C., Mandar,M. Pivoted Document Language Normalization, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.21-29, 1996;

Appendix 1.
Tested text for a single document pre-processing (news story from *Washington Post*, 08/01/1999)

LOS ANGELES (AP) - For three days, President Clinton has stressed the need to invest in the poorest areas of the country. But here, on economically deprived turf that is still trying to recover from deadly race riots, he is arguing for investing in the poorest people themselves.

The president was visiting the community of Watts, scarred by riots nearly 30 years apart, to make the case today for private firms to help disadvantaged young people gain the skills they need for high-tech jobs in the new millennium.

Among those lined up to accompany Clinton was Magic Johnson, the former Los Angeles Lakers star who has revitalized South Central L.A. and other inner cities with multiplex movie theaters.

Clinton was touring the Transportation Career Academy Program within Alain Leroy Locke High School, named for the first black Rhodes scholar. The facility helps prepare students for careers in transportation-related fields from urban planning to architecture. Since 1994, 1,800 students have participated in the academy, and 90 percent who graduate go on to college.

Later, Clinton was going to nearby Anaheim for the annual conference of the National Academy Foundation - where chief executives were huddling to discuss ways to connect employers with disadvantaged youth, especially those ages 16 to 24. The White House estimated that 10 million people in that age group are out of school, and 4 million of them lack a high school diploma.

"They have to pay attention to and care about the development of the work force," said deputy White House chief of staff Maria Echaveste. "They can't be competitive, they can't stay profitable, if they don't have a work force that is skilled and that is trained. At the conference, Clinton was to announce an 8 million initiative to help create "information academies" within inner-city and rural schools. The initiative is a partnership between the Department of Labor and companies such as AT&T, Lucent Technologies and Cisco Systems.

That announcement closes out Clinton's tour, but he will remain in Los Angeles through Saturday to watch the U.S. women's soccer team compete for the World Cup.

Today's visit to Los Angeles' south side takes Clinton back to an area he first canvassed as a presidential candidate in May 1992, just days after riots in the wake of police acquittals in the beating of motorist Rodney King left 55 people dead and 720 buildings destroyed or damaged by fire.

Part of the complaint then - as it was in 1965, when 34 people died in Watts rioting- was the need for better access to jobs and social investment to eliminate the economic isolation of the inner city.

The president will see a changed South L.A. After the 1992 riots, banks and federal redevelopment programs have made millions of dollars in loans and grants to local businesses, and many damaged areas have been rebuilt. And while residents welcome the government's help, there is a thriving spirit of free-enterprise and self-reliance.

The Baldwin Hills Crenshaw Plaza mall, boosted by a Magic Johnson movie theater and its proximity to affluent black neighborhoods, is booming with an occupancy rate of more than 90 percent. In other areas, retailers have opened new stores. Three banks, Washington Mutual, Wells Fargo and Hawthorne Savings, partnered with Operation Hope to open banking centers where residents can apply for loans and take classes on managing their finances.

The president flew to Los Angeles from Phoenix, where he toured the facilities of La Canasta, a successful food producer, to highlight the needs of the Latino community on that city's south side.

He strolled through the plant with owner Carmen Abril Lopez and watched as thousands of tortillas coursed past him on conveyor belts. While workers in white shirts and baseball caps removed the flawed ones, Clinton took a tortilla in his hands and inspected it, marveling at the fact that the plant produces and sells 840,000 tortillas each day.

"Our country has been really blessed by these good economic times," Clinton said. "But we know, as blessed as America has been, not every American has been blessed by this recovery. All you have to do is drive down the streets of South Phoenix to see that."

Appendix 2.
Tested collection of documents (news stories from *Washington Post*, 03/18/2000)

Document 1.

President Clinton today asked Congress to create an emergency heating oil reserve in the Northeast to blunt future shortages, but said there is "no overnight solution" to the current high oil prices.
Clinton said he was concerned about the hardships facing many Americans because of the soaring gasoline and heating oil costs. Still, he said in his weekly radio address, "we also need to take a longer view" to protect consumers and "strengthen America's energy security."
None of the president's proposals was aimed directly at alleviating the high prices and low supplies. Meantime, energy officials have suggested gasoline prices may climb even more this summer.

Document 2.

America Online Inc. announced yesterday that it would pay up to $8.25 billion to buy out its partner in AOL Europe and AOL Australia, German media conglomerate Bertelsmann AG.
The move raises questions about the fate of AOL's other international partnerships in light of the Internet frontrunner's pending marriage to Time Warner Inc.
Dulles-based AOL shared equal ownership of the European and Australian entities with Bertelsmann, a privately held company whose holdings include Random House and the Arista and RCA record labels.
Bertelsmann is one of the leading rivals of New York-based Time Warner.

Document 3.

High-tech millionaire Harry Hopper wanted a new kitchen in his 19th-century Old Town Alexandria home. But the plans he drew up for a 250-square-foot addition left some neighbors aghast.
The architectural style, one of them complained at a public hearing, was "wingding" at best. What's more, neighbor Charles Trozzo said, the addition "would loom over Duke Street like Gheorghe Muresan looms over regular-size people," referring to the 7-foot-7 pro basketball player.
Hopper's new wing, which the City Council eventually approved, is just one episode in an increasingly sour tiff that is dividing certain well-heeled precincts of Old Town, where new money and old houses are proving a combustible combination.

Document 4.

Mounting evidence suggests that increasing numbers of Americans are falling seriously ill or even dying after taking dietary supplements that promise everything from extra energy to sounder sleep.
The victims include men and women of all ages as well as children whose parents are feeding them snacks, drinks and nostrums made with herbal supplements that are neither regulated by the federal government nor tested for their effects on the young.
While the Food and Drug Administration issues periodic warnings about the dangers of individual supplements, no organization or agency has ever made a comprehensive analysis of the sickness and death associated with them.

Table 1.

Pre-processing of a single document (news article, Appendix 1). Results are given for the words occurring in the text more than once

3 - opened	2 - watts	2 - conference
3 - closed	2 - make	2 - chief
2 - prayer	2 - today	2 - age
4 - try	4 - help	3 - white
2 - constructor	2 - disadvantaged	2 - house
3 - carry	2 - skills	3 - work
2 - lake	3 - high	2 - force
2 - emotional	2 - jobs	2 - say
3 - relations	2 - new	2 - announce
3 - kind	2 - magic	2 - initiative
2 - develop	2 - johnson	2 - partnership
2 - associate	2 - star	2 - watch
2 - write	5 - south	2 - side
5 - los	3 - inner	3 - take
5 - angeles	3 - cities	2 - damaged
3 - day	2 - movie	2 - economic
4 - president	2 - theater	2 - see
9 - clinton	3 - tour	3 - banks
3 - invest	2 - transportation	2 - loans
2 - poorest	3 - care	2 - residents
4 - area	4 - academy	2 - open
2 - country	2 - program	2 - phoenix
2 - recover	4 - school	2 - producer
2 - dead	2 - black	2 - plant
5 - riots	2 - facility	3 - tortilla
5 - people	2 - students	3 - blessed
2 - visit	2 - percent	
2 - community	2 - go	

Table 2.
Normalized weights for words occurred in the tested document more than twice.

FREQUANCY	WORD	WEIGHT
5	los	0.179374250
5	angeles	0.179374250
3	day	0.107624546
4	president	0.143499400
9	clinton	0.322873650
3	invest	0.107624546
4	area	0.143499400
5	riots	0.179374250
5	people	0.179374250
4	help	0.143499400
3	high	0.107624546
5	south	0.179374250
3	inner	0.107624546
3	cities	0.107624546
3	tour	0.107624546
3	care	0.107624546
4	academy	0.143499400
4	school	0.143499400
3	white	0.107624546
4	million	0.143499400
3	work	0.107624546
3	take	0.107624546
3	banks	0.107624546
3	tortilla	0.107624546
3	blessed	0.107624546

Table 3.
Output matrix for the documents collection (news stories, Appendix 2).

Document 1	Document 2	Document 3	Document 4	Word
0.20739026	0.00000000	0.00000000	0.00000000	president
0.20739026	0.00000000	0.00000000	0.00000000	clinton
0.10369513	0.00000000	0.00000000	0.00000000	today
0.10369513	0.00000000	0.00000000	0.00000000	asked
0.10369513	0.00000000	0.00000000	0.00000000	congress
0.10369513	0.00000000	0.00000000	0.00000000	create
0.10369513	0.00000000	0.00000000	0.00000000	emergency
0.20739026	0.00000000	0.00000000	0.00000000	heating
0.3110854	0.00000000	0.00000000	0.00000000	oil
0.10369513	0.00000000	0.00000000	0.00000000	reserve
0.10369513	0.00000000	0.00000000	0.00000000	northeast
0.10369513	0.00000000	0.00000000	0.00000000	blunt
0.10369513	0.00000000	0.00000000	0.00000000	future
0.10369513	0.00000000	0.00000000	0.00000000	shortages
0.3110854	0.00000000	0.10540925	0.00000000	say
0.10369513	0.00000000	0.00000000	0.00000000	overnight
0.10369513	0.00000000	0.00000000	0.00000000	solution
0.10369513	0.00000000	0.00000000	0.00000000	current
0.20739026	0.00000000	0.00000000	0.00000000	high
0.3110854	0.00000000	0.00000000	0.00000000	prices
0.10369513	0.00000000	0.00000000	0.00000000	concerned
0.10369513	0.00000000	0.00000000	0.00000000	hardships
0.10369513	0.00000000	0.00000000	0.00000000	facing
0.10369513	0.00000000	0.00000000	0.121267796	americans
0.10369513	0.00000000	0.00000000	0.00000000	soaring
0.20739026	0.00000000	0.00000000	0.00000000	gasoline
0.10369513	0.00000000	0.00000000	0.00000000	costs
0.10369513	0.00000000	0.00000000	0.00000000	weekly
0.10369513	0.00000000	0.00000000	0.00000000	radio
0.10369513	0.00000000	0.00000000	0.00000000	address
0.10369513	0.00000000	0.00000000	0.00000000	take
0.10369513	0.00000000	0.00000000	0.00000000	longer
0.10369513	0.00000000	0.00000000	0.00000000	view
0.10369513	0.00000000	0.00000000	0.00000000	protect
0.10369513	0.00000000	0.00000000	0.00000000	consumers
0.10369513	0.00000000	0.00000000	0.00000000	strengthen
0.10369513	0.106600344	0.00000000	0.00000000	america
0.20739026	0.00000000	0.00000000	0.121267796	energy
0.10369513	0.00000000	0.00000000	0.00000000	security
0.10369513	0.00000000	0.00000000	0.00000000	proposals
0.10369513	0.00000000	0.00000000	0.00000000	aimed
0.10369513	0.00000000	0.00000000	0.00000000	directly
0.10369513	0.00000000	0.00000000	0.00000000	alleviating
0.10369513	0.00000000	0.00000000	0.00000000	low
0.10369513	0.00000000	0.00000000	0.00000000	supplies
0.10369513	0.00000000	0.00000000	0.00000000	meantime
0.10369513	0.00000000	0.00000000	0.00000000	officials
0.10369513	0.00000000	0.00000000	0.121267796	suggested
0.10369513	0.00000000	0.00000000	0.00000000	climb
0.10369513	0.00000000	0.00000000	0.24253559	even
0.10369513	0.00000000	0.00000000	0.00000000	summer
0.00000000	0.106600344	0.00000000	0.00000000	online
0.00000000	0.21320069	0.00000000	0.00000000	inc

0.00000000	0.106600344	0.00000000	0.00000000	announced
0.00000000	0.106600344	0.00000000	0.00000000	yesterday
0.00000000	0.106600344	0.00000000	0.00000000	pay
0.00000000	0.106600344	0.00000000	0.00000000	billion
0.00000000	0.106600344	0.00000000	0.00000000	buy
0.00000000	0.106600344	0.00000000	0.00000000	partner
0.00000000	0.42640138	0.00000000	0.00000000	aol
0.00000000	0.106600344	0.00000000	0.00000000	europe
0.00000000	0.21320069	0.00000000	0.00000000	australia
0.00000000	0.106600344	0.00000000	0.00000000	german
0.00000000	0.106600344	0.00000000	0.00000000	media
0.00000000	0.106600344	0.00000000	0.00000000	conglomerate
0.00000000	0.31980103	0.00000000	0.00000000	bertelsmann
0.00000000	0.106600344	0.00000000	0.121267796	ag
0.00000000	0.106600344	0.00000000	0.00000000	move
0.00000000	0.106600344	0.00000000	0.00000000	raises
0.00000000	0.106600344	0.00000000	0.00000000	questions
0.00000000	0.106600344	0.00000000	0.00000000	fate
0.00000000	0.106600344	0.00000000	0.00000000	international
0.00000000	0.106600344	0.00000000	0.00000000	partnerships
0.00000000	0.106600344	0.00000000	0.00000000	light
0.00000000	0.106600344	0.00000000	0.00000000	internet
0.00000000	0.106600344	0.00000000	0.00000000	frontrunner
0.00000000	0.106600344	0.00000000	0.00000000	pending
0.00000000	0.106600344	0.00000000	0.00000000	marriage
0.00000000	0.21320069	0.00000000	0.00000000	time
0.00000000	0.21320069	0.00000000	0.00000000	warner
0.00000000	0.106600344	0.00000000	0.00000000	dulles
0.00000000	0.106600344	0.00000000	0.00000000	shared
0.00000000	0.106600344	0.00000000	0.00000000	equal
0.00000000	0.106600344	0.00000000	0.00000000	ownership
0.00000000	0.106600344	0.00000000	0.00000000	european
0.00000000	0.106600344	0.00000000	0.00000000	entities
0.00000000	0.106600344	0.00000000	0.00000000	privately
0.00000000	0.21320069	0.00000000	0.00000000	hold
0.00000000	0.106600344	0.00000000	0.00000000	company
0.00000000	0.106600344	0.00000000	0.121267796	include
0.00000000	0.106600344	0.00000000	0.00000000	random
0.00000000	0.106600344	0.10540925	0.00000000	house
0.00000000	0.106600344	0.00000000	0.00000000	arista
0.00000000	0.106600344	0.00000000	0.00000000	rca
0.00000000	0.106600344	0.00000000	0.00000000	record
0.00000000	0.106600344	0.00000000	0.00000000	labels
0.00000000	0.106600344	0.00000000	0.00000000	leading
0.00000000	0.106600344	0.00000000	0.00000000	rivals
0.00000000	0.106600344	0.31622776	0.00000000	new
0.00000000	0.106600344	0.00000000	0.00000000	york
0.00000000	0.00000000	0.10540925	0.00000000	high-tech
0.00000000	0.00000000	0.10540925	0.00000000	millionaire
0.00000000	0.00000000	0.10540925	0.00000000	harry
0.00000000	0.00000000	0.2108185	0.00000000	hopper
0.00000000	0.00000000	0.10540925	0.00000000	wanted
0.00000000	0.00000000	0.10540925	0.00000000	kitchen
0.00000000	0.00000000	0.31622776	0.00000000	old
0.00000000	0.00000000	0.2108185	0.00000000	town
0.00000000	0.00000000	0.10540925	0.00000000	alexandria
0.00000000	0.00000000	0.10540925	0.00000000	home
0.00000000	0.00000000	0.10540925	0.00000000	plans
0.00000000	0.00000000	0.10540925	0.00000000	drew
0.00000000	0.00000000	0.10540925	0.00000000	square
0.00000000	0.00000000	0.2108185	0.00000000	addition
0.00000000	0.00000000	0.10540925	0.00000000	left

0.00000000	0.00000000	0.2108185	0.00000000	neighbor
0.00000000	0.00000000	0.10540925	0.00000000	aghast
0.00000000	0.00000000	0.10540925	0.00000000	architectural
0.00000000	0.00000000	0.10540925	0.00000000	style
0.00000000	0.00000000	0.10540925	0.00000000	complained
0.00000000	0.00000000	0.10540925	0.00000000	public
0.00000000	0.00000000	0.10540925	0.00000000	hearing
0.00000000	0.00000000	0.10540925	0.00000000	wingding
0.00000000	0.00000000	0.10540925	0.00000000	best
0.00000000	0.00000000	0.10540925	0.00000000	charles
0.00000000	0.00000000	0.10540925	0.00000000	trozzo
0.00000000	0.00000000	0.2108185	0.00000000	loom
0.00000000	0.00000000	0.10540925	0.00000000	duke
0.00000000	0.00000000	0.10540925	0.00000000	street
0.00000000	0.00000000	0.10540925	0.00000000	like
0.00000000	0.00000000	0.10540925	0.00000000	gheorghe
0.00000000	0.00000000	0.10540925	0.00000000	muresan
0.00000000	0.00000000	0.10540925	0.00000000	regular
0.00000000	0.00000000	0.10540925	0.00000000	people
0.00000000	0.00000000	0.10540925	0.00000000	referring
0.00000000	0.00000000	0.10540925	0.00000000	foot
0.00000000	0.00000000	0.10540925	0.00000000	pro
0.00000000	0.00000000	0.10540925	0.00000000	basketball
0.00000000	0.00000000	0.10540925	0.00000000	player
0.00000000	0.00000000	0.10540925	0.00000000	wing
0.00000000	0.00000000	0.10540925	0.00000000	city
0.00000000	0.00000000	0.10540925	0.00000000	council
0.00000000	0.00000000	0.10540925	0.00000000	eventually
0.00000000	0.00000000	0.10540925	0.00000000	approved
0.00000000	0.00000000	0.10540925	0.00000000	episode
0.00000000	0.00000000	0.10540925	0.00000000	increasingly
0.00000000	0.00000000	0.10540925	0.00000000	sour
0.00000000	0.00000000	0.10540925	0.00000000	tiff
0.00000000	0.00000000	0.10540925	0.00000000	dividing
0.00000000	0.00000000	0.10540925	0.00000000	certain
0.00000000	0.00000000	0.10540925	0.121267796	well
0.00000000	0.00000000	0.10540925	0.00000000	precincts
0.00000000	0.00000000	0.10540925	0.00000000	money
0.00000000	0.00000000	0.10540925	0.00000000	proving
0.00000000	0.00000000	0.10540925	0.00000000	combustible
0.00000000	0.00000000	0.10540925	0.00000000	combination
0.00000000	0.00000000	0.00000000	0.121267796	mounting
0.00000000	0.00000000	0.00000000	0.121267796	evidence
0.00000000	0.00000000	0.00000000	0.121267796	promise
0.00000000	0.00000000	0.00000000	0.121267796	extra
0.00000000	0.00000000	0.00000000	0.121267796	sleep
0.00000000	0.00000000	0.00000000	0.24253559	make
0.00000000	0.00000000	0.00000000	0.121267796	young
0.00000000	0.00000000	0.00000000	0.121267796	drug
0.00000000	0.00000000	0.00000000	0.121267796	periodic
0.00000000	0.00000000	0.00000000	0.121267796	comprehensive
0.00000000	0.00000000	0.00000000	0.121267796	death