



Cross-Language Ideology Detection Using Linguistics-Based Knowledge Extraction

Department 06641
September 13, 2006

Peter Chew, Stephen Verzi, JT McClain, Travis Bauer
Sandia National Laboratories



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.





Background

Words, or language, persuade people to join movements which can in turn become threats.



threats



movements



ideas



words



Our approach

- Representative sample of web pages, with focus towards blogs, commentaries
 - Multi-lingual (English, Russian, Arabic, French, Spanish)
- Collect in a SQL Server database
- Analyze, based on language and/or links, the prominence of different clusters
 - We expect approach to be ‘blind’ to language
- Analyze changes over time



What we are aiming for – an example

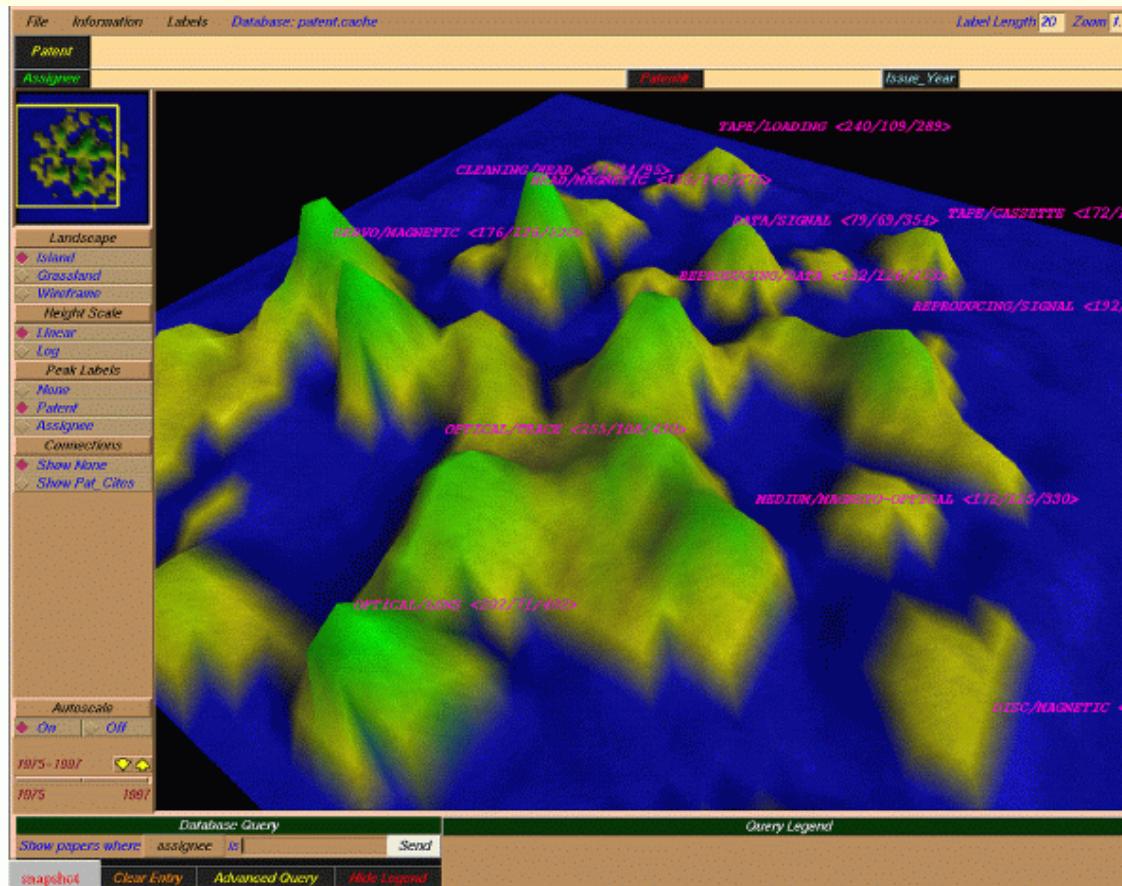
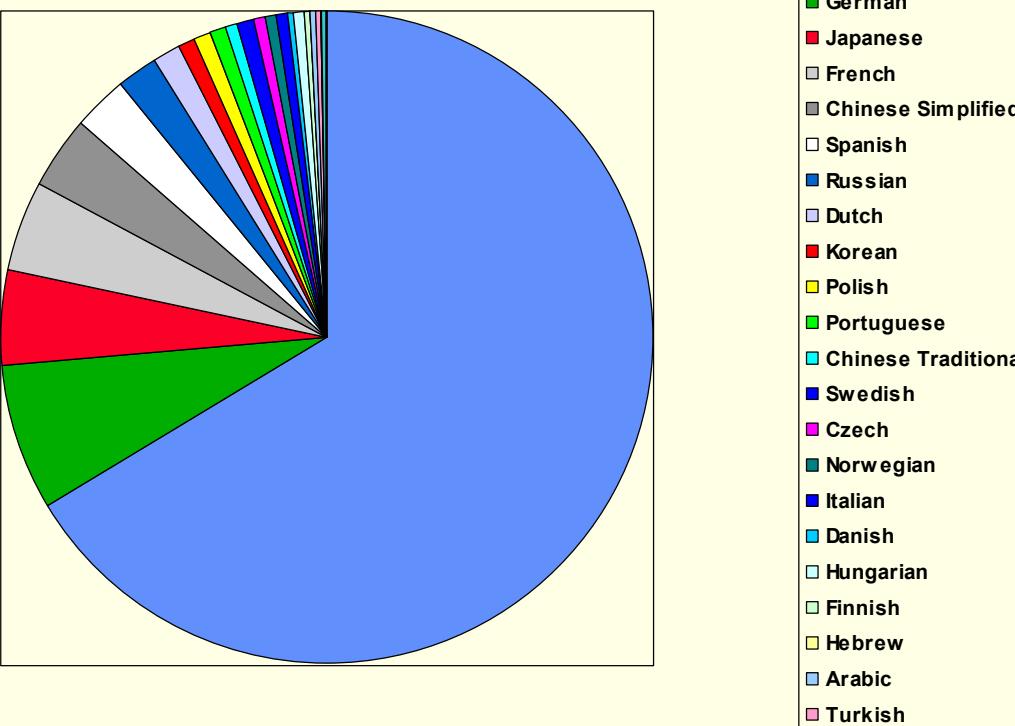


Figure 2: The mountain terrain metaphor provides an intuitive exploration environment.



Distribution of internet content by language





Data collection to date

	Number of documents	% in sample	approx. % in population
English	176,307	50.7	66.5
Russian	62,954	18.1	2.0
Arabic	30,248	8.7	0.2
Spanish	39,898	11.5	2.6
French 	38,479	11.1	4.4
Others	-	-	24.3
TOTAL	347,886	100.0	100.0



Approaches to cross-language information retrieval

- Translate the query
 - Efficacy is constrained by quality of machine translation
- Train algorithm on parallel corpora
 - Translations should:
 - Be available in target languages
 - Be reliable
 - Be sufficiently large in size
 - Cover target subject domain
 - Be free of undue copyright restrictions
 - Be electronically available
 - Be alignable



```
health of animals act ===> loi
health of the economy ===> santé
heavy burden ===> lourd fardeau
heavy equipment ===> matériel lourd
heritage day ===> jour du patrimoine
higher premiums ===> cotisation
highest bidder ===> plus offrant
historic document ===> document
historic event ===> événement
house management committee ===>
house of commons standing commit
```



The Bible as a Parallel Corpus

- **Resnik, Olsen & Diab (1999)** showed that the Bible fulfills all of these criteria and is surprisingly suitable as a parallel corpus
 - Translations in > 2,400 languages and rising
 - Great care taken over translations
 - Respectably large compared to other corpora
 - Covers many modern genres
 - Covers up to 85% of modern vocabulary
 - Generally free of copyright restrictions
 - Electronically available
 - Alignable



Language coverage - detail

A STATISTICAL SUMMARY OF LANGUAGES WITH THE SCRIPTURES

A summary, by geographical area and type of publication, of the number of different languages and dialects in which publication of at least one book of the Bible had been registered as of December 31, 2005.

Continent/Region	Portions	Testaments	Bibles	Bibles, DC*	Total
Africa	223	301	159	(29)	683
Asia	218	244	131	(28)	593
Australia/New Zealand/ Pacific Islands	148	234	38	(9)	420
Europe	114	36	61	(47)	211
North America	39	30	7	(0)	76
Caribbean Islands / Central America / Mexico/South America	118	270	29	(9)	417
Constructed Languages	2	0	1	(0)	3
TOTALS	862	1,115	426	(122)	2,403

* This column is a sub-section of the Bibles column – for example, there is a translation of the Deuterocanon for 47 of the 61 languages of Europe in which the Bible has been translated.

[A few corrections were made to our language databases and are reflected in this statistical summary]

Per <http://www.biblesociety.org/latestnews/latest341-slr2005stats.html>



The 'Unbound Bible'

The Unbound Bible

[English] [Русский] [한국어]
[Translate] [Nederlands] [Français] [עברית] [English]

Bible Search Bible Study Tools Know God Downloads Translate
Bible in a Year Search Settings FAQ Links Contact
Download Unicode Fonts and Bibles

Unicode Bibles

Afrikaans 1953 † Download

Afrikaans 1953 †
Albanian †
Amharic NT
Arabic: Smith & Van Dyke †
Aramaic NT: Peshitta †
Armenian (Eastern): (Genesis, Exodus, Gospels)
Armenian (Western): NT (Dwight/Riggs, 1853)
Basque (Navarro-Labourdin): NT
Breton Gospels
Chamorro (Psalms, Gospels, Acts)
Chinese: NCV (Traditional) †
Chinese: Union (Simplified) †
Chinese: NCV (Simplified) †
Chinese: Union (Traditional) †
Croatian 2.0
Czech BKR †
Czech CEP †
Czech KMS †
Czech NKB †
Danish †
Dutch Staten Vertaling †
English: King James Version 2.0
English: American Standard Version
English: Basic English Bible
English: Darby Version
English: Douay-Rheims 2.0
English: Webster's Bible
English: Weymouth NT
English: World English Bible
English: Young's Literal Translation

mir Rybant has written a free software program that works with Bibles in the Unbound Bible format. You can look up a passage, and then copy and paste it into any Unicode-aware application.

UniBible - A Multilingual Bible Reader for PalmOS

UniBible is a multilingual Bible-reader program for the PalmOS platform. It uses Unicode to display Bible texts in various languages, including:

Armenian, Arabic, Chinese, English, French, German, Hebrew, Italian, Japanese, Korean, Portuguese, Russian, Spanish, and many others.

UniBible is available for download for free!

Unicode Fonts

UniBible uses Unicode fonts to display the Bible text. These fonts are available for download for free!

GB18030 Support Package

UniBible supports the GB18030 character encoding, which is used in China. This package contains the necessary fonts and files to use UniBible with GB18030 text.

**85 translations (some partial) in
51 languages, in common format**





The 'Unbound Bible' – a sample

web_utf8.txt - Notepad

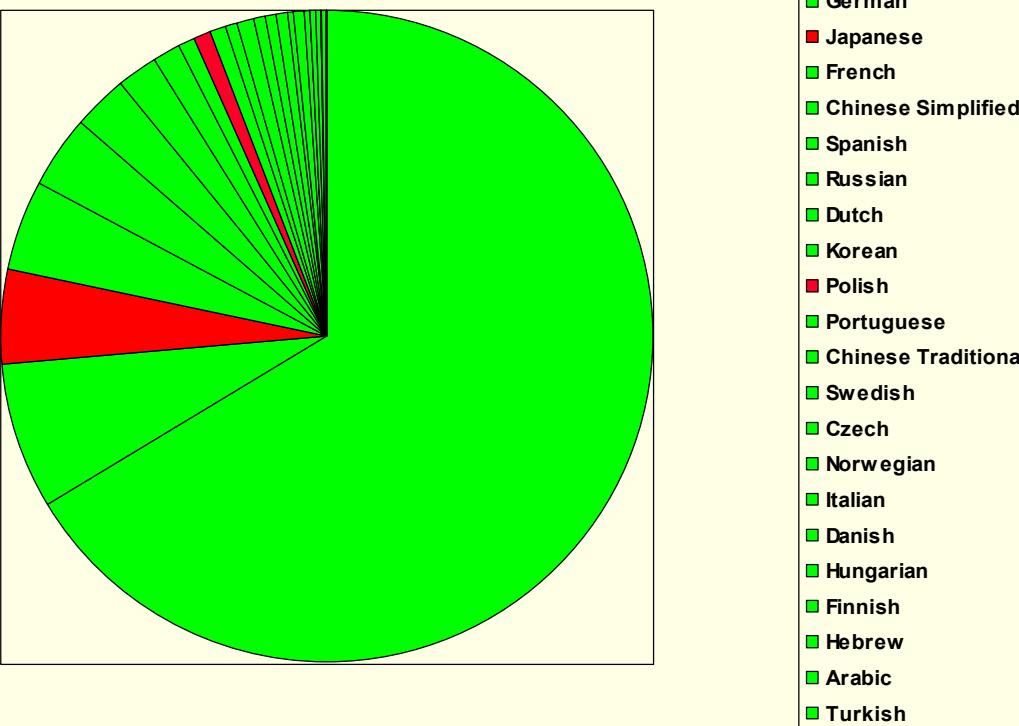
File Edit Format View Help

43N	1	1	In the beginning was the Word, and
43N	1	2	The same was in the beginning with
43N	1	3	All things were made through him.
43N	1	4	In him was life, and the life was
43N	1	5	The light shines in the darkness,
43N	1	6	There came a man, sent from God, v
43N	1	7	The same came as a witness, that h
43N	1	8	He was not the light, but was sent
43N	1	9	The true light that enlightens eve
43N	1	10	He was in the world, and the world
43N	1	11	He came to his own, and those who
43N	1	12	But as many as received him, to th
43N	1	13	who were born, not of blood, nor
43N	1	14	The Word became flesh, and lived
43N	1	15	John testified about him. He cried
43N	1	16	From his fullness we all received
43N	1	17	For the law was given through Mose
43N	1	18	No one has seen God at any time.
43N	1	19	This is John's testimony, when the
43N	1	20	He confessed, and didn't deny, but
43N	1	21	They asked him, "What then? Are yo





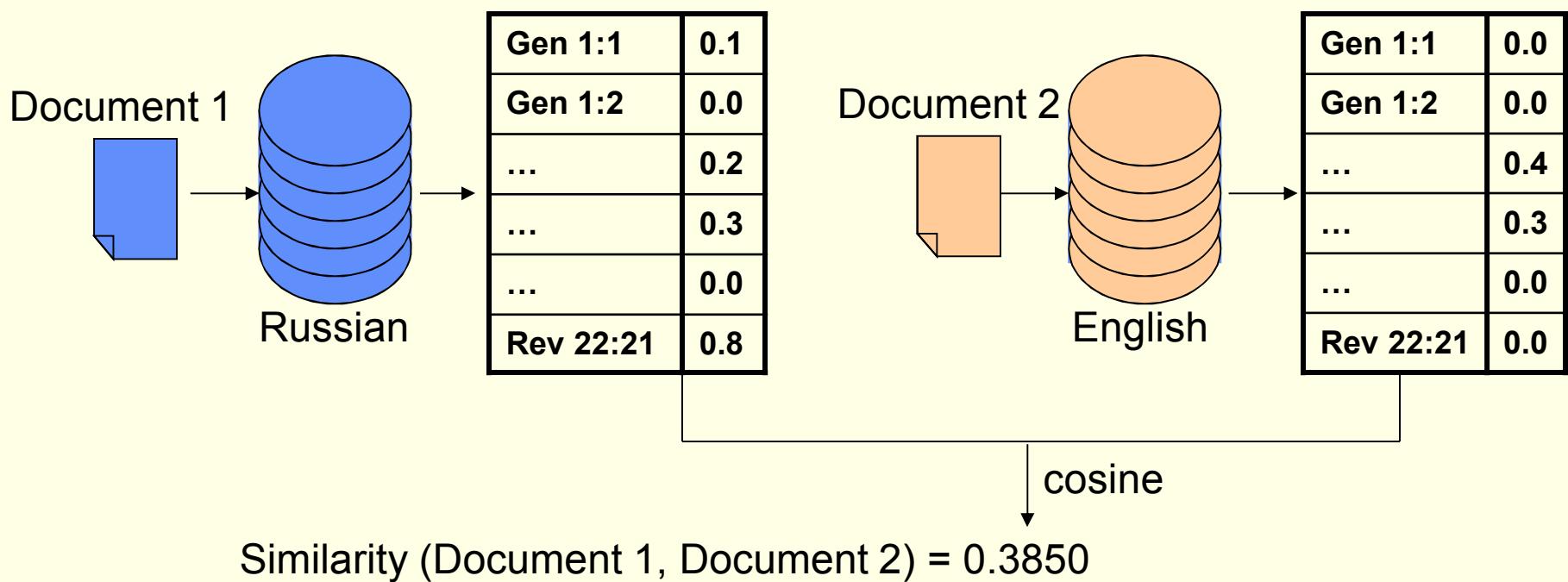
Coverage of internet content based on 'Unbound Bible'





Cross-Language Comparison – Method 1

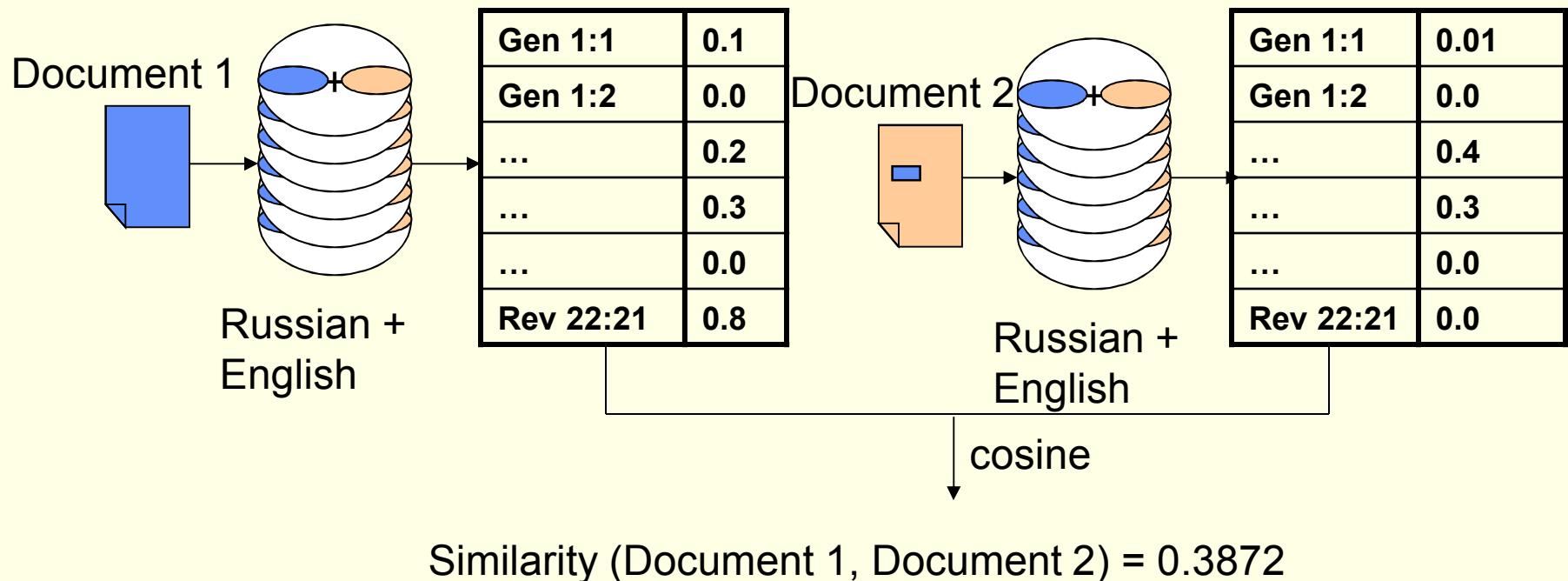
- Use separate STANLEY textual model for each language





Cross-Language Comparison – Method 2

- Use single STANLEY textual model of concatenated language document chunks





Validation results: Test set in training set

Method: Index on entire Bible, measure average uninterpolated precision at doc. 1 for 66 books of Bible

		To				
		Arabic	English	French	Russian	Spanish
From	Arabic	1.00	1.00	1.00	1.00	1.00
	English	1.00	1.00	1.00	1.00	1.00
	French	1.00	1.00	1.00	1.00	1.00
	Russian	1.00	.99	1.00	1.00	1.00
	Spanish	1.00	1.00	1.00	1.00	1.00



Validation results: Test set **not** in training set

Method: Index on entire Bible, obtain matrix of similarity measures for 5 conference abstracts where English and Spanish translations exist

		English				
		Doc 1	Doc 2	Doc 3	Doc 4	Doc 5
Spanish	Doc 1	.607	.043	.045	.035	.022
	Doc 2	.049	.397	.038	.082	.166
	Doc 3	.030	.050	.045	.101	.049
	Doc 4	.102	.096	.080	.189	.105
	Doc 5	.035	.131	.039	.042	.168

We have also used the framework successfully for Maori, to distinguish between the Treaty of Waitangi and the New Zealand National Anthem



Validation results: Test set **not** in training set

Method: Index on entire Bible, measure Mean Average Precision for 114 suras of Quran in English, Arabic, Russian, and Spanish (results comparable to McNamee & Mayfield 2004)

- LSA - no removal of stopwords		To					
		Arabic	English	Russian	Spanish		
From	Arabic	1.00	.71 .60	.62 .33	.72 .46		
	English	.71 .49	1.00	.90 .75	.90 .53		
	Russian	.56 .40	.92 .68	1.00	.67 .45		
	Spanish	.66 .46	.87 .78	.74 .62	1.00		

.35

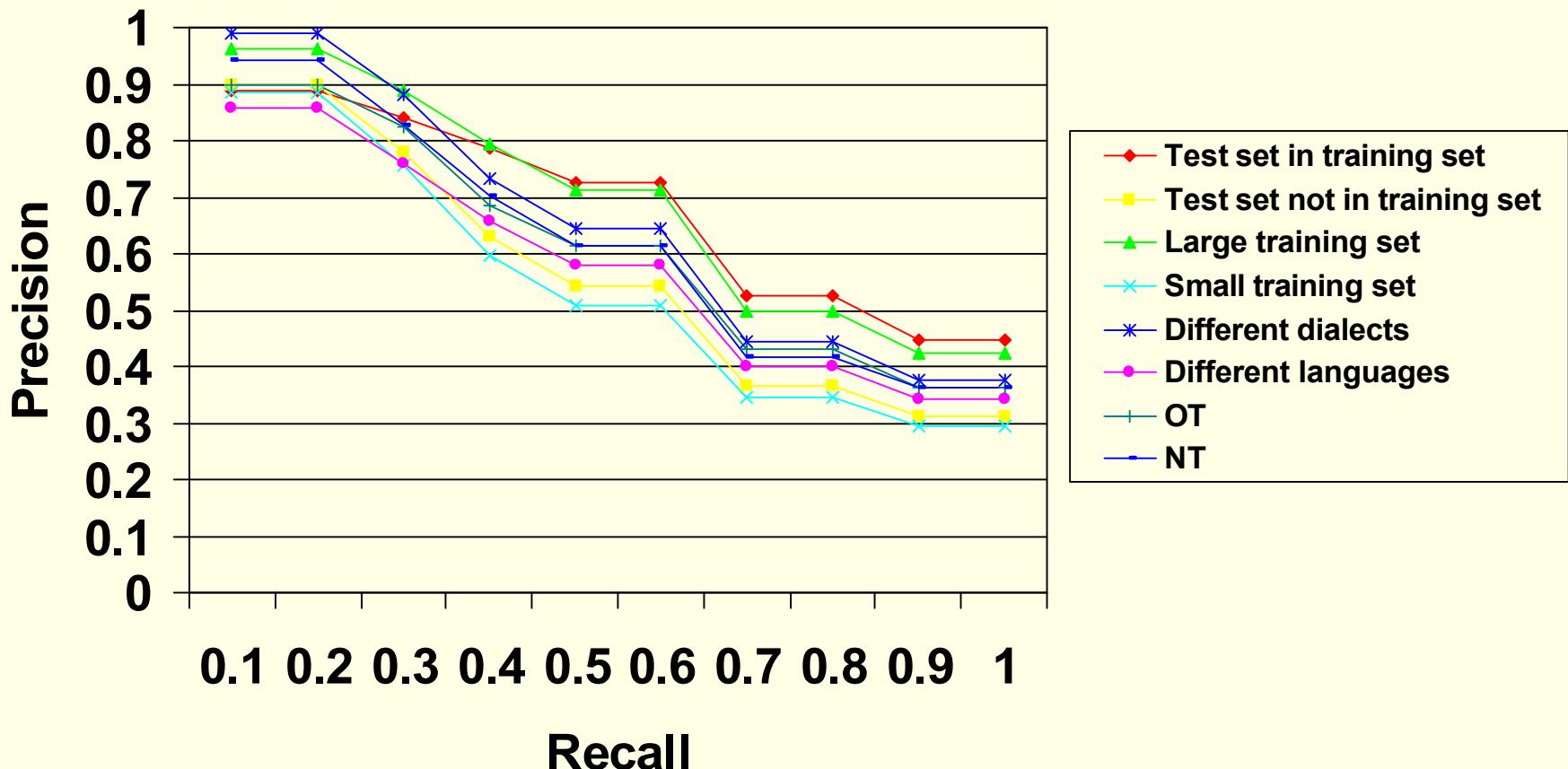
(using 5-
grams)

Method 1: Separate index for each language

Method 2: Single index for all languages



Precision-recall graphs for different test parameters



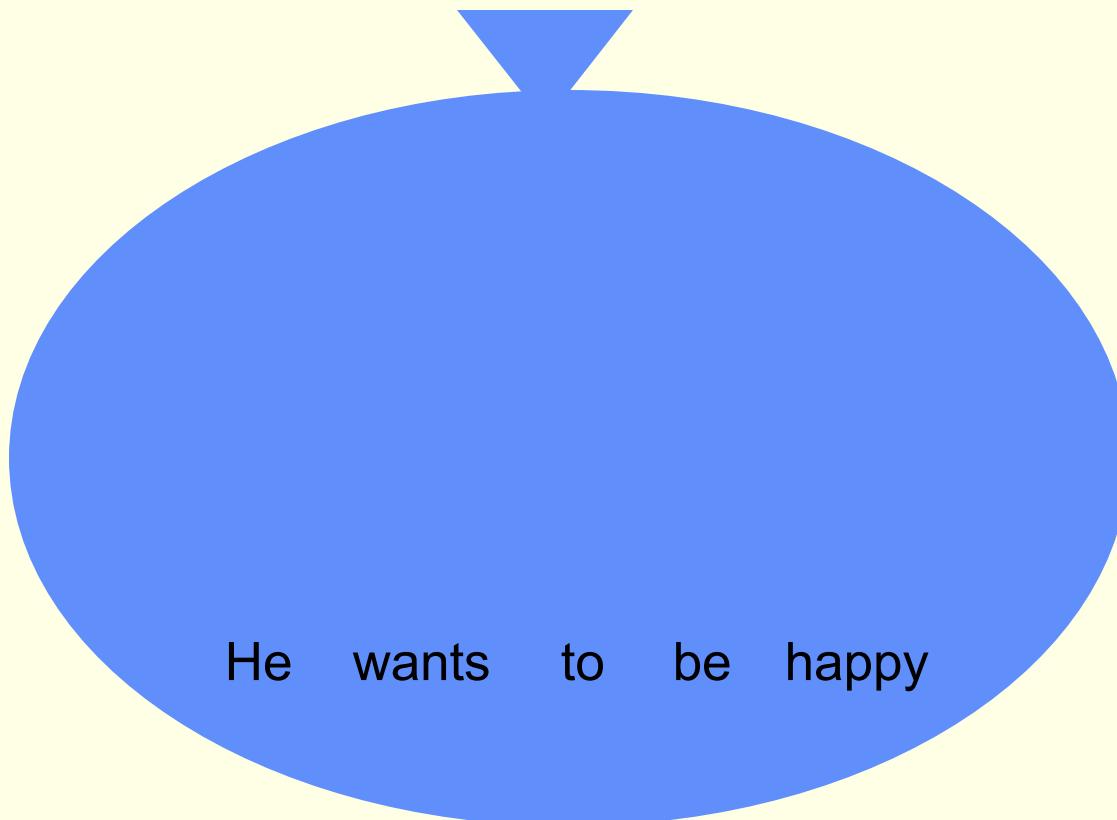


Observations

- Cross-language approach is easily extensible to new languages and corpora, including minority languages
- Resources exist which allow large parallel corpora to be built up from scratch in hours, at no monetary cost
- Unsurprisingly, the larger the training set, the better the precision-recall results
- Results appear to be comparable to, or better than, those achieved by other methods and reported recently in research literature; further testing may be needed
- Our best results were obtained using LSA, a single index for all languages, and without removing stopwords. This has the advantage of requiring no language-specific expertise to set up.



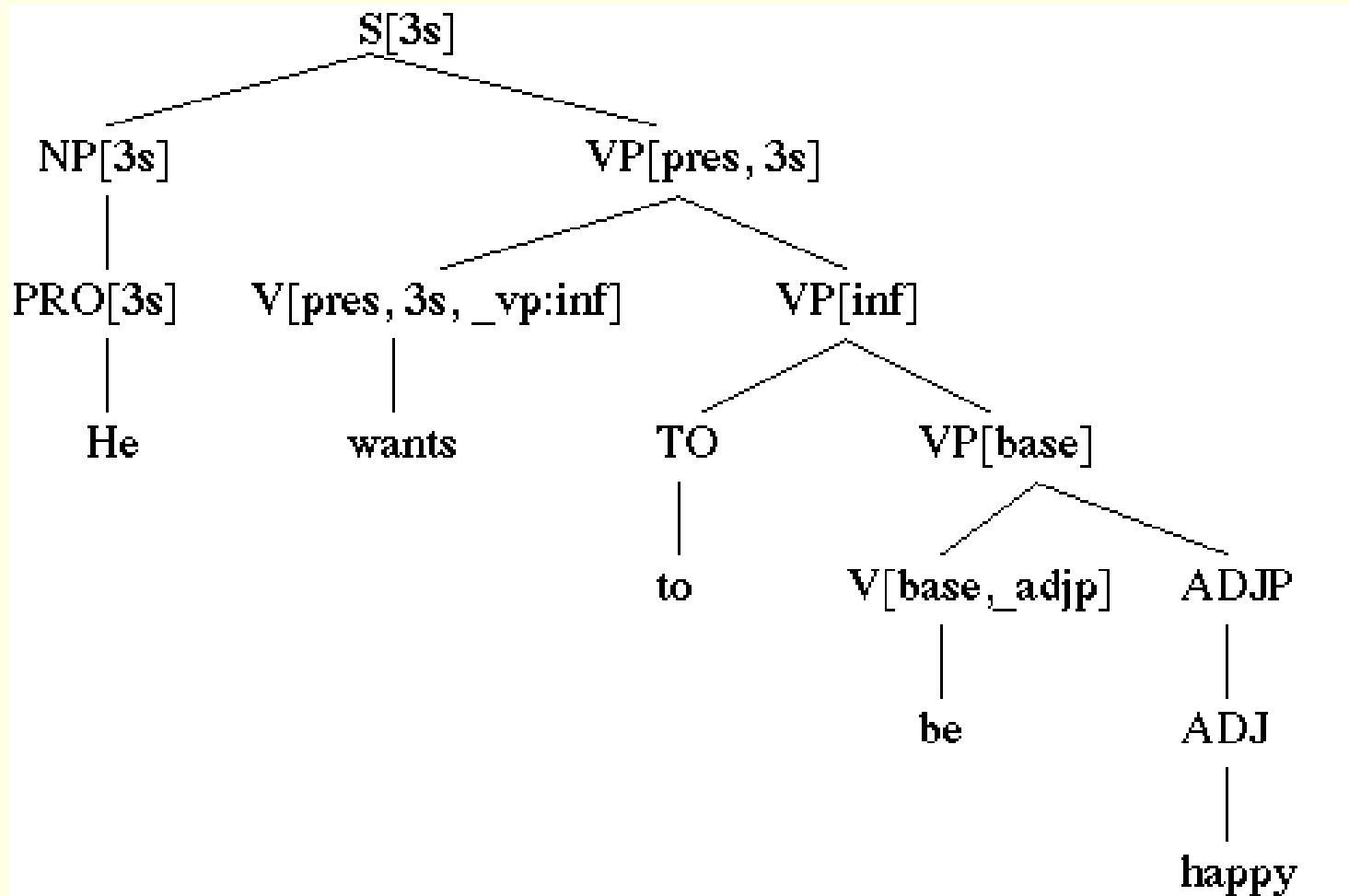
Beyond the 'bundle of words'



The 'bundle of words'



Of course, order does matter in language





Why does the bundle of words approach work at all?

Assume:

- vocabulary of 100,000 (10^5) words
- sentences are 20 words long
- word order important only within sentences

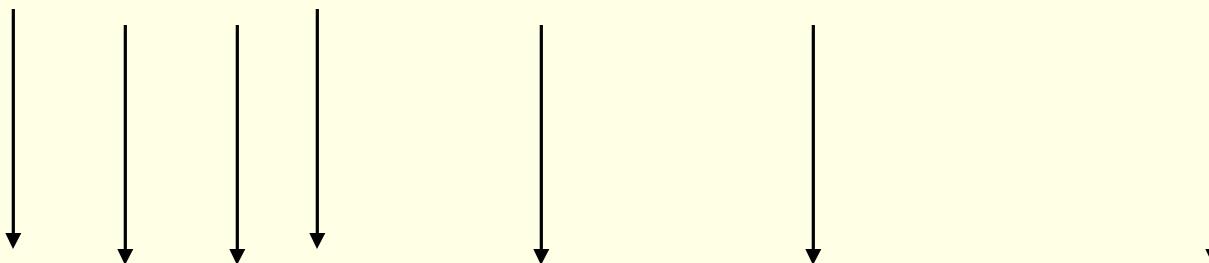
⇒ Contribution (in bits) to passage ‘information’:

- | | | | | |
|--|-------------------------------|----------|----------|---------|
| – From word <i>choice</i> : $\log_2(20!)$ | ≈ | 61.0774 | 84.47% | |
| – From word <i>order</i> : $\log_2((10^5)^{20})$ | ≈ | 332.1928 | 15.53% | |
| – Total | $\log_2(20! \times 10^{100})$ | ≈ | 393.2702 | 100.00% |



N-Word-Grams

- **Consecutive N-Word Lexical Entities**
- **Statistical Windowing Technique**
 - **window width**
 - **step size**

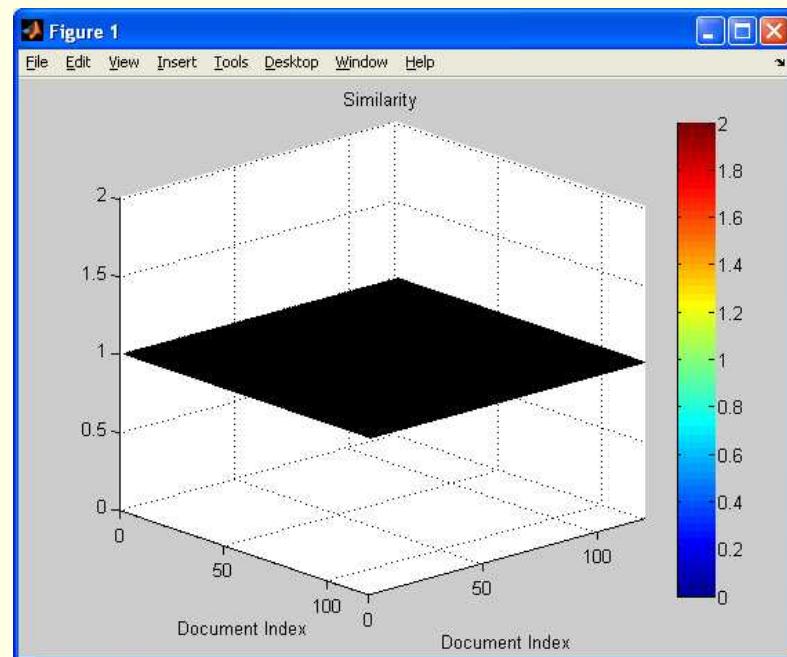
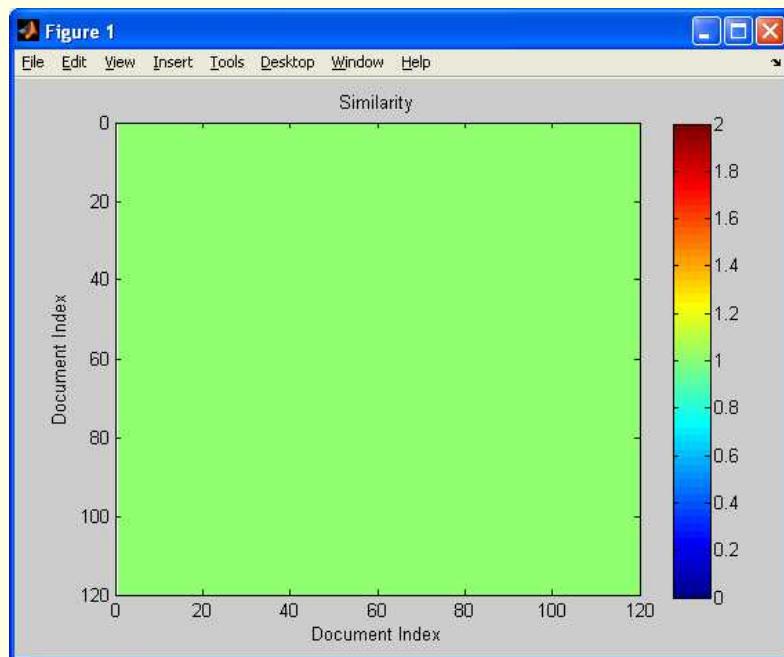


- **We are at Sandia National Laboratories ...**



Document-to-Document Similarities

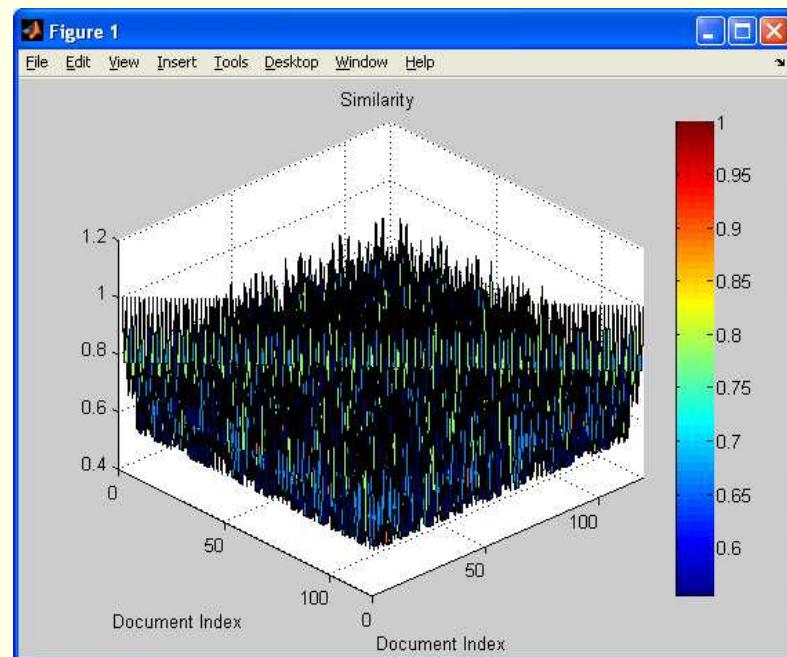
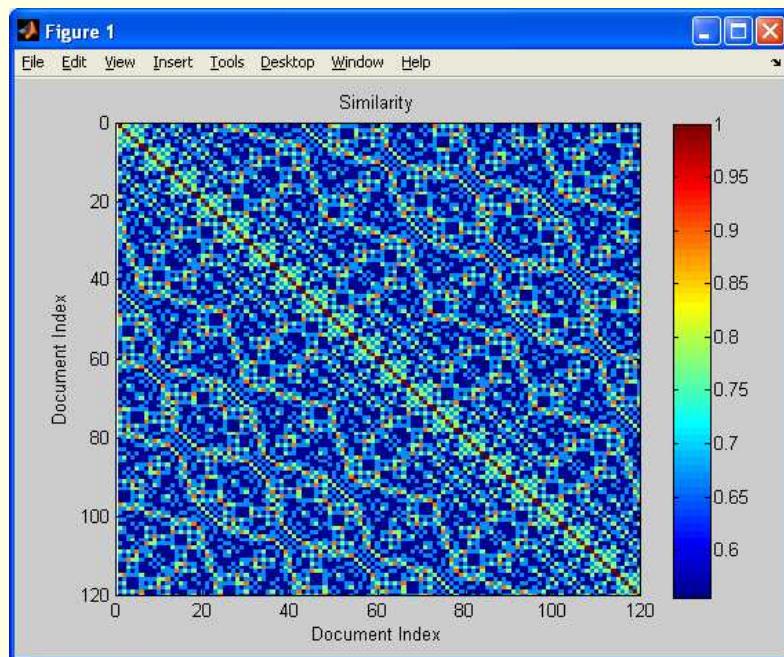
- All 120 permutations ($n = 1$ word-grams)





Document-to-Document Similarities

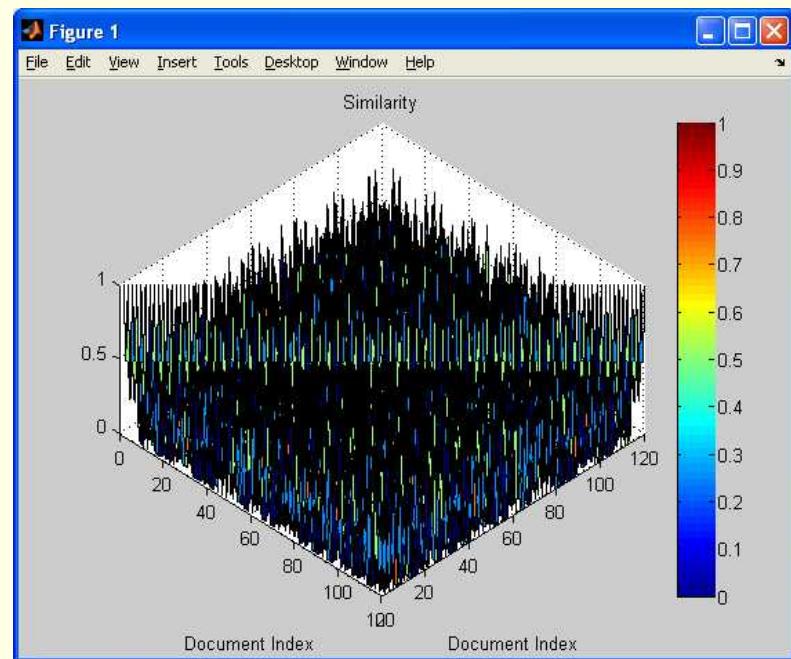
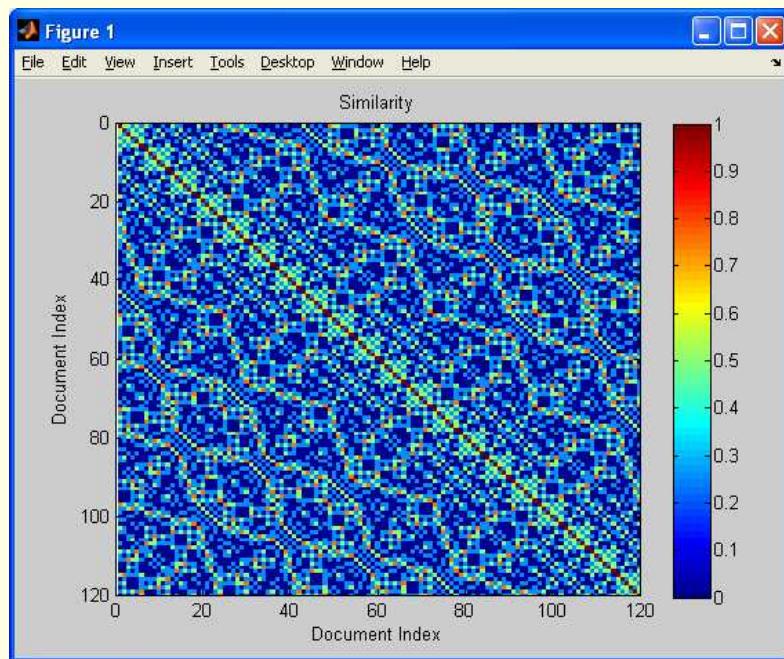
- All 120 permutations ($n \leq 2$ word-grams)





Document-to-Document Similarities

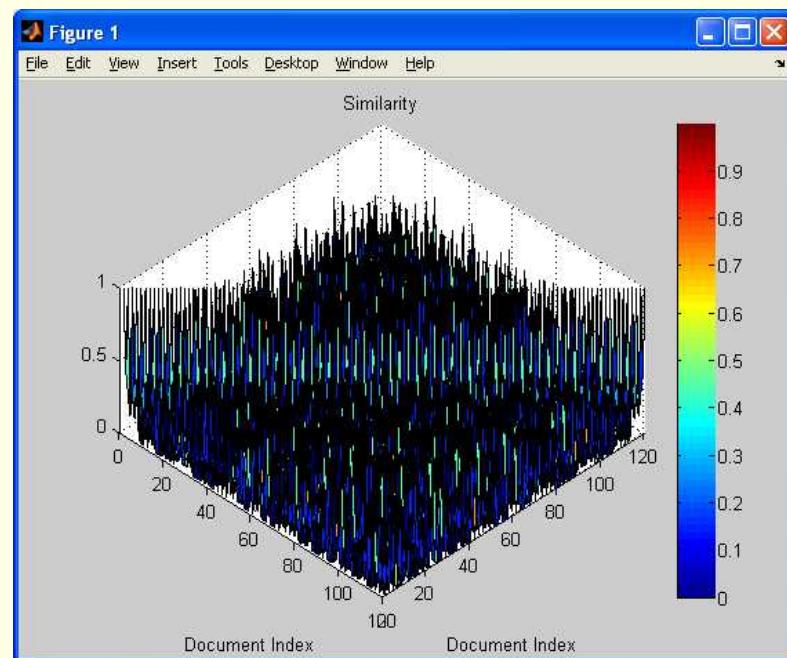
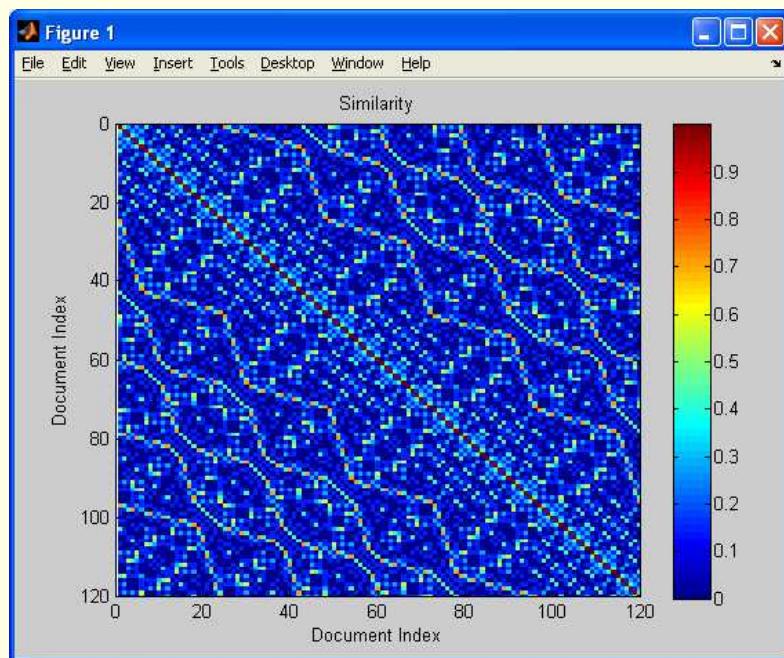
- All 120 permutations ($n = 2$ word-grams)





Document-to-Document Similarities

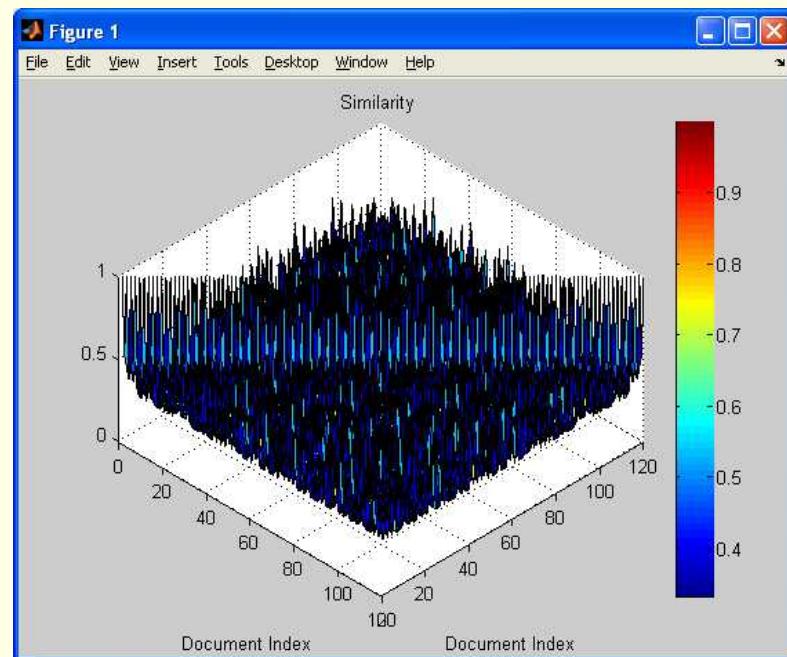
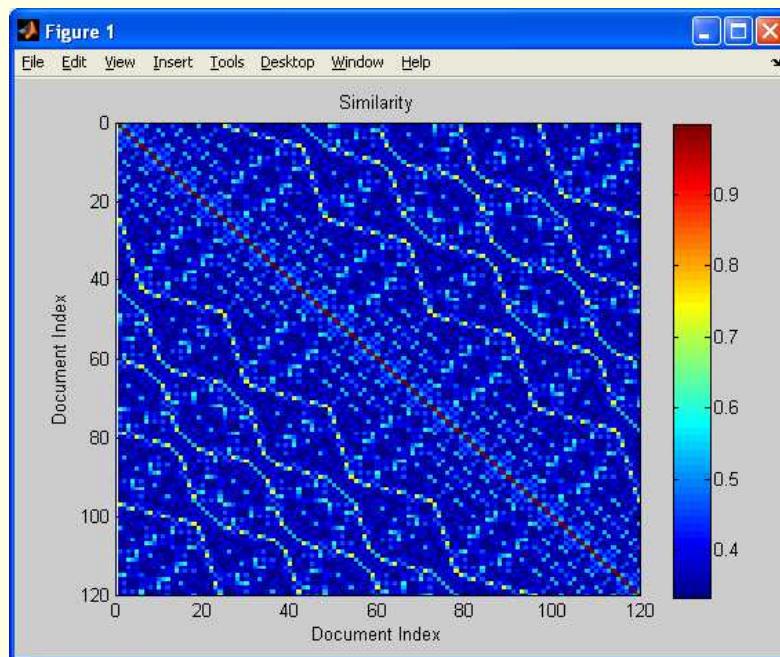
- All 120 permutations ($2 \leq n \leq 3$ word-grams)





Document-to-Document Similarities

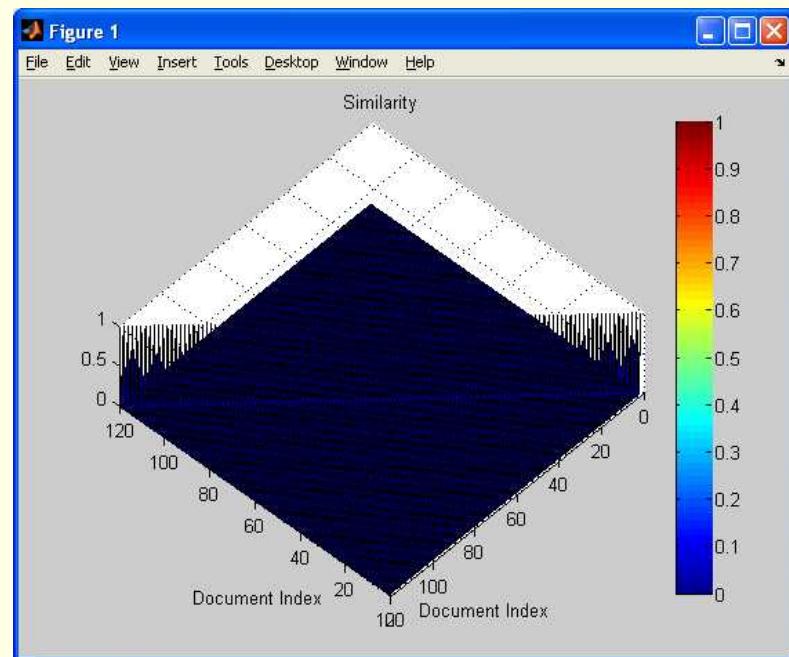
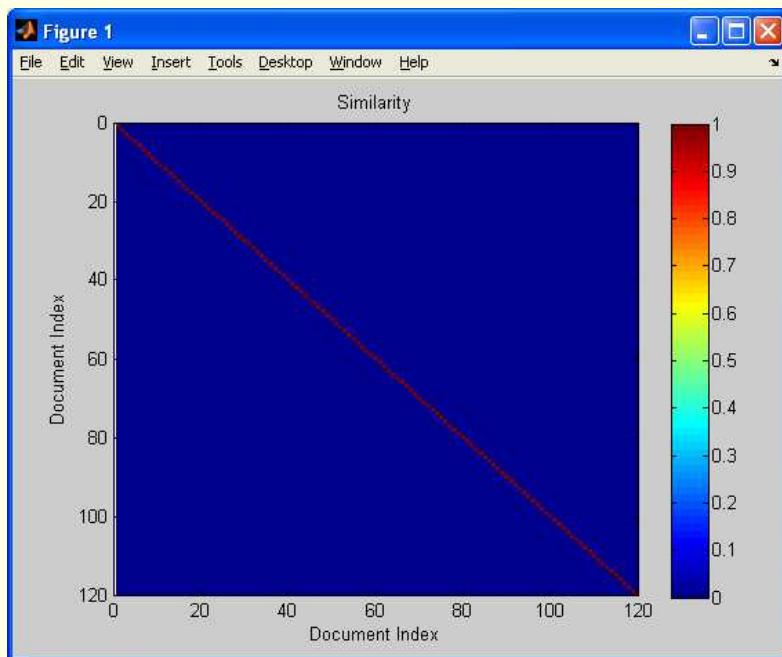
- All 120 permutations ($n \leq 5$ word-grams)





Document-to-Document Similarities

- All 120 permutations ($n = 5$ word-grams)





Ideological documents of known threat level

E.O. 13224: Identified Terrorists and Groups - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Stop Home Search Favorites Address http://www.state.gov/ct/rls/fs/2001/6531.htm

Google [Search](#) 165 blocked ABC Check AutoLink AutoFill Options State department terrorist list Go Links

Iran: Diplomatic Solution Needed To Resolve Nuclear Issue | Daily Press Briefing | Other News...

U.S. DEPARTMENT of STATE SEARCH Go Subject Index

Secretary Rice To Be

Home Issues & Press Travel & Business Youth & Education About State Department

Bureaus/Offices Reporting Directly to the Secretary > Office of the Coordinator for Counterterrorism > Releases > Fact Sheets > 2001

Chronology
The Office of the Coordinator for Counterterrorism
Washington, DC
December 31, 2001

Comprehensive List of Terrorists and Groups Identified Under Executive Order 13224

Executive Order 13224, signed by President Bush on September 23, blocks the assets of organizations and individuals linked to terrorism. There are now 189 such groups, entities, and individuals covered by the Executive Order. Following is a complete listing.

Original Annex: September 23, 2001

1. Al-Qaida/Islamic Army*
2. Abu Sayyaf Group*
3. Armed Islamic Group (GIA)*
4. Harakat ul-mujahidin (HUM)*
5. Al-Jihad (Egyptian Islamic Jihad)*
6. Islamic Movement of Uzbekistan (IMU)*
7. Asbat al-Ansar
8. Salafist Group for Call and Combat (GSPC)
9. Libyan Islamic Fighting Group
10. Al-Itta'ad al-Islamiya (AIA)
11. Islamic Army of Aden
12. Usama bin Laden ("Most Wanted" Terrorist)
13. Muhammad Atta/Subhi Abu Sitta/Abu Hafs al-Masri ("Most Wanted" Terrorist – killed in Afghanistan)
14. Sayf al-Adl ("Most Wanted" Terrorist)
15. Shaykh Salaf/Mustafa Muhammad Ahmad
16. Abu Hafs the Mauritanian/Mahfouz Ould al-Valid/Khalid al-Shanqiti
17. Ibn al-Shaykh al-Libi
18. Abu Zubaydar/Zayn al-Abidin Muhammad Husayn Tariq
19. Abd al-Hadi al-Iraqi/Abu Abdullah
20. Ayman al-Zawahiri ("Most Wanted" Terrorist)
21. Thirwan Salih Shihata/Muhammad Ali
22. Tariq Anwar al-Sayid Ahmad/Fathi/Amr al-Fatih
23. Muhammad Salah/Nasr Fahmi Nasr Hasanayn
24. Makhtab al-Khidamat/Al-Khifaf
25. Wifa Humanitarian Organization
26. Al-Rashid Trust
27. Mamoun Darkazani Import-Export Company

(Note: Groups with asterisks are also designated as Foreign Terrorist Organizations under the Anti-terrorism and Effective Death Penalty Act)

Internet





Ideological documents of known threat level

Microsoft SQL Server Management Studio

File Edit View Query Project Tools Window Community Help

New Query Execute

threat_id

SQLQuery1.sql* Summary

Object Explorer

Connect

SELECT dbo.ideological_organization.ideal...
FROM dbo.government_organization INNER JOIN
dbo.terrorist_proscription ON
dbo.government_organization.government_organization_id = dbo.terrorist_proscription.government_organization_id INNER JOIN
dbo.ideal...
dbo.terrorist_proscription.ideal... = dbo.ideal...
dbo.web_document ON dbo.ideal... = dbo.web_document.ideal...

Results Messages

	Expr1	organization_na...	ideological_organization_description	web_document_text	web_document_url
1	Real IRA	US State Depar...	NULL	Ireland's OWN: History	http://irelandsown.net/whyrira.html
2	Real IRA	US State Depar...	NULL	The Real IRA -by Michealín Ni Dhochartaigh	http://irelandsown.net/RIRA.html
3	Kahane Chai (Kach)	US State Depar...	NULL	Kahane.org - Principles and Philosophy The Kahane m...	http://www.kahane.org/philosophy.htm
4	Palestinian Islamic Jihad...	US State Depar...	PIJ is a Shi'a group which aims to end the Israeli...	In the Name of Allah, the Merciful, the Mercy Giving ...	http://www.qudsway.com/Links/English_Neda/AboutUs/Link_AboutUs.htm
5	Palestinian Islamic Jihad...	US State Depar...	PIJ is a Shi'a group which aims to end the Israeli...	PALESTINE HISTORY Early History The character ...	http://www.qudsway.com/Links/English_Neda/PalestinianFacts/Htm_Palestinian/hp...
6	Al-Aqsa Martyrs Brigade	US State Depar...	NULL	من أخبارنا تفاصيل خطيرة حول عمليات الاغتيال الأخيرة	http://www.kataebqasal.com/arabic/index.php
7	HAMAS (Hamas Izz al-...	US State Depar...	Hamas aims to end Israeli occupation in Palestin...	Palestine was predestined by God Almighty to be the lan...	http://www.palestine-info.co.uk/am/publish/article_13.shtml
8	HAMAS (Hamas Izz al-...	US State Depar...	Hamas aims to end Israeli occupation in Palestin...	THE CHARTER OF THE HAMAS: Ideological Starting...	http://72.14.203.104/search?q=cache:qbHTeNowA98J:www.someplacesomewhere....
9	Real IRA	EU Official Jour...	NULL	Ireland's OWN: History	http://irelandsown.net/whyrira.html
10	Real IRA	EU Official Jour...	NULL	The Real IRA -by Michealín Ni Dhochartaigh	http://irelandsown.net/RIRA.html
11	Kahane Chai (Kach)	EU Official Jour...	NULL	Kahane.org - Principles and Philosophy The Kahane m...	http://www.kahane.org/philosophy.htm
12	Palestinian Islamic Jihad...	EU Official Jour...	PIJ is a Shi'a group which aims to end the Israeli...	In the Name of Allah, the Merciful, the Mercy Giving ...	http://www.qudsway.com/Links/English_Neda/AboutUs/Link_AboutUs.htm
13	Palestinian Islamic Jihad...	EU Official Jour...	PIJ is a Shi'a group which aims to end the Israeli...	PALESTINE HISTORY Early History The character ...	http://www.qudsway.com/Links/English_Neda/PalestinianFacts/Htm_Palestinian/hp...
14	Al-Aqsa Martyrs Brigade	EU Official Jour...	NULL	من أخبارنا تفاصيل خطيرة حول عمليات الاغتيال الأخيرة	http://www.kataebqasal.com/arabic/index.php
15	HAMAS (Hamas Izz al-...	EU Official Jour...	Hamas aims to end Israeli occupation in Palestin...	Palestine was predestined by God Almighty to be the lan...	http://www.palestine-info.co.uk/am/publish/article_13.shtml
16	HAMAS (Hamas Izz al-...	EU Official Jour...	Hamas aims to end Israeli occupation in Palestin...	THE CHARTER OF THE HAMAS: Ideological Starting...	http://72.14.203.104/search?q=cache:qbHTeNowA98J:www.someplacesomewhere....
17	Palestinian Islamic Jihad...	UK Home Office	PIJ is a Shi'a group which aims to end the Israeli...	PALESTINE HISTORY Early History The character ...	http://www.qudsway.com/Links/English_Neda/PalestinianFacts/Htm_Palestinian/hp...
18	HAMAS (Hamas Izz al-...	UK Home Office	Hamas aims to end Israeli occupation in Palestin...	Palestine was predestined by God Almighty to be the lan...	http://www.palestine-info.co.uk/am/publish/article_13.shtml
19	HAMAS (Hamas Izz al-...	UK Home Office	Hamas aims to end Israeli occupation in Palestin...	THE CHARTER OF THE HAMAS: Ideological Starting...	http://72.14.203.104/search?q=cache:qbHTeNowA98J:www.someplacesomewhere....
20	Palestinian Islamic Jihad...	UK Home Office	PIJ is a Shi'a group which aims to end the Israeli...	In the Name of Allah, the Merciful, the Mercy Giving ...	http://www.qudsway.com/Links/English_Neda/AboutUs/Link_AboutUs.htm

Query executed successfully.

00:00:01 20 rows

Ready

Ln 17 Col 4 INS



Detecting Ideology

Words, or language, persuade people to join movements which can in turn become threats.



threats



movements



ideas



words



Similar Web Documents

Communist Manifesto

Mein Kampf

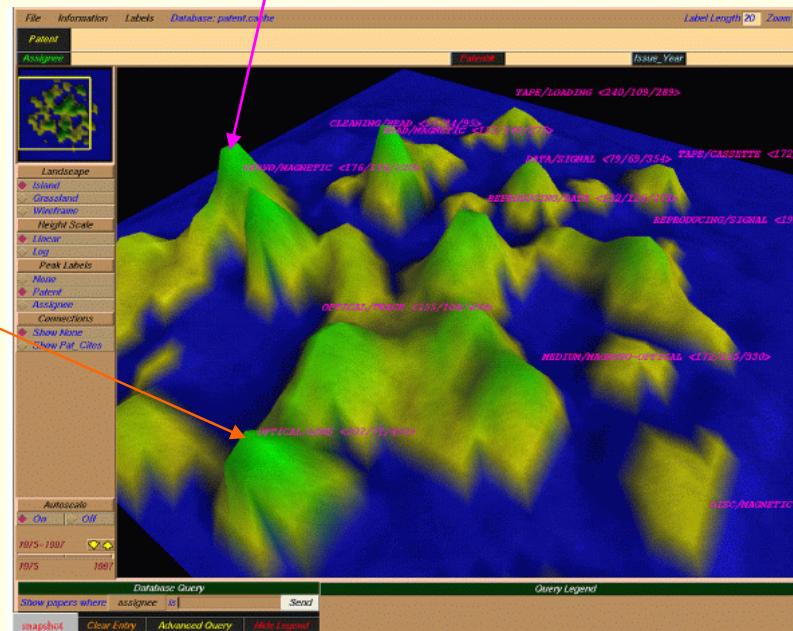


Figure 2: The mountain terrain metaphor provides an intuitive exploration environment.



Future directions

- Continue to test CLIR algorithm to identify its strengths and weaknesses
- What chunk size yields best cross-language IR results, and why?
- Can we use the output of cross-language comparison to characterize documents by their ideology?
- Add further languages for more coverage of WWW
- Linguistic analysis beyond the ‘bundle of words’ approach



QUESTIONS?



Selected References

- Ackland, R. 2005. Mapping the U.S. Political Blogosphere: Are Conservative Bloggers More Prominent? mimeo., The Australian National University.
- Chomsky, Noam. 1988 *Language and Politics*. C.P. Otero (ed.). Montreal: Black Rose Books.
- Landauer, Thomas. 1998. An Introduction to Latent Semantic Analysis. In *Discourse Processes* 25, 259-284.
- McNamee, Paul, and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7:73-97.
- Resnik, Philip, Mari Olsen, and Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the Book of 2000 Tongues. In *Computers and the Humanities* 33, 1-2. pp 129-153.
- Resnik, Philip, and Noah Smith. 2003. The Web as a Parallel Corpus. In *Computational Linguistics* 29, No. 3, pp. 349-380.



Selected References – “Bundle of Words”

- Thomas K. Laudauer, Darrell Laham, Bob Rehder, and M. E. Schreiner, “How Well Can Passage Meaning be Derived without Using Word-Order? A Comparison of Latent Semantic Analysis and Humans,” Proceedings of the 19th Annual Meeting of the Cognitive Science Society, 1997.
- Fernando Pereria, “Formal Grammar and Information Theory: Together Again?,” Philosophical Transactions of the Royal Society Series A 358, 1239-1253, 2000.
- Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger, “Deep Read: A Reading Comprehension System,” Proceedings of the ACL, 1999.



Selected References – “Word-Grams”

- Roger Zhang, “A Simple C++ N-gram Extraction Package”, Dalhousie University, Nova Scotia, Canada, November 28, 2004.

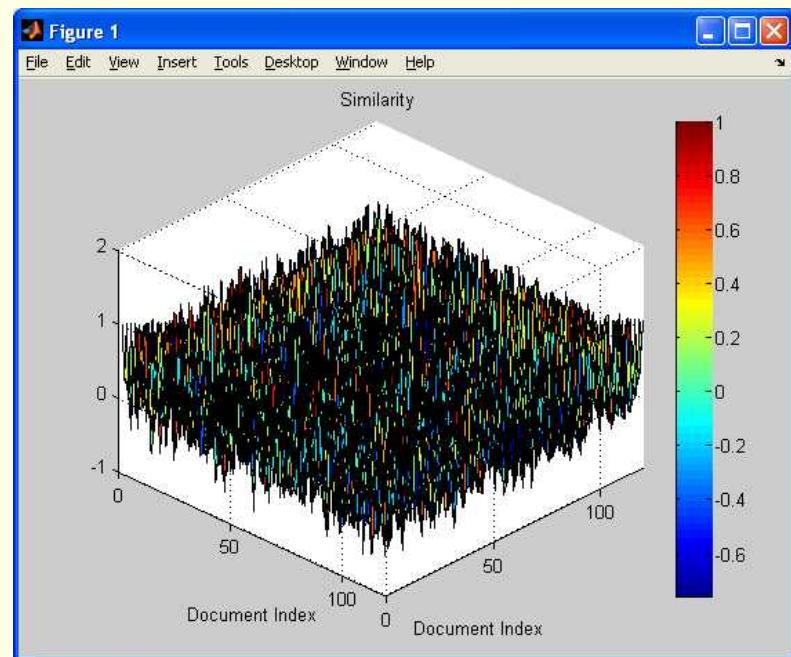
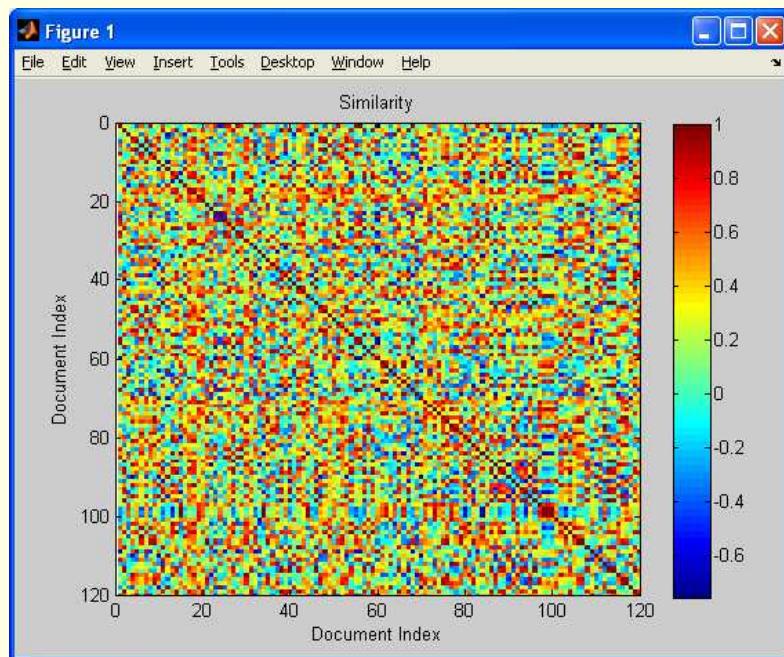


Backup Slides



Document-to-Document Similarities

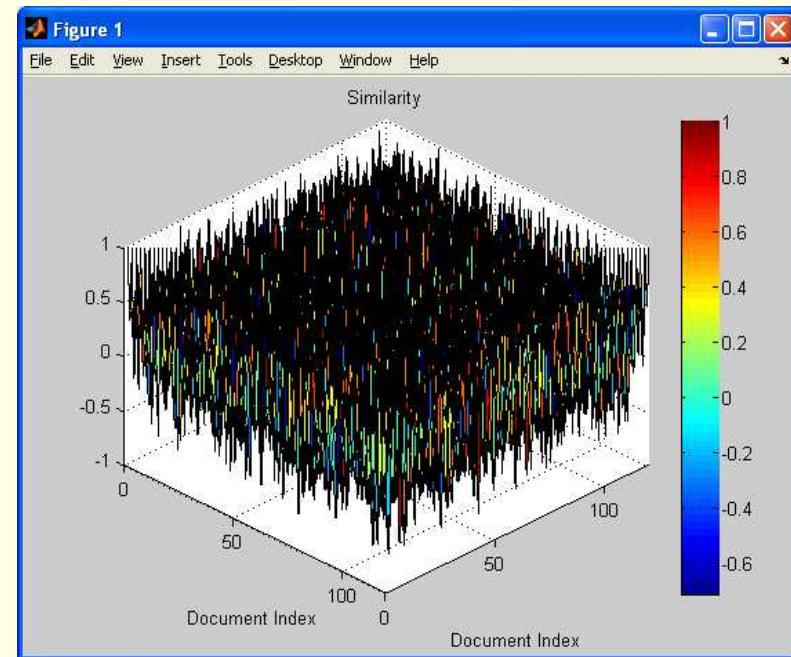
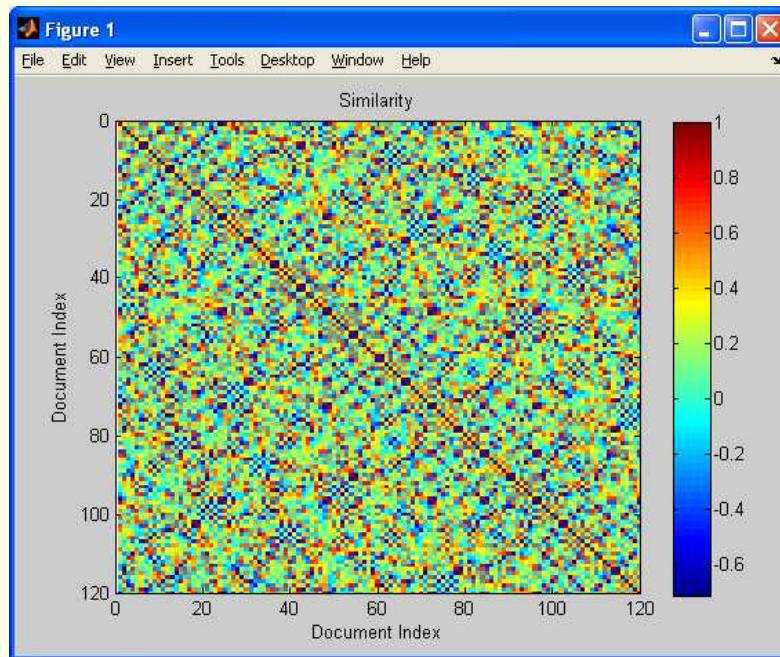
- All permutations ($n = 1$ word-grams) - LSA





Document-to-Document Similarities

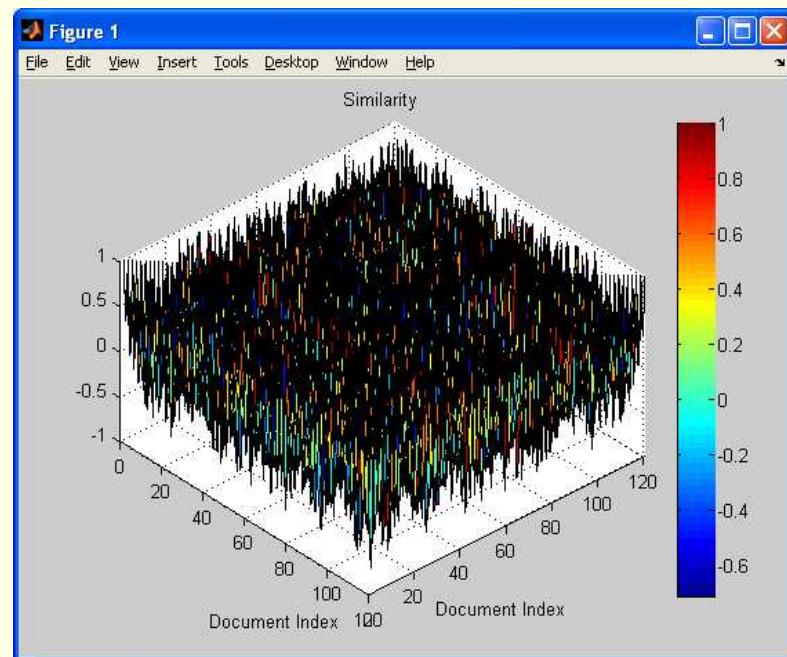
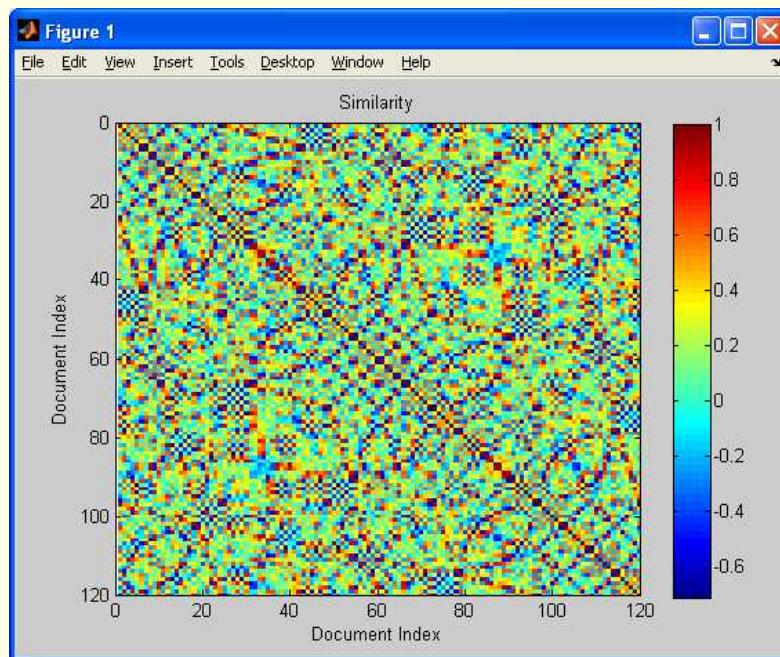
- All permutations ($n \leq 2$ word-grams) - LSA





Document-to-Document Similarities

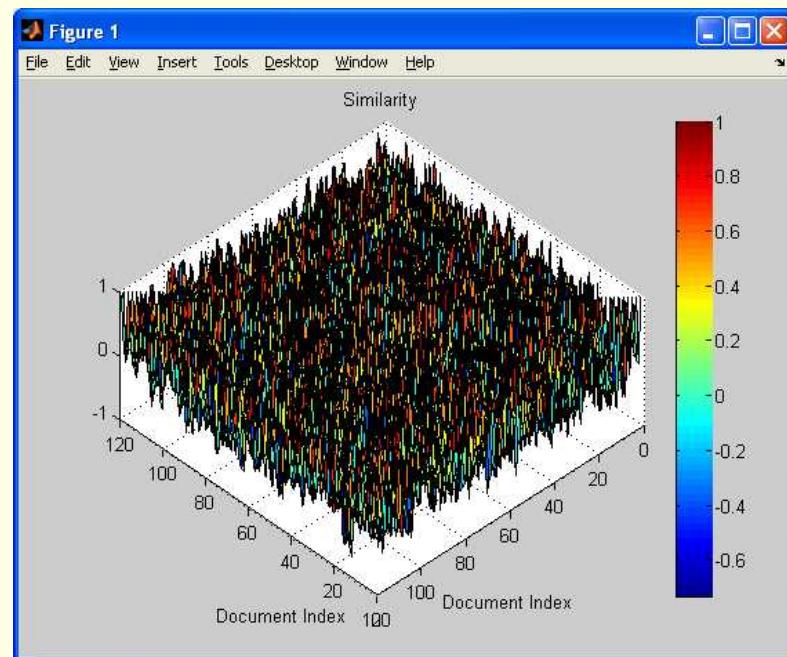
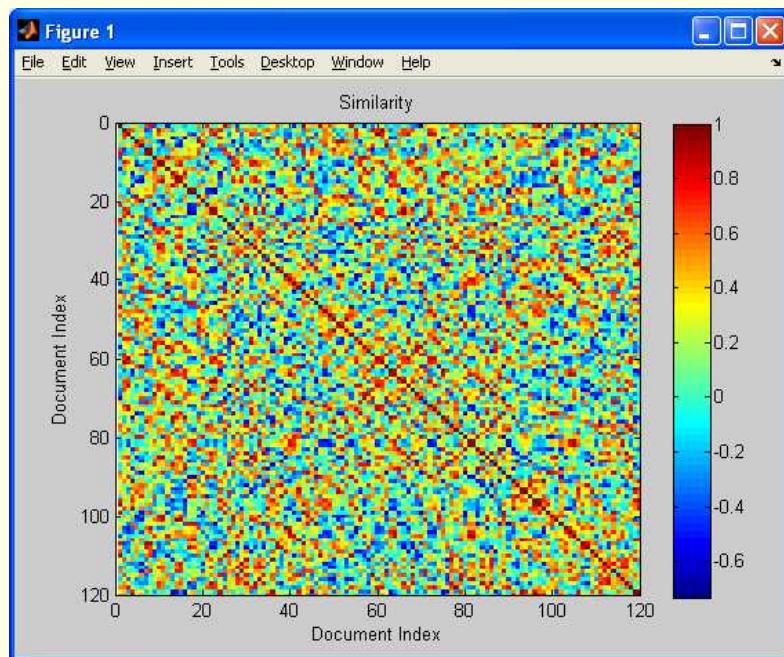
- All permutations (n = 2 word-grams) - LSA





Document-to-Document Similarities

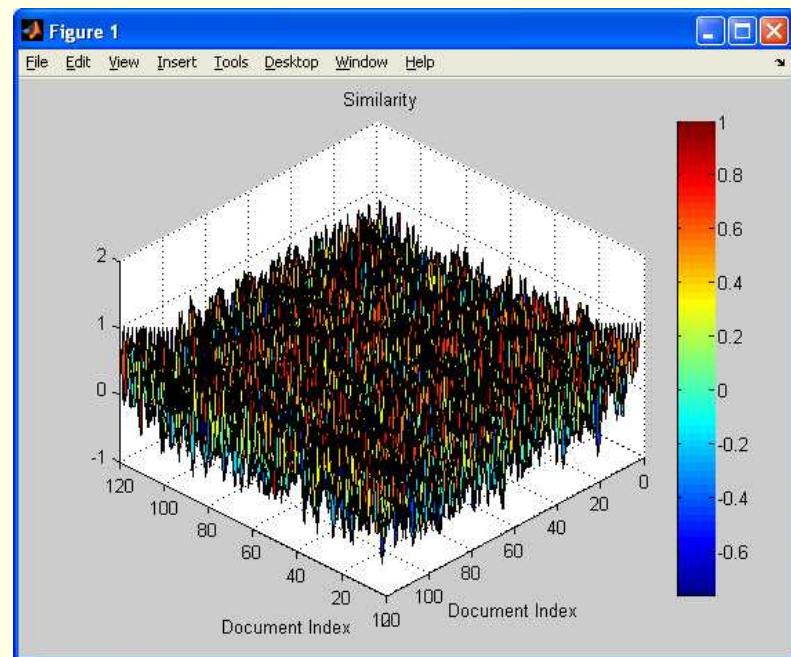
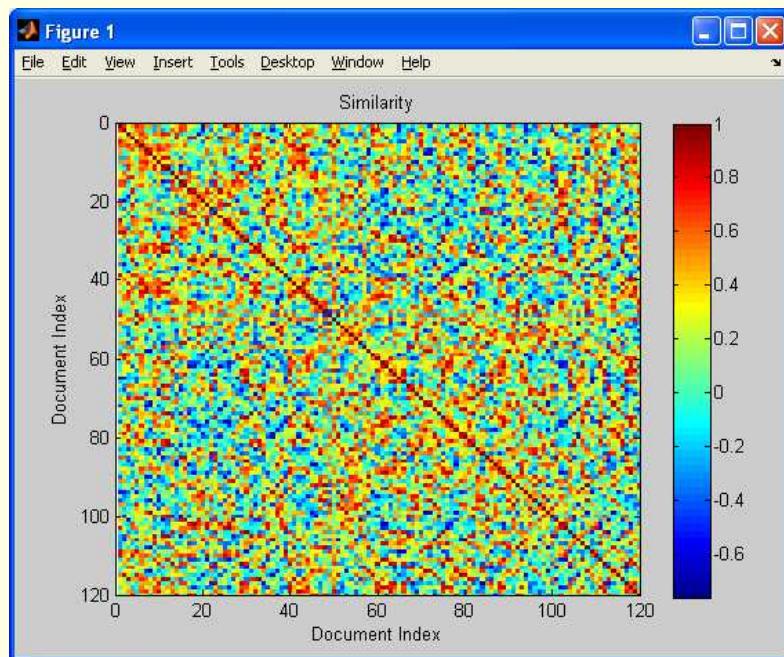
- All permutations ($2 \leq n \leq 3$ word-grams) - LSA





Document-to-Document Similarities

- All permutations ($n \leq 5$ word-grams) - LSA





Document-to-Document Similarities

- All permutations (n = 5 word-grams) - LSA

