# Temporal Information Retrieval

Wintersemester 2015/16

http://www.l3s.de/~anand/tir

**Introduction and Course Overview** 

# Who are we?

- Instructor: Dr. Avishek Anand
- TA: Jaspreet Singh
- L3S Research Center





- Fun with Web stuff News, Social Media, Wikipedia, Blogs, Archives, Flickr, Amazon reviews ......
- Analysis, Retrieval, Mining, Learning, Crawling,...
- Cutting edge research
- Building information systems for Web based data

Digital content is growing continuously...



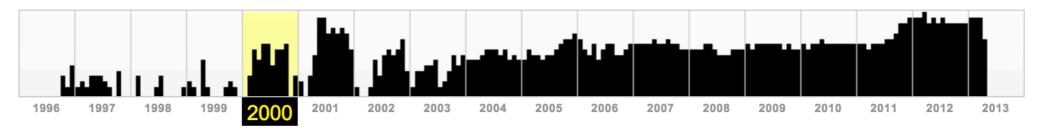


http://yahoo.com

Go Wayback!

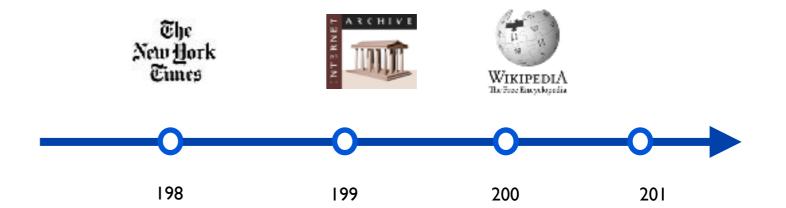
http://yahoo.com has been crawled 14,160 times going all the way back to October 17, 1996.

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. FAQ



JAN							FEB								MAR								APR							
						1			1	2	3	4	5					1	2	3	4								1	
2	3	4	5	6	7	8	6	7	8	9	10	11	12		5	6	7	8	9	10	11		2	3	4	5	6	7	8	
9	10	11	12	13	14	15	13	14	15	16	17	18	19		12	13	14	15	16	17	18		9	10	11	12	13	14	15	
16	17	18	19	20	21	22	20	21	22	23	24	25	26		19	20	21	22	23	24	25		16	17	18	19	20	21	22	
23	24	25	26	27	28	29	27	28	29						26	27	28	29	30	31			23	24	25	26	27	28	29	
30	31																						30							
MAY						JUN							JUL								AUG									
	1	2	3	4	5	6					1	2	3								1				1	2	3	4	5	
7	8	9	10	11	12	13	4	5	6	7	8	9	10		2	3	4	5	6	7	8		6	7	8	9	10	11	12	
14	15	16	17	18	19	20	11	12	13	14	15	16	17		9	10	11	12	13	14	15		13	14	15	16	17	18	19	
24		00	24	25	26	27	18	19	20	21	22	23	24		16	17	18	19	20	21	22		20	21	22	23	24	25	26	
21	22	23	24																											
28	22	30	31				25	26	27	28	29	30			23	24	25	26	27	28	29		27	28	29	30	31			

Digital content is growing continuously...

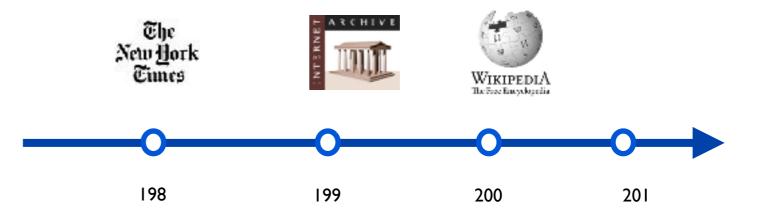


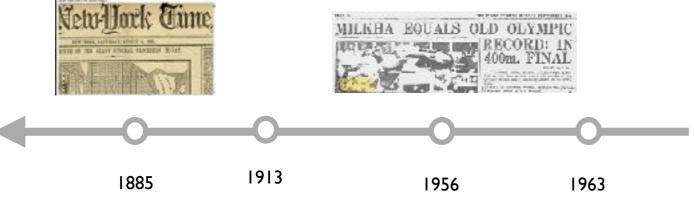


The Clinical and Laborated

Thank Glant White Stor Lines, Smire After Column Wildskellerg on Her Maiden Voyage, and 1,800 Lines Are Reported Lost a World's Greater Marine Disast KENNEDY DEAD treach failer in Date Lunder School Product Content increasingly getting digitized...

Digital content is growing continuously...



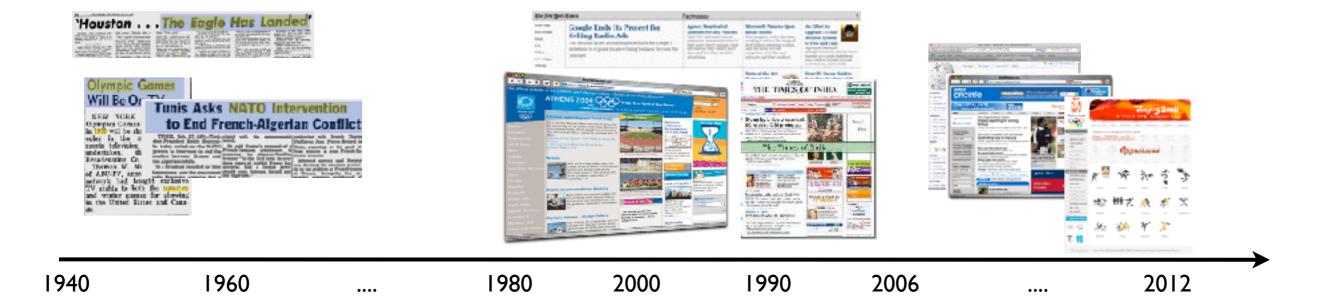


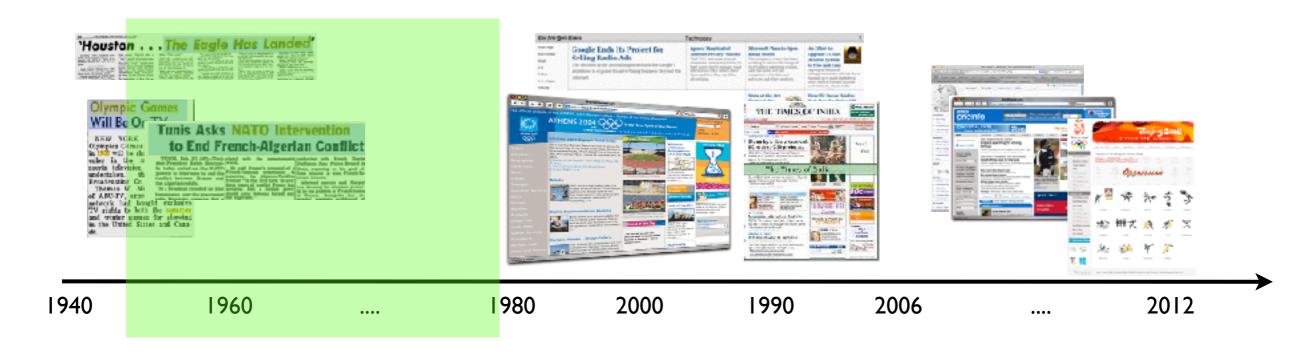
There Change at Large Sansa After Column
Wilskeberg on Her Maiden Voyage, and 1,600 Lives
Are Reported Lost as World's Greater Marine Planes
at the Column C

Content increasingly getting digitized...

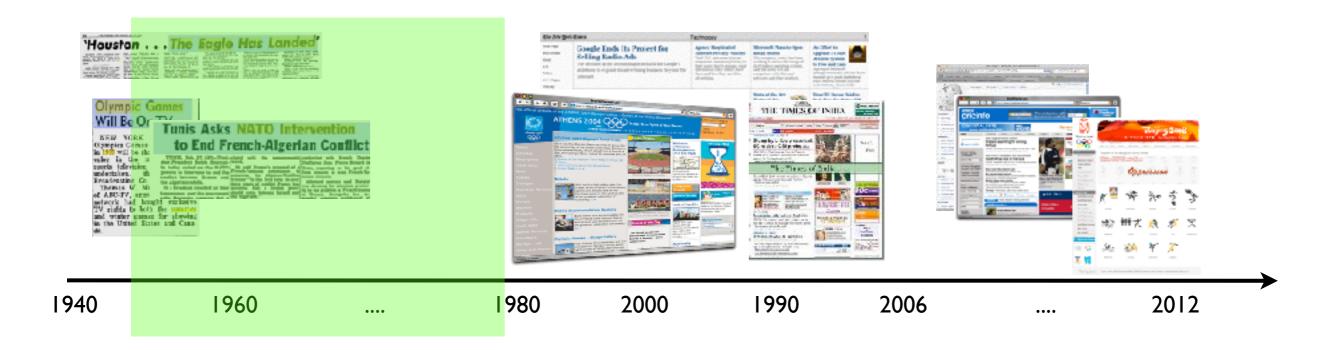
Content is spanning long time periods and growing

KENNEDY DEAD





nato intervention @ 1950-1980



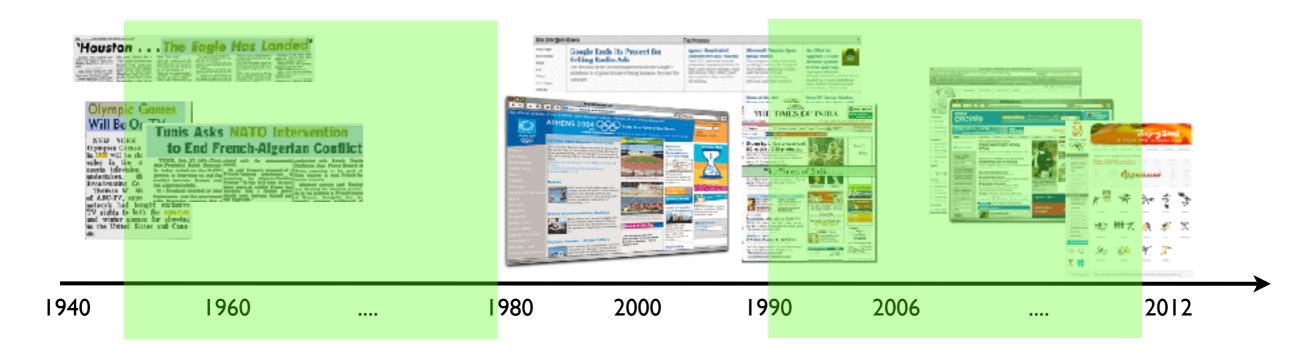
### nato intervention @ 1950-1980

### Tunis Asks Nato Intervention To End French-algerian...

Spokane Daily Chronicle - Feb 27, 1958

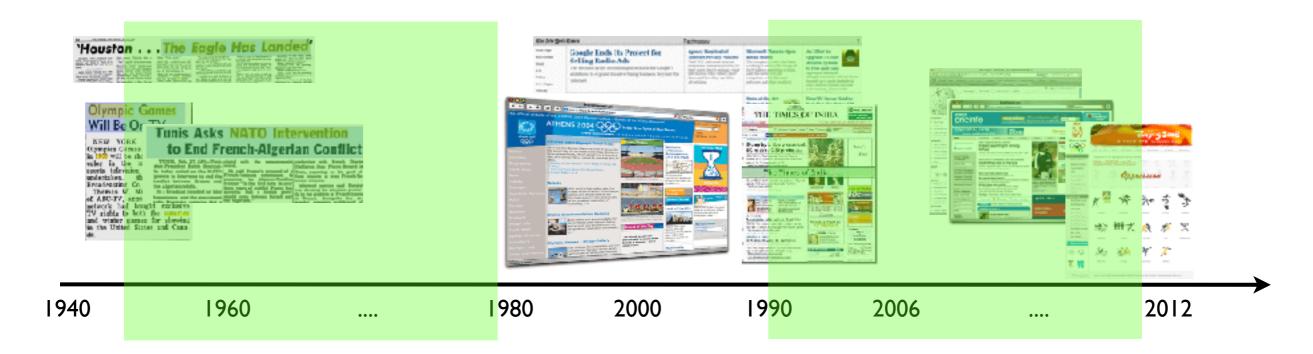
Habib Bourguiba today called on the **NATO** powers to **intervene** to end the conflict between France and the Algerian rebels. In a broadcast recorded for later ...

Nato .Intervention Ashed In French,... Reading Eagle



nato intervention @ 1950-1980

nato intervention @ 1990-2012



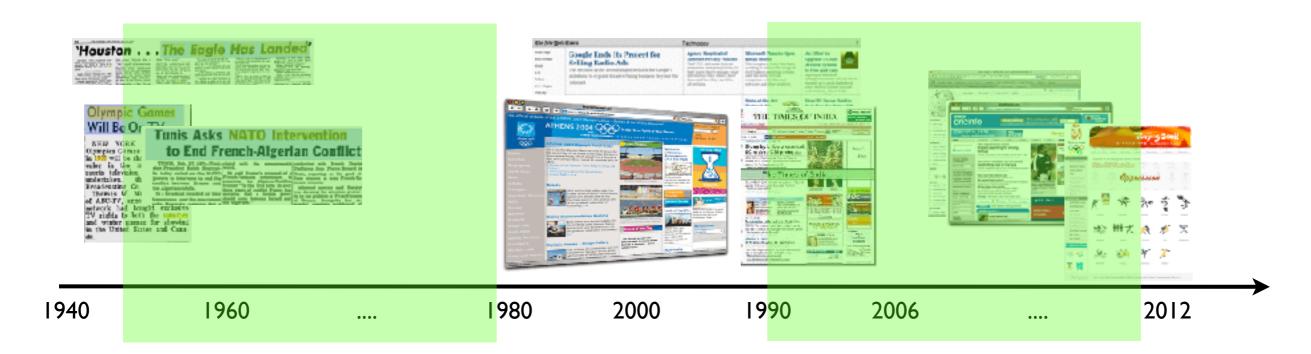
nato intervention @ 1950-1980

nato intervention @ 1990-2012

Time-travel queries: Keyword queries with time of interest

nato intervention @ 25/12/2001 nato intervention @ 1950-1980

time points vs time ranges



nato intervention @ 1950-1980

nato intervention @ 1990-2012

Time-travel queries: Keyword queries with time of interest

nato intervention @ 25/12/2001 nato intervention @ 1950-1980

time points vs time ranges

- Retrieve documents from the past
- Relevance based on time of interest

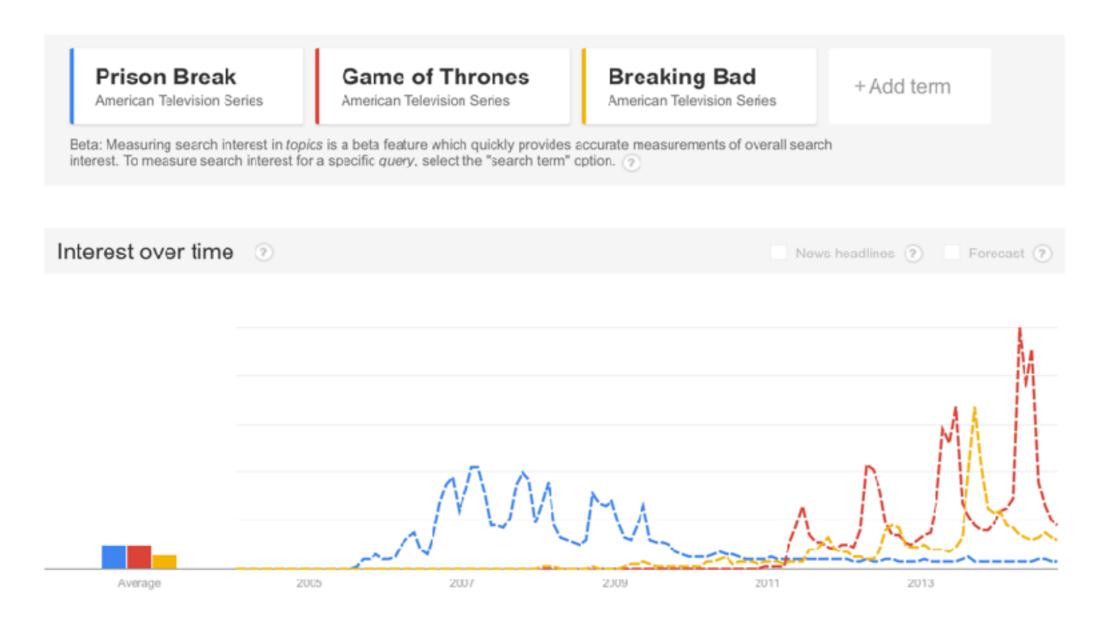
# Issues in Temporal IR

Temporal Collections



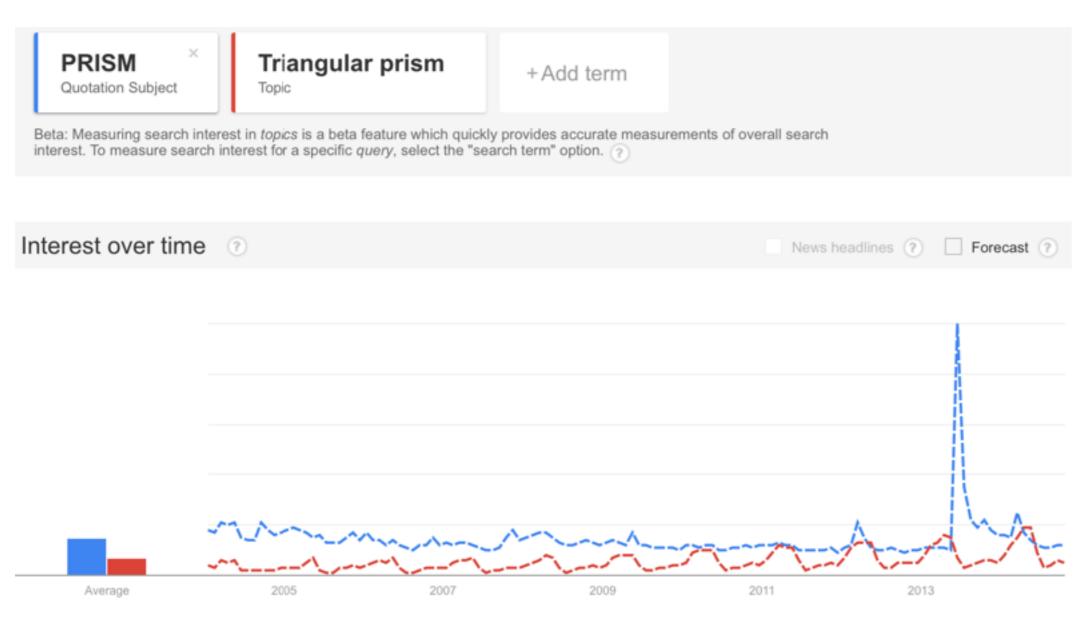
- Web Archives, News Archives, Wikipedia Versions....
- Temporal information needs nato intervention @ 25/12/2001
  - Historical vs Recency
- Document relevance with temporal relevance
  - Is the doc relevant vs Is the doc relevant at the time of interest?

# Issues in Temporal IR



Infer popular terms to improve ranking & suggestions

# Issues in Temporal IR

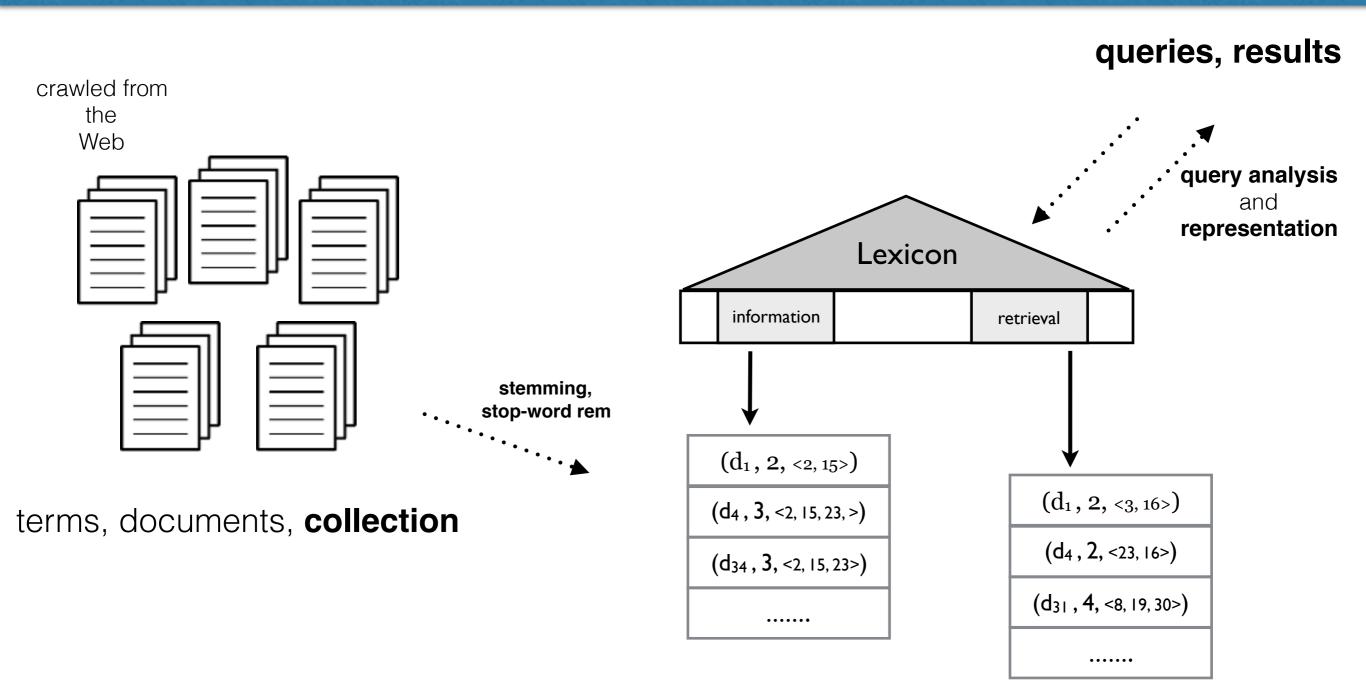


Infer popular terms to improve ranking & suggestions

# Course Outline

- Foundations of Temporal Analysis
- Indexing for Temporal Retrieval
- Retrieval Models
- Query Analysis
- Word Representations

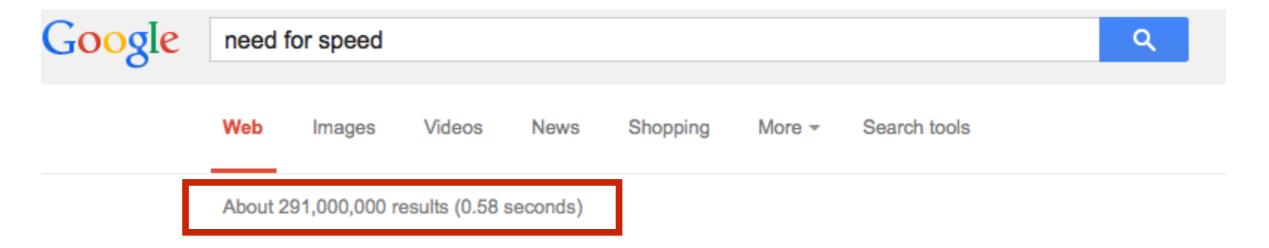
### Simplistic Overview - Web IR System



**Inverted Index** 

# Need for Speed





Sub-second retrieval



- Interactive search
- Search as a primitive for further mining and post processing

# Indexing Temporal Collections



- How do you search these collections efficiently?
- How do you build and maintain indexes over evolving text collections?
- What will you learn:
  - Indexing large text collections,
  - Compression and how to handle updates, QP
  - Software: Lucene

# Temporal Ranking



- Ranking a large set of results
- Rank results based on relevance to
  - Text in the document
  - Temporal intent
- Temporal Diversity
- Test Collections and Evaluation

# Temporal Ranking



- What will you learn?
  - Statistical Language Models
  - Support Vector Machines
  - Learning to Rank
  - Software: RankLib, SVMRank

# Query Understanding

- What does the user mean by the query?
  - Query: jaguar price

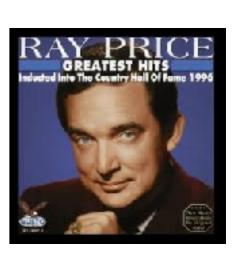






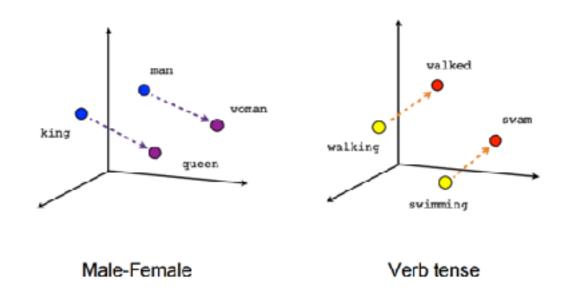




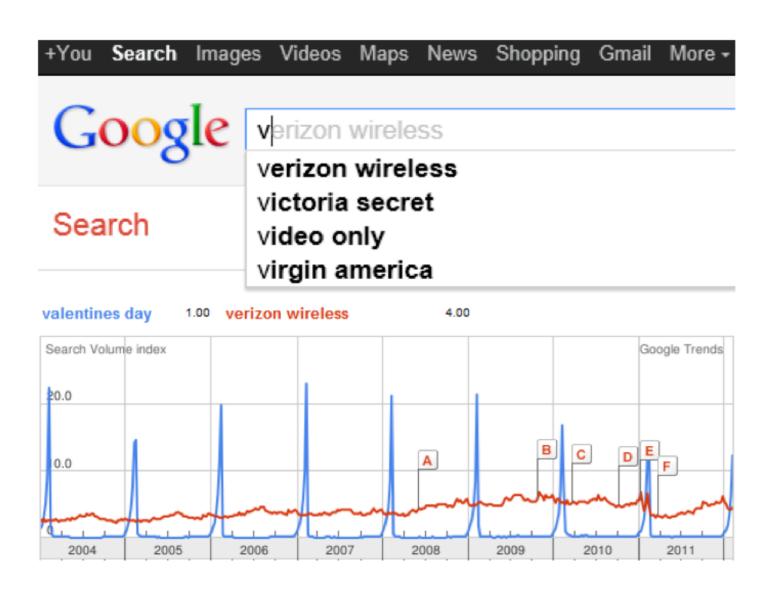


# Query Understanding

- Word embeddings
  - Similar words share similar contexts
  - Compositionality
- What will you learn?
  - Neural Networks (basics)
  - Software: Word2Vec



## Query Analysis



- Query auto-completion
- Query Suggestion
- Query Expansions

# Building a Temporal Search Engine

hannover HistDiv Results returned: 72 out of 72 Query: hannover - HistDiv The New York Times 1-6-2005 1 The New York Times 23-6-1987 2 The New York Times 24-2-2004 5 The New York Times 15-4-2003 3 The New York Times 22-3-2005 4 Director of Offshore A.I.G. Unit Blossoming Indoors in M.I.S.L. MUSIC: GERMAN CHOIR World Business Briefing | Europe: A Home With the MetroStars Was Cited in Australian Inquiry Germany: Insurer's Profit Rises Cultural Desk Sports Desk Sports Desk Business/Financial Desk Business/Financial Desk ...LEAD: THE Knabenchor ... about \$5 million with Hannover ...; and in Germany, striker Clint ... at a unit of the German reinsurer ...A leading reinsurance company, Hannover -which sang a program 96. BECKHAM STILL BENDING IT Mathis scored his third goal in four Hannover Re, was appointed to Hannover Rc, said it carned 267 from the German Baroque at St. - David Beckham, the English star games for Hannover 96. SOCCER. the board of Union Excess in June million euros (\$287 million) last REPORT Bartholomew's... Hannover who plays 2000... through several reinsurers, year, up from from 11 million euros which sang a program from the including Hannover Re, that made last year but short of its target of Baroque it appear financially stronger than it 300 million euros. Hannover said. Bartholomew's Church Friday was. The questionable transactions write-downs on the value of its night... Hannover -about 25 boys The New York Times 18-8-1990 6 The New York Times 30-8-2005 7 arranged by Hannover Re allowed capital investments hurt its 2002 and slightly fewer young men for Arts, Briefly HIH to overstate its before-tax...

Royal Commission. Mr. Elers declined repeated requests for

comment A molecularion for The New York Times 13-9-2005 8

Cost Estimates By Reinsurers Are Climbing

Business/Financial Desk

Hannover Re of Germany

the bass and tenor parts - brought to mind... "Der Geist hilft." The Knabenchor Hannover impresses

The New York Times 5-7-2005 9

In Germany, the Jobless Work to Keep Their Benefits

### Business/Financial Desk

...In January, Benjamin Dath's

results. The year-earlier income was held down by the Sept. 11 terrorist attacks in the United States. In its outlook for noon Hannorm sold

The New York Times 26-7-2006 10

2 Connected To Reinsurer May Face U.S. Action

### Business/Financial Desk

... included the American Re-

### GERMANS ARRESTED ON IRAQI EXPORTS

### Foreign Desk

... prosecutor, said the suspects were arrested in raids in Hamburg, Hannover and Dreieich, a small... Preussag A.G., a leading West German construction and heavy engineering concern in Hannover,

The New York Times 26-11-1080 11 The New York Times 5-11-1080 12

### The Arts/Cultural Desk

... reported. Mr. Oue, who leads the NDR Symphony Orchestra in Hannover, will be replaced by Peter Schneider

Move the mouse over the graph to reveal details. Click and Drag to select time interval.



# Assignments

- Tutorials:
  - Programming assignments
    - Languages: Python, Java
    - 2 weeks to solve
  - Theoretical questions
- **★** Assignment 1 [Indexing]
- ★ Assignment 2 [Retrieval Models LM]
- **★** Assignment 3 [Learning to Rank]
- **★** Assignment 4 [Word Embeddings]

# Final Exam

- Last lecture: 24th January
  - Conclusions and Exam revisions
- 30th January : Q&A
- Final Exam: Oral Exam
  - 1st Week of February



# Administrivia and Dates

- Webpage: <a href="http://www.l3s.de/~anand/tir">http://www.l3s.de/~anand/tir</a>
- Assignments, slides and references will be put up every week after the lecture
- Office hours: 30 mins after the lecture

# Demos

- Systems in action
  - Wayback Machine <u>link</u>
  - Portugese Web Archive <u>link</u>
  - British library <u>link</u>
  - L3S HistDiv System <u>link</u>