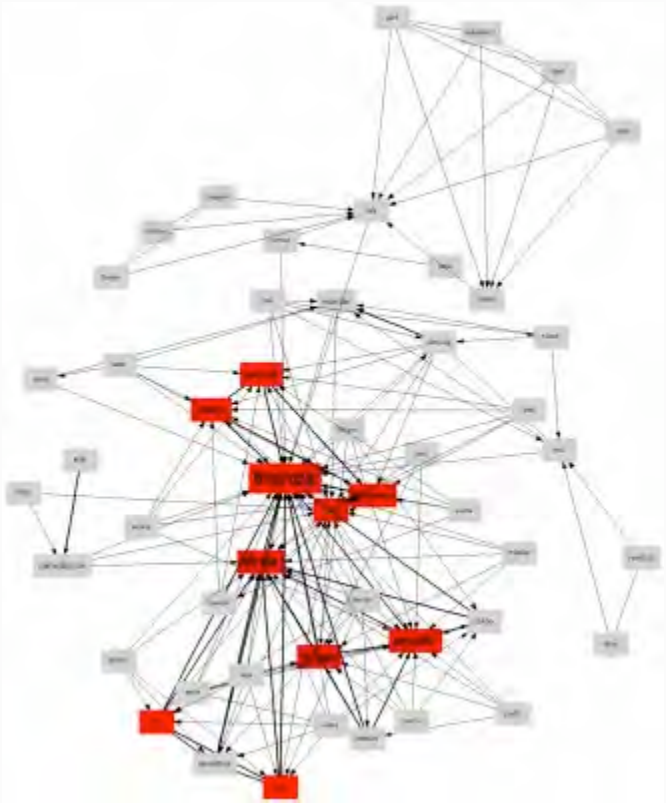


Sid #	Name	Year	GPA
1	Smith	3	3.0
2	Jones	2	3.5
3	Doe	1	1.2
4	Varda	4	4.0
5	Carey	4	0.5

Student Relation

Fid #	Name	Position	Dept
9	Henry	Prof.	Math
2	Jackson	Assist. Prof	Hist
14	Schuh	Assoc. Prof	Chem
21	Lerner	Assist. Prof	CS

Faculty Relation

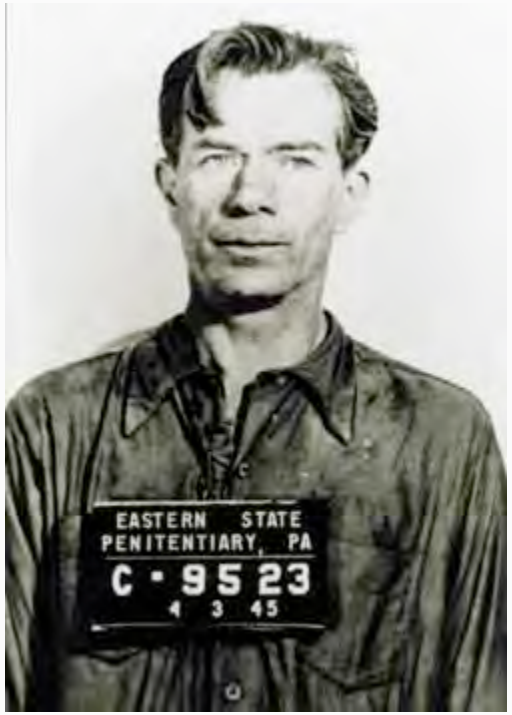
Course #	Course Name	Cr	Dept
223	Calculus	5	Math
302	Intro Prog	3	CS
302	Organic Chem	3	Chem
542	Asian Hist	2	Hist
222	Calculus	5	Math

Course Relation

Extracting Semantics from the Web

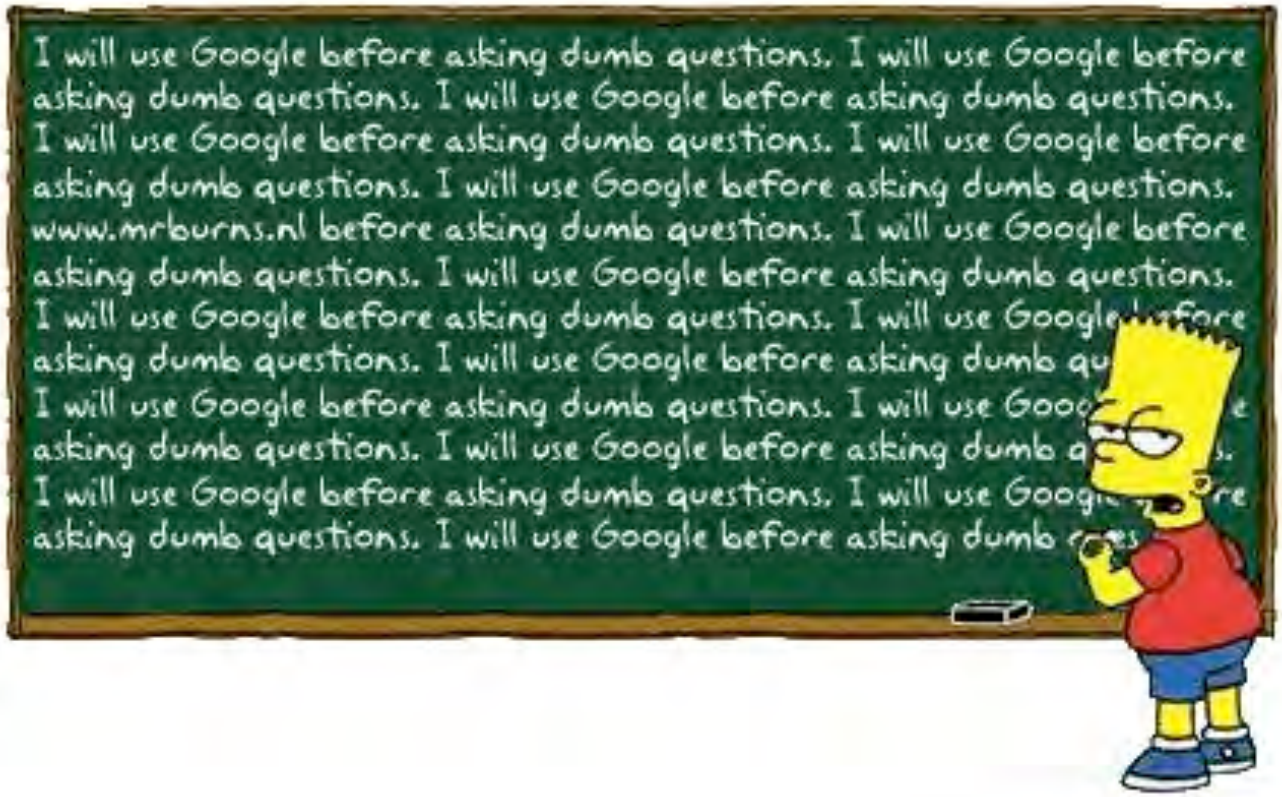
Peter Norvig

Google



Willie Sutton

Google


I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. www.mrburns.nl before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions. I will use Google before asking dumb questions.

Google

What to Extract

FROM

- Unstructured text
- Semi-structured
- Structured databases
- Link patterns
- Usage patterns
- …

TO

- Probability distributions
- Info. retrieval models
- Language models
- Relational databases
- Semantic networks
- …

Google

Effect of large corpus size

Banko & Brill, 2001

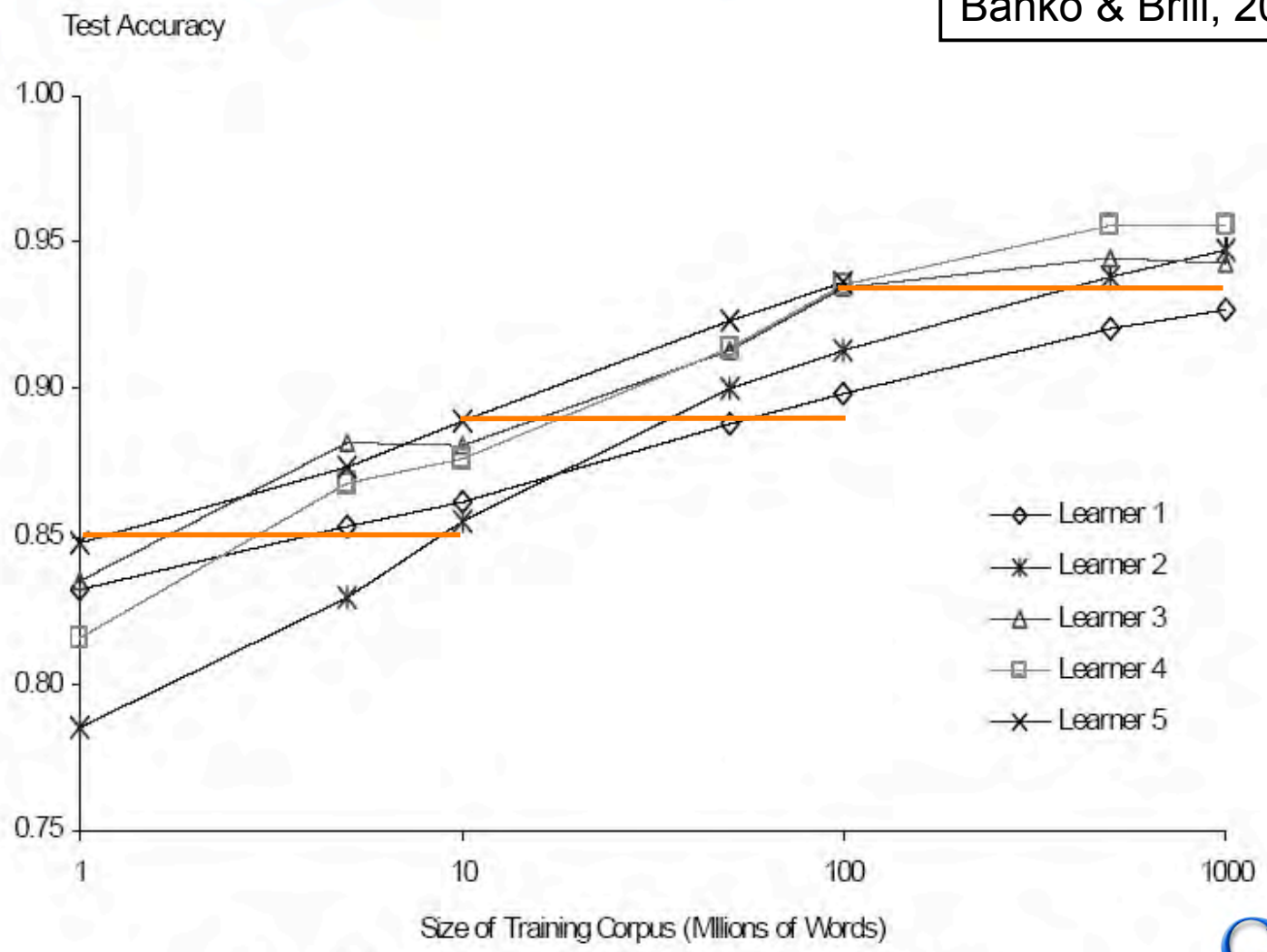
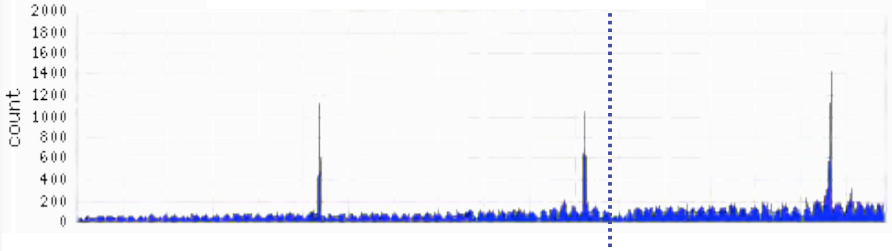


Figure 2. Learning Curves for Confusable Disambiguation

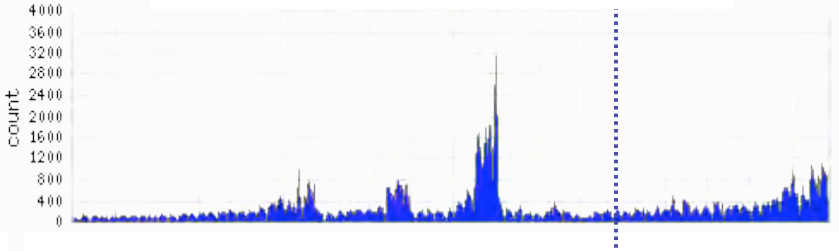
Google

Extraction from usage patterns

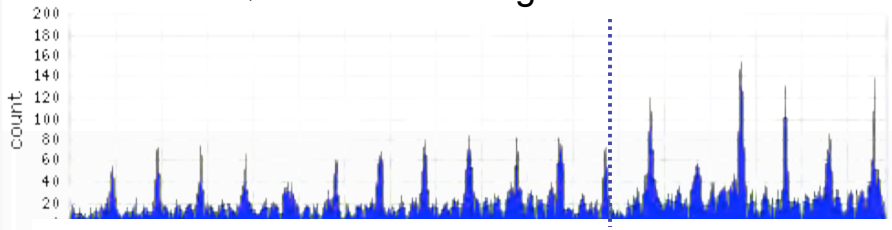
Queries containing “eclipse”



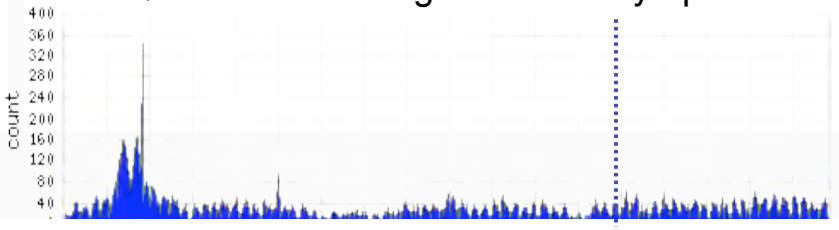
Queries containing “world series”



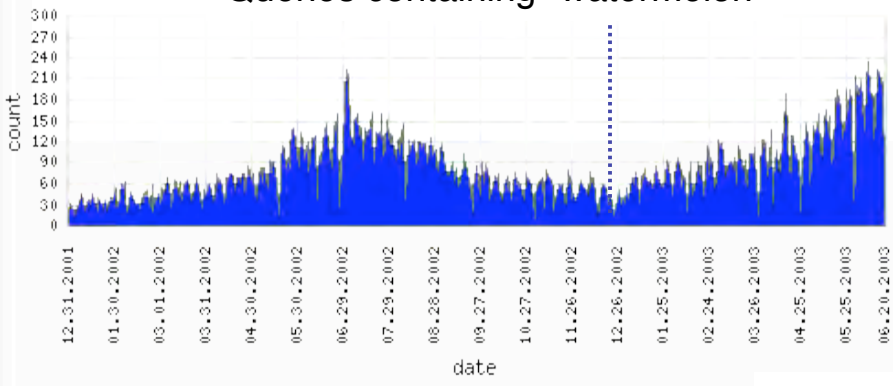
Queries containing “full moon”



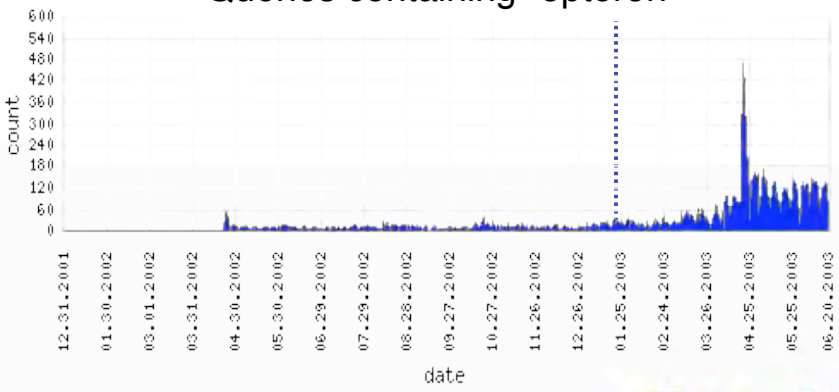
Queries containing “summer olympics”



Queries containing “watermelon”



Queries containing “opteron”



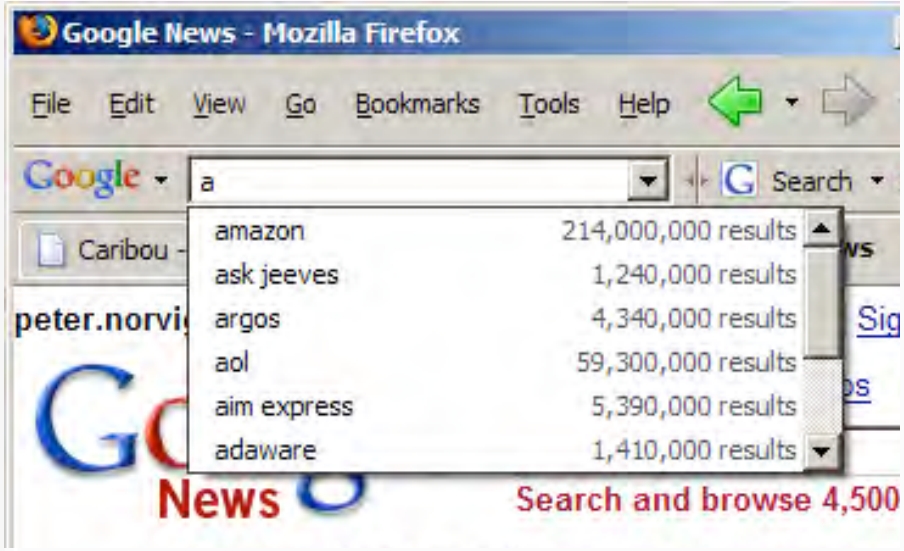
Google

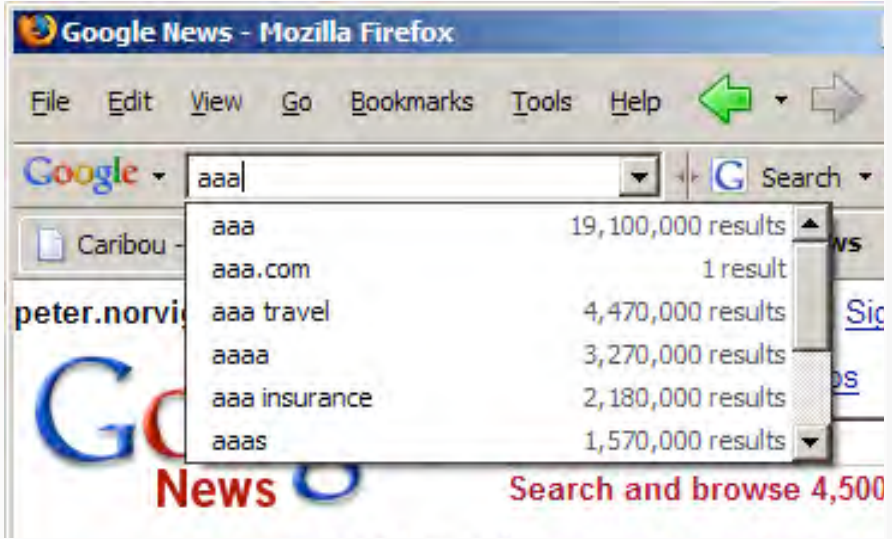
Spelling model from usage patterns

Searching for Britney Spears…

Google

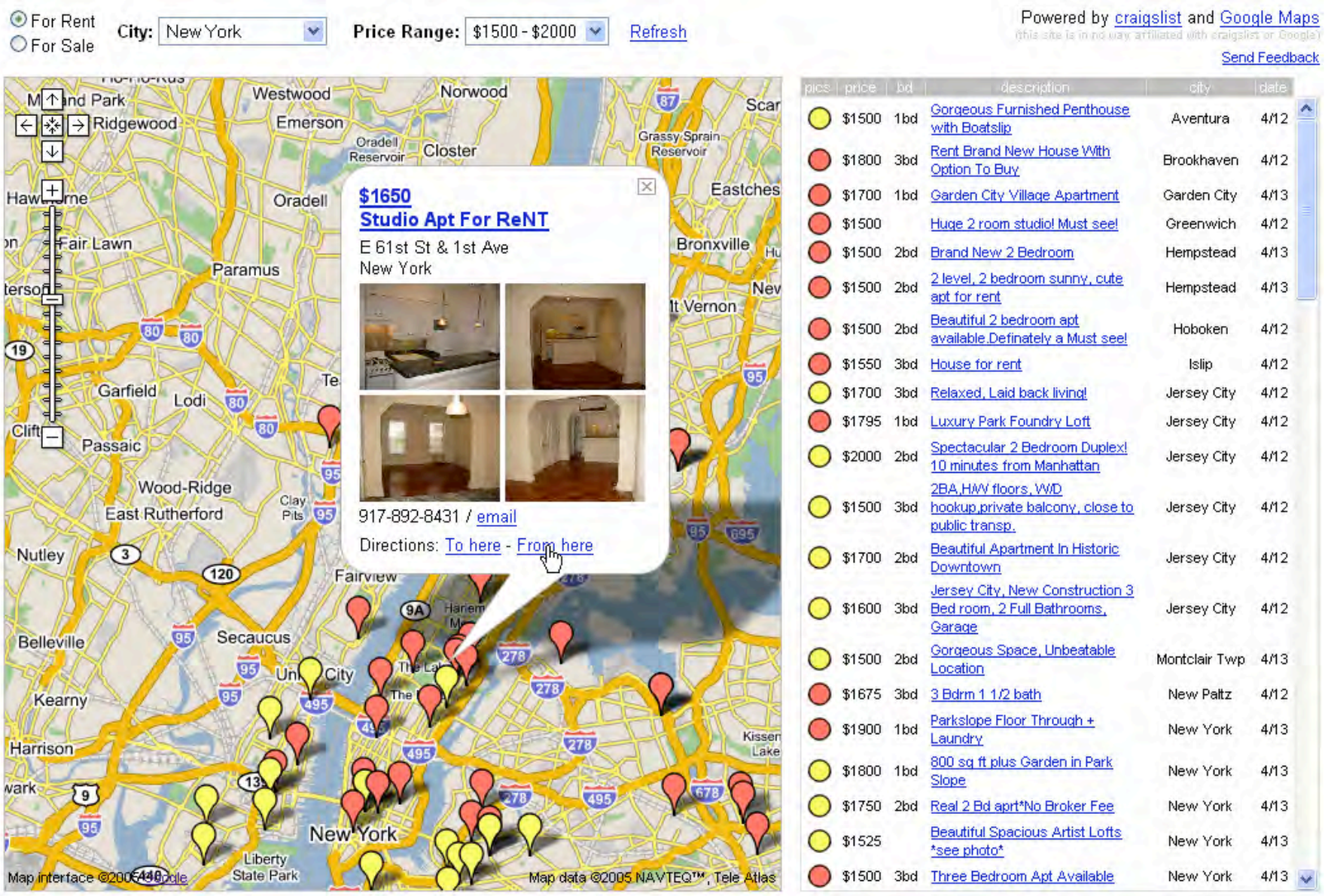
Suggestions





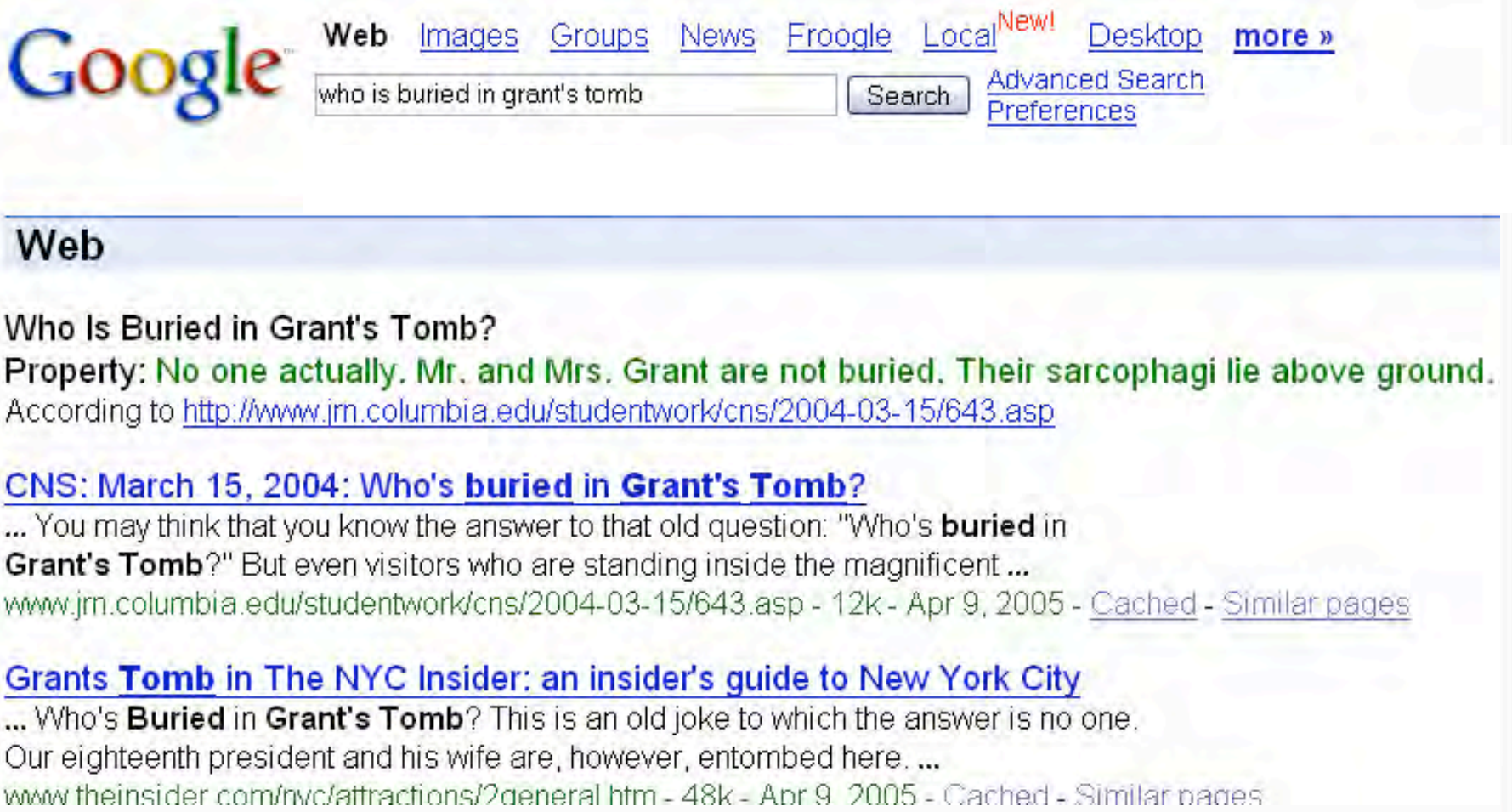
Google

Extraction and remixing



Extraction from semi-structured text

Extraction from semi-structured text



Named Entity Extraction (Pasca et al.)

1. Filter HTML, break text into sentences, and mark POS with tagger (Brants' TnT).
2. Match sentences against patterns like
   … *C* ["such as" | "including"] *N* …
3. Retain (*N* isa *C*) if *C* is simple, ends in a plural noun.
4. Discard noisy data; regularize over names.
5. Plug (*N* isa *C*) pairs back into text to find new patterns. Iterate.

Google

Named Entity Extraction

- Hybrid cars: Toyota Prius, Honda Insight
- High-speed networks: ATM, Gigabit Ethernet, B-ISDN, Myrinet, Frame Relay, Fast Ethernet, …
- Rappers: Eminem, Jay-Z, Nas, Dmx, Snoop Dogg, Dr. Dre, Ja Rule, Mos Def, Nelly, 50-cent …
- Anti-depressants: Prozac, Zoloft, Paxil, Wellbutrin, Effexor, Elavil, Luvox, SSRIs, …

Google

Statistical Machine Translation

- Data: aligned parallel corpus
- Model: $\Pr(e|f) = p_{\lambda_1^M}(e|f) \propto \exp[\sum_{m=1}^{M} \lambda_m h_m(e,f)]$

- Training: maximize BLEU scores directly
- Inference: beam search
- Evaluation: BLEU

Google

Statistical Machine Translation

- Model features (h_m):
  - All based on word[i..j] statistics
- **Not** used:
  - Syntax: parser, chunker, POS tagger
  - Semantics: Wordnet, ontologies
  - Annotated data: TreeBank, FrameBank

Google

Statistical Machine Translation

- ركذي ان سلجم نمال وه ديحولا لوخملا ضرف تابوقعدض ايروك ةيلاملشا يتلا ترذح نم اهنا ربتعتسلت بوقعلا ةباثمب نالعابرح.

- صربقو ةمسقم ىلا نيرطش ذنم ماعلا1974 امدنعل خا شيجلا يكرتلا ثلثلا يلاملشا نم ةريزجلا ادر ىلعبلقنا هذفن نوياموق ةصرابق فدهب مضةريزجلا ىلانانويلا.

- ركذي ان ناريا يتلا تضاخ ابرح دض قارعلا ترمتساينا مث تاونس(1980-1988) ضراعت نش موجه يركسع يكريماعلا قارعلا يذلا همهتت نطنشاو كالتمابةحلسا رامدللها نئا هنابوطبترمباهرالاب.

Statistical Machine Translation

- It is noteworthy that the Security Council is the only authorized to impose sanctions against North Korea, which warned that it would consider sanctions a declaration of war.

- Cyprus has been divided into two parts since the year 1974 when the Turkish army invaded the northern third of the island in response to a coup by Greek nationalists with the aim of annexing the island to Greece.

- It is worth mentioning that Iran, which fought a war against Iraq lasted eight years (1980 - 1988) opposes American military attack on Iraq, which Washington accuses of possessing weapons of mass destruction and that it was linked to terrorism.

Google

Statistical Machine Translation

- 新华社大马士革4月15日电(记者拱振喜)叙利亚总统巴沙尔·阿萨德15日在此间与来访的美国国务卿鲍威尔举行了会谈，双方讨论了中东局势的最新发展，特别是巴勒斯坦的严重局势以及黎以边界地区的紧张局势等问题。

- 叙利亚通讯社报道，巴沙尔总统在会谈中说："在巴勒斯坦发生的事件使(中东)和平进程走进了死胡同，如果不能认识到这一点，事情的发展有可能达到无法挽回的程度，那时，我们只能再等待一代人的时间。"

- 他指出，只有在以色列从它占领的巴勒斯坦领土撤军，停止屠杀巴勒斯坦人以后，才可以谈和平进程的问题。

Google

Statistical Machine Translation

- Xinhua News Agency, Damascus, April 15 (Reporter Gong Zhenxi) Syrian President Bashar Assad 15th here with visiting US Secretary of State Colin Powell held talks, the two sides discussed the latest development of the situation in the Middle East, especially the serious situation in Palestine and the tension in the border region between Lebanon and Israel and other issues.

- According to the Syrian News Agency, President Bashar during the talks, said: "In the incident to the Palestinian (Middle East) peace process into a dead end, if not realize that this is happening may not be able to restore to the extent that time, we can only wait for the generation of time."

- He pointed out that only in the Israeli withdrawal from the occupied Palestinian territories, stop massacre of Palestinians, can talk about the peace process.

Google

Results, NIST competition

Arabic-to-English Task, *Large* Data Track

Table 1	
Site	**BLEU-4 Score**
GOOGLE	0.5131
ISI	0.4657
IBM	0.4646
UMD	0.4497
JHU-CU	0.4348
EDINBURGH	0.3970
SYSTRAN	0.1079
MITRE	0.0772
FSC	0.0037

Arabic-to-English Task, *Unlimited* Data Track

Table 2	
Site	**BLEU-4 Score**
GOOGLE	0.5137
SAKHR	0.3403
ARL	0.2257

Chinese-to-English Task, *Large* Data Track

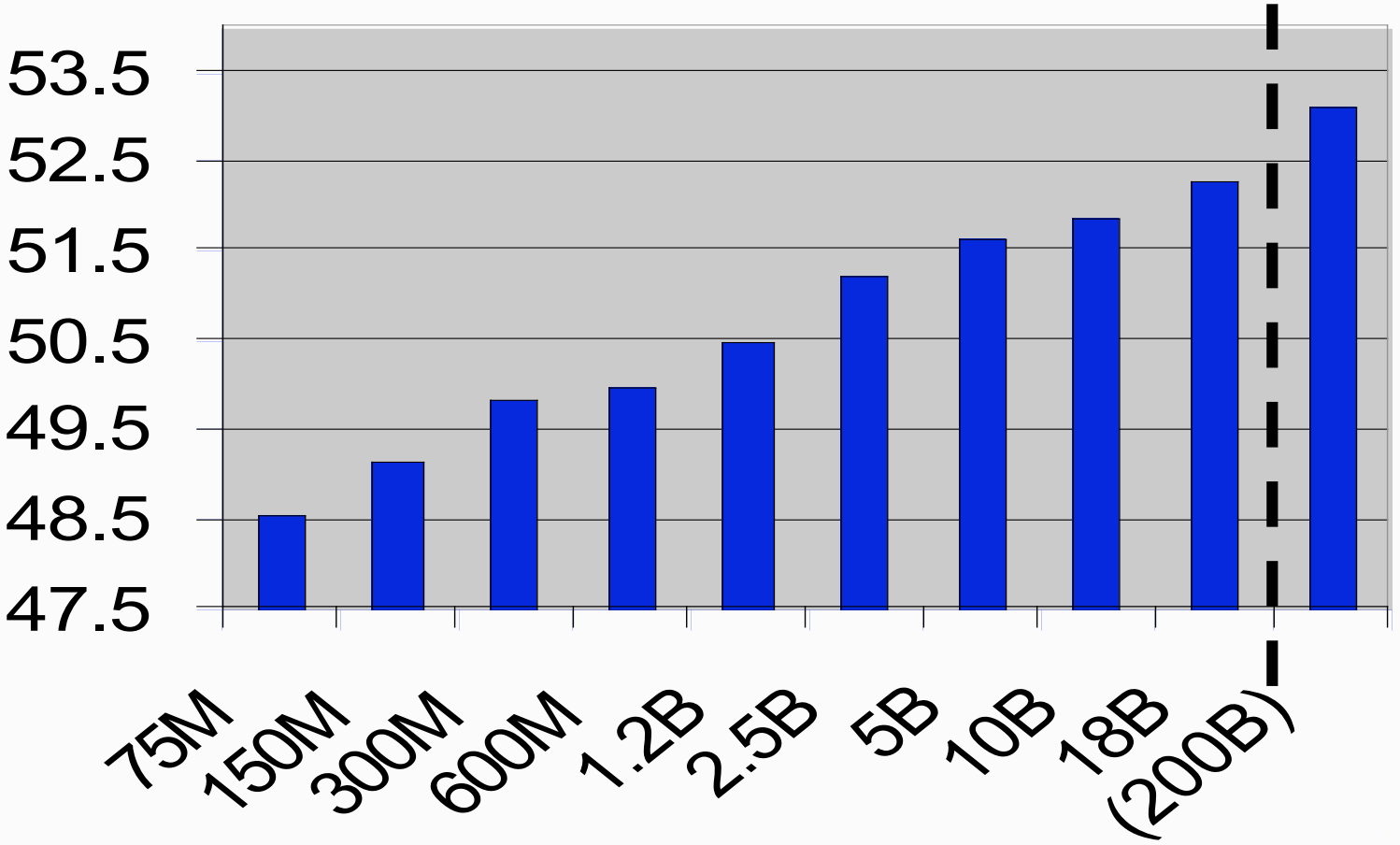
Table 4	
Site	**BLEU-4 Score**
GOOGLE	0.3531
ISI	0.3073
UMD	0.3000
RWTH	0.2931
JHU-CU	0.2827
IBM	0.2571
EDINBURGH	0.2513
ITCIRST	0.2445
NRC	0.2323
NTT	0.2321
ATR	0.1822
SYSTRAN	0.1471
SAAR	0.1310
MITRE	0.0542

Chinese-to-English Task, *Unlimited* Data Track

Table 5	
Site	**BLEU-4 Score**
GOOGLE	0.3516
ICT	0.1293
HIT	0.0797

Google

More Data Helps



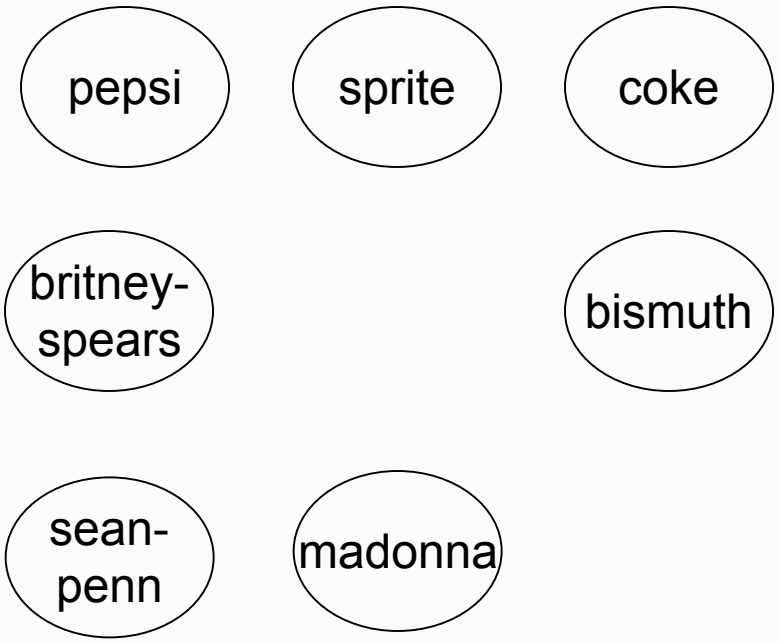
Five-gram language model, no count-cutoff, integrated into search

Google

Questions?

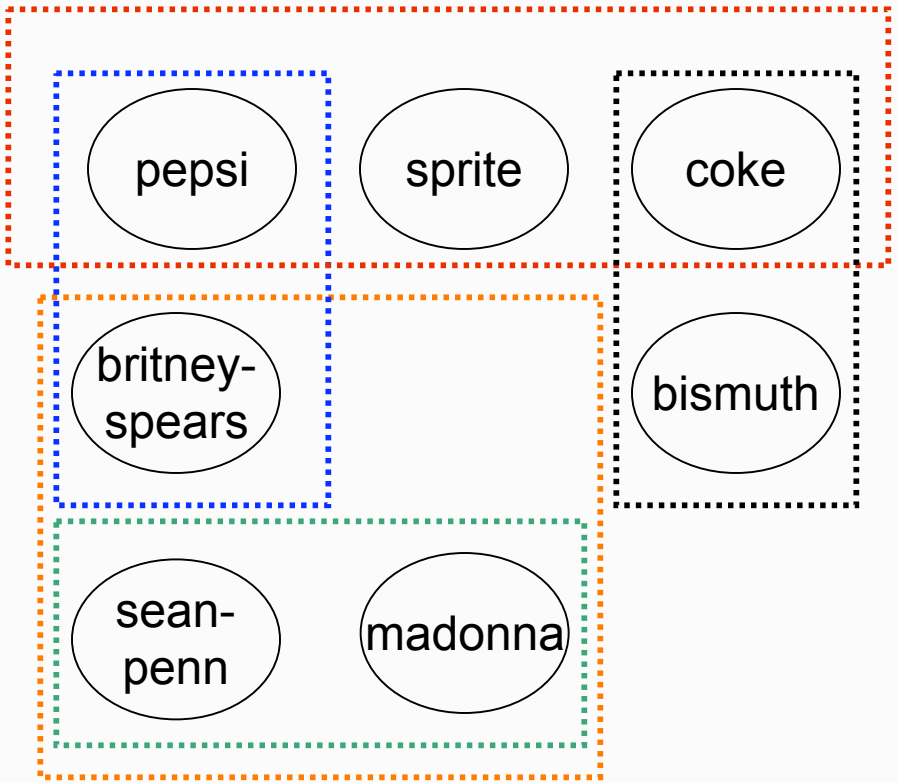
Google

Word Clustering



Google

Word Clustering



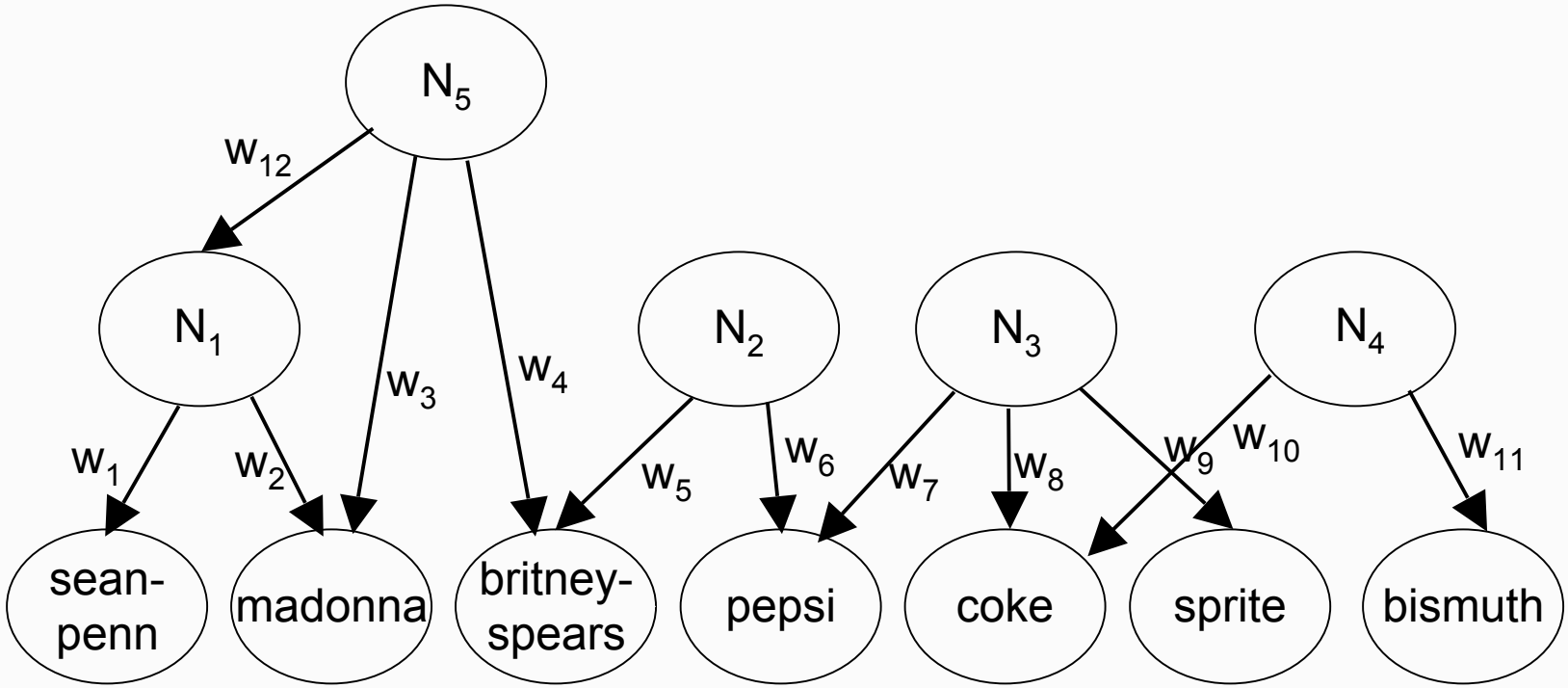
Google

Word Clustering

- Data: Billions of short word subsequences
- Model: Noisy-Or Bayes Net
  - Millions of terminal nodes (words)
  - 100,000s of non-terminals (topics) (Sparse.)
- Training: EM
- Inference: Greedy Deterministic Sampling
- Evaluation: ???

Google

Word Clustering



Google

/search?q=visa

visa Search

QUERY: visa

visa 15.624945/4.039471/8.826189 [TOP]

Probability	Activation	Cluster
0.582235	0.525590	[visa for visas uk to tourist]
0.359859	0.214223	[credit-card capital-one credit-cards american-express mbna visa]
0.149869	0.124134	[visa card platinum first first-usa united-mileage-plus]
0.051598	0.042750	[visa 三井-住友 カード 住友 viza マスター]
0.097412	0.012965	[ins green-card visa us status application]
0.021811	0.012400	[visa requirements embassy tourist requirement citizens]
0.006395	0.005442	[visa pour formulaire france demande visas]
0.006069	0.003239	[visa h1 h4 l1 h1b b1]
0.042917	0.001117	[カード jcb nicos セゾン-カード dc 日本-信販]
1.000000	0.000050	TOP
0.582595	0.000000	[in embassy consulate usa us-embassy us]
0.040458	0.000000	L.ja

Query matched 13 clusters

Google

/search?q=visa+h1b

visa h1b Search

QUERY: visa h1b
h1b 19.241566/2.164040/12.102268 [TOP]
visa 14.624947/1.298397/8.826189 [TOP]

Probability	Activation	Cluster
0.877054	1.613480	[h1b visa h1 usa h-1b sponsor]
0.104648	0.281994	[visa h1 h4 l1 h1b b1]
0.203830	0.240929	[h1b h1-b visa h1 transfer h-1b]
0.018289	0.051459	[visa h1 sponsorship h1visa h1b sponsoring]
0.061363	0.008124	[ins green-card visa us status application]
0.023886	0.003838	[visa us-consulate us chennai consulate us-embassy]
1.000000	0.000021	TOP
0.023871	0.000000	[india indian in of delhi bombay]
0.868206	0.000000	[jobs job in employment job-search careers]
0.008232	0.000000	[yellow-pages white-pages area-codes zip-codes mapquest people-search]
0.011326	0.000000	L.en-ca
0.047433	0.000000	L.es
0.083118	0.000000	L.de

Query matched 15 clusters

Google

/search?q=visa+card

visa card Search

QUERY: visa card

card 5.355414/2.596364/8.062965 [[credit-card capital-one credit-cards american-express mbna v

visa 5.221890/1.154132/8.826189 [[credit-card capital-one credit-cards american-express mbna v

Probability	Activation	Cluster
0.931197	1.830076	[visa card platinum first first-usa united-mileage-plus]
0.050427	0.117189	[card debit-card cards visa debit-cards gift]
0.952259	0.042121	[credit-card capital-one credit-cards american-express mbna visa]
0.014113	0.030300	[card visa credit-card master-card card-services nederland]
0.029524	0.000450	[payment online payments services electronic pay]
1.000000	0.000338	TOP
0.008299	0.000073	[miles frequent-flyer frequent-flyer-miles program reward rewards]
0.014806	0.000000	[abbey-national natwest lloyds-tsb hsbc barclays bank]
0.013837	0.000000	L.en-ca
0.016587	0.000000	L.en-au
0.018484	0.000000	L.it
0.057745	0.000000	L.es
0.013925	0.000000	L.en-gb
0.049068	0.000000	L.ja
0.101071	0.000000	L.de

Query matched 17 clusters

Google™

/search?q=visa+application

visa application Search

QUERY: visa application
visa 14.624947/2.741010/8.826189 [TOP]
application 13.903603/5.769062/8.515065 [TOP]

Probability	Activation	Cluster
0.802836	1.679451	[visa for visas uk to tourist]
0.172430	0.513658	[credit-card capital-one credit-cards american-express mbna visa]
0.178909	0.091250	[ins green-card visa us status application]
0.024841	0.084685	[credit-cards credit-card visa secured application canadian]
0.046922	0.036228	[visa card platinum first first-usa united-mileage-plus]
0.012266	0.005756	[passport us passport-application passport-renewal passports renew]
1.000000	0.001210	TOP
0.024840	0.000000	[credit first peoples chexsystems banks chex-systems]
0.803012	0.000000	[in embassy consulate usa us-embassy us]
0.012100	0.000000	L.en-ca
0.014410	0.000000	L.en-au
0.055389	0.000000	L.en-gb

Query matched 14 clusters

Google