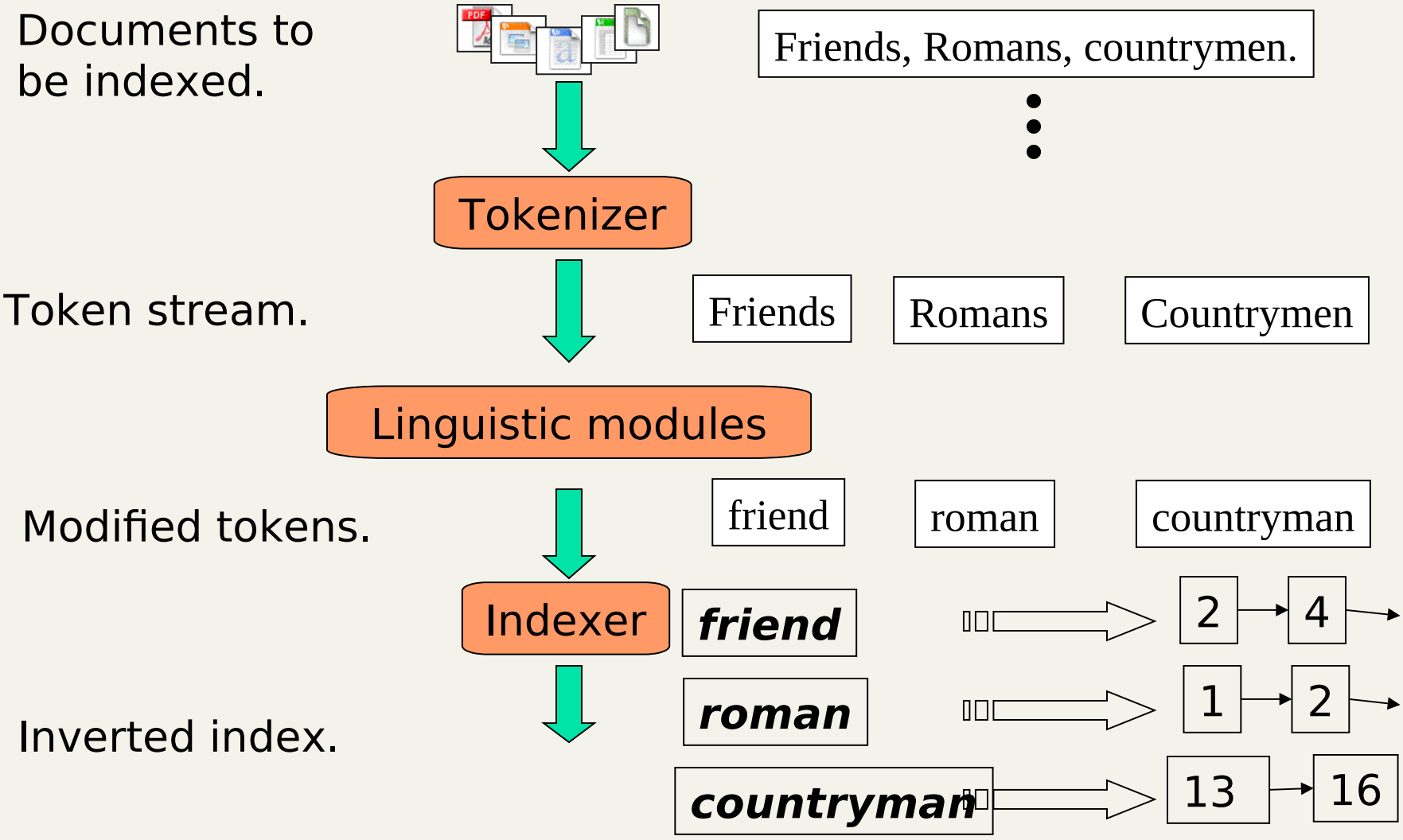
Document Parsing

Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

Inverted index construction



Google friend roman countryman

Search About 1,740,000 results (0.39 seconds)

Web
Images
Maps
Videos
News
Shopping
More

Show search tools

Friends Romans Countrymen...
www.angelfire.com/moz/useless/**Romans**.html
"**Friends**, **Romans**, **Countryman**, lend me your ears; I come to bury Caesar not to praise him. The evil that men do lives after them, the good is oft interred with ...

Monologue: **Friends**, **Romans**, **Countryman** Speech from William ...


www.youtube.com/watch?v...
6 Jun 2010 - 2 min - Uploaded by th3m0vingshad0w
Yay for my first VA attempt. I decided to go with probably one of my favorite plays and monologues of all time ...

More videos for **friend roman countryman** »

Friends, **Romans**, countrymen, lend me your ears - Wikipedia, the ...
en.wikipedia.org/.../**Friends**,_**Romans**,_countrymen,_lend_me...
Friends, **Romans**, countrymen, lend me your ears is the first line of a famous and often-quoted speech by Mark Antony in the play Julius Caesar, by William ...
About the speech - Setting - Relevance and cultural impact - External links

SCENE II. The Forum.
shakespeare.mit.edu/julius_caesar/julius_caesar.3.2.html
was no less than his. If then that **friend** demand why Brutus rose against Caesar, this is my answer: --Not that I loved Caesar less, but that I loved **Rome** more.

Parsing a document

- What format is it in?
  - pdf/word/excel/html?
- What language is it in?
- What character set is in use?

Each of these is a **classification problem**.

But these tasks are often done heuristically …

Tokenization

- Input: "***Friends, Romans and Countrymen***"
- Output: Tokens
  - ***Friends***
  - ***Romans***
  - ***Countrymen***
- A token is an instance of a sequence of characters
- Each such token is now a candidate for an index entry, after further processing

- But what are valid tokens to emit?

Tokenization: terms and numbers

- Issues in tokenization:
  - ***Barack Obama:*** one token or two?
  - ***San Francisco?***
  - ***Hewlett-Packard***: one token or two?
  - ***B-52, C++, C#***
  - ***Numbers ? 24-5-2010***
  - ***192.168.0.1***

Google

san paolo

Search

About 27,100,000 results (0.28 seconds)

Web

Images

Maps

Videos

News

Shopping

More

Show search tools

Tip: Search for **English** results only. You can specify your search language in Preferences

See results

Servizi bancari e consulenza per famiglie e ... - Intesa **Sanpaolo**
www.intesa**sanpaolo**.com/.../RetailIntesa**San**... - Translate this page
I servizi bancari e la consulenza di Intesa **Sanpaolo** per famiglie e imprese: conto corrente, bancomat, carte di credito, prestiti, internet banking, investimenti, ...

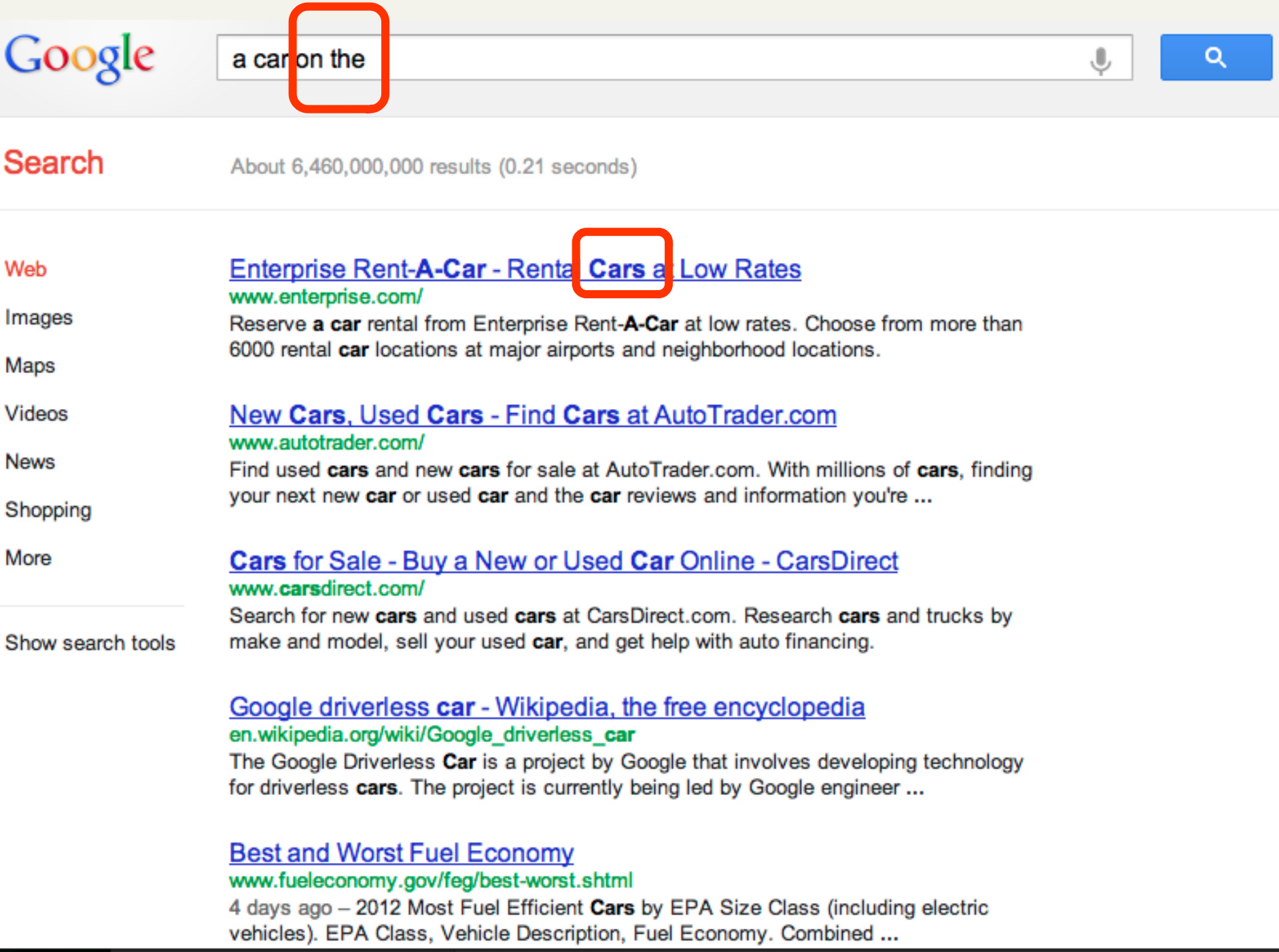
Intesa **SanPaolo** SpA
www.intesa**sanpaolo**.com/ - Translate this page
Presenta il gruppo, illustra profilo, prodotti e servizi online ed offline.
Intesa San Paolo - Intesa Sanpaolo. - Servizi bancari e assistenza ... - Carte
You've visited this page many times. Last visit: 4/6/12

São Paulo - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**São_Paulo**
São Paulo is the largest city in Brazil, the largest city in the southern hemisphere and Americas, and the world's seventh largest city by population.
São Paulo FC - São Paulo (state) - São Paulo-Guarulhos ... - Santo André

Intesa **Sanpaolo** Private Banking
www.intesa**sanpaolo**privatebanking.it/ - Translate this page
Intesa **Sanpaolo** Private Banking: la vostra banca personale d'investimento, per proteggere, accrescere e accompagnare nel tempo il vostro patrimonio.
You've visited this page many times. Last visit: 8/2/12

Stop words

- We exclude from the dictionary the most common words (called, stopwords). Intuition:
  - They have little semantic content: *the, a, and, to, be*
  - There are a lot of them: ~30% of postings for top 30 words
- But the trend is away from doing this:
  - Good compression techniques (lecture!!) means the space for including stopwords in a system is very small
  - Good query optimization techniques (lecture!!) mean you pay little at query time for including stop words.
  - You need them for phrase queries or titles. E.g., “As we may think”


Google
a car on the
Search
About 6,460,000,000 results (0.21 seconds)
Web
Images
Maps
Videos
News
Shopping
More
Show search tools
Enterprise Rent-A-Car - Rental Cars at Low Rates
www.enterprise.com/
Reserve a car rental from Enterprise Rent-A-Car at low rates. Choose from more than 6000 rental car locations at major airports and neighborhood locations.
New Cars, Used Cars - Find Cars at AutoTrader.com
www.autotrader.com/
Find used cars and new cars for sale at AutoTrader.com. With millions of cars, finding your next new car or used car and the car reviews and information you're ...
Cars for Sale - Buy a New or Used Car Online - CarsDirect
www.carsdirect.com/
Search for new cars and used cars at CarsDirect.com. Research cars and trucks by make and model, sell your used car, and get help with auto financing.
Google driverless car - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Google_driverless_car
The Google Driverless Car is a project by Google that involves developing technology for driverless cars. The project is currently being led by Google engineer ...
Best and Worst Fuel Economy
www.fueleconomy.gov/feg/best-worst.shtml
4 days ago – 2012 Most Fuel Efficient Cars by EPA Size Class (including electric vehicles). EPA Class, Vehicle Description, Fuel Economy. Combined ...

Normalization to terms

- We need to “normalize” terms in indexed text and query words into the same form
  - We want to match ***U.S.A.*** and ***USA***
- We most commonly implicitly define equivalence classes of terms by, e.g.,
  - deleting periods to form a term
    - ***U.S.A., USA → USA***
  - deleting hyphens to form a term
    - ***anti-discriminatory, antidiscriminatory → antidiscriminatory***
- ***C.A.T. → cat ?***

Google C.A.T.

Search 5 personal results. 2,650,000,000 other results.

Web
Images
Maps
Videos
News
Shopping
More

Show search tools

Cat - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Cat**
The domestic **cat** (Felis catus or Felis silvestris catus) is a small, usually furry, domesticated, carnivorous mammal. It is often called the housecat when kept as an ...
List of cat breeds - Cat intelligence - Behavior - Feral cat

Cat Products & Services
www.**cat**.com/
Cat machines & engines set the standard for the industries we serve. Our extensive product line reflects our increased focus on our customers' success.
+ Show stock quote for CAT
Cat Products - Cat Dealer Locator - Parts & Service - About The Company

CAT
www.**cat**iim.in/
Registration for **CAT** 2012 is now closed. Registered candidates may log on to https://iim.prometric.com to print a copy of their Admit Card until the end of the ...
Iim cat result - CAT Eligibility - Selection Process of IIMs - CAT 2012 Test Sites

CAT: Summary for Caterpillar, Inc. Common Stock- Yahoo! Finance
finance.yahoo.com/q?s=**CAT**
2 hours ago – View the basic **CAT** stock chart on Yahoo! Finance. Change the date range, chart type and compare Caterpillar, Inc. Common Stock against ...

See result

Case folding

- Reduce all letters to lower case
  - exception: upper case in midsentence?
    - e.g., ***General Motors***
    - ***SAIL*** vs. ***sail***
    - ***Bush*** vs. ***bush***
- Often best to lower case everything, since users will use lowercase regardless of ‘correct’ capitalization…

bush

Search

About 449,000,000 results (0.43 seconds)

Web

Images

Maps

Videos

News

Shopping

Blogs

More

Show search tools

George W. Bush - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/George_W._Bush
George Walker **Bush** (born July 6, 1946) is an American politician and businessman who was the 43rd President of the United States from 2001 to 2009 and the ...
George Bush - Bush administration - Jenna Bush Hager - Laura Bush

Bush (band) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Bush_(band)
Bush is a rock band formed in London in 1992 shortly after vocalist/guitarist Gavin Rossdale and guitarist Nigel Pulsford met. It was not long before they recruited ...
Discography - Sixteen Stone - Razorblade Suitcase - The Sea of Memories

BUSH Official Website
www.bushofficial.com/
Official Site for **BUSH**. Music of Gavin Rossdale, Chris Traynor, Corey Britz, Robin Goodridge.
Tour - Store - Photos - Video

Decision Points by George W. Bush
www.georgewbush.com/
Shattering the conventions of political autobiography, Decision Points by George W. **Bush** offers a strikingly candid journey through the defining decisions in the ...

Thesauri

- **Do we handle synonyms and homonyms?**
  - E.g., by hand-constructed equivalence classes
    - ***car*** = ***automobile*** ***color*** = ***colour***
  - We can rewrite to form equivalence-class terms
    - When the document contains ***automobile***, index it under ***car-automobile*** (and vice-versa)
  - Or we can expand a query
    - When the query contains ***automobile***, look under ***car*** as well

Google automobile

Search About 492,000,000 results (0.28 seconds)

Web
Images
Maps
Videos
News
Shopping
Books
Blogs
More

Show search tools

Ad related to **automobile**

automobile.it - Offerte Auto Usate e Km 0.
www.**automobile**.it/
Scopri il nuovo **automobile**.it di eBay

Auto Usate Auto Nuove
Auto Km 0 Vendi la tua Auto

Automobile - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Automobile** Share
It shows the significant growth in BRIC. World map of passenger cars per 1000 people. An **automobile**, autocar, motor car or car is a wheeled motor vehicle used ...
History of the automobile - Karl Benz - List - Crossover

New Cars & Car Reviews, Concept Cars & Auto Shows - **Automobile** ...
www.**automobile**mag.com/
Find new cars as well as in-depth ca reviews, photos, videos, and the latest concept cars from auto shows across the world at **Automobile** Magazine. Research a ...
Car Reviews - Rumors - Used Cars - Contact Us

New Cars, Used Cars, Car Reviews and Pricing - Edmunds.com
www.edmunds.com/
Edmunds car buying guide lists new car prices, used car prices, car comparisons, car buying advice, car ratings, car values, auto leasing.

People relate

Porter's algorithm

Stemming

- Reduce terms to their "roots" before indexing
- "Stemming" suggest crude affix chopping
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

automated production

Search

About 9,260,000 results (0.39 seconds)

Web

Images

Maps

Videos

News

Shopping

More

Show search tools

Automation - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Automation**
Automation is the use of machines, control systems and information technologies to optimize productivity in the **production** of goods and delivery of services.
Category:Industrial automation - Building automation - Pop music automation

AP | Automated Production Systems
www.**automatedproduction**.com/
Automated Production Systems is a world class manufacturer and provider of swine, dairy, and horticulture equipment and services throughout the world.
Swine Systems - Manuals - Sales & Support - About AP

Home | **Automated Production**
automatedproduction.biz/
Automated Production specializes in DNV design and fabrication projects including reels, lift frames, positioning systems, baskets, living quarters, storage ...

What are **automated production** systems
wiki.answers.com › ... › Engineering › Industrial Engineering
Automated production systems consist of automated workstations connected by a material handling system whose actuation is coordinated with the stations.

Lemmatization

- Reduce inflectional/variant forms to base form
- E.g.,
  - *am, are, is* → *be*
  - *car, cars, car's, cars'* → *car*
- Lemmatization implies doing “proper” reduction to dictionary headword form

to be or not to are

Search

About 23,790,000 results (0.34 seconds)

Web

Images

Maps

Videos

News

Shopping

More

Show search tools

To Be or Not to Be (1942 film) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**To_Be_or_Not**_to_Be_(1942_film)
To Be or Not to Be is a 1942 American comedy directed by Ernst Lubitsch, about a troupe of actors in Nazi-occupied Warsaw who use their abilities at disguise ...

To be or not to be (Shakespeare) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**To_be_or_not**_to_be_(Shakespeare)
"**To be or not** to be" is the opening phrase of a soliloquy in William Shakespeare's play Hamlet. It is perhaps the most famous of all literary quotations but there is ...

Cell - Diabetic β Cells: **To Be or Not** To Be?
www.cell.com/abstract/S0092-8674(12)01019-7
14 Sep 2012 – Diabetic β Cells: **To Be or Not** To Be? Summary; Main Text · References; Comments (0). To view the full text, please login as a subscribed user ...

"**To Be Or Not** To Be" (Ep. 101) from Chrissy & Mr. Jones | Full - Vh1
www.vh1.com/video/chrissy.../**to-be-or-not**-to.../playlist.jhtml
3 days ago – Chrissy wants to take their relationship to the next level, but feels like Jim isn't on the same page. Meanwhile, their house has been sold and the couple has ...

Europe: **To Be or Not** To Be - Empire - Al Jazeera English
www.aljazeera.com/.../empire/.../2012615122134208504.html
15 Jun 2012 – As Europe's crisis worsens without any solution in sight is a shift in political power a sign of hope on the horizon?

Language-specificity

- Many of the above features embody transformations that are
  - Language-specific and
  - Often, application-specific
- These are “plug-in” addenda to indexing
- Both open source and commercial plug-ins are available for handling these

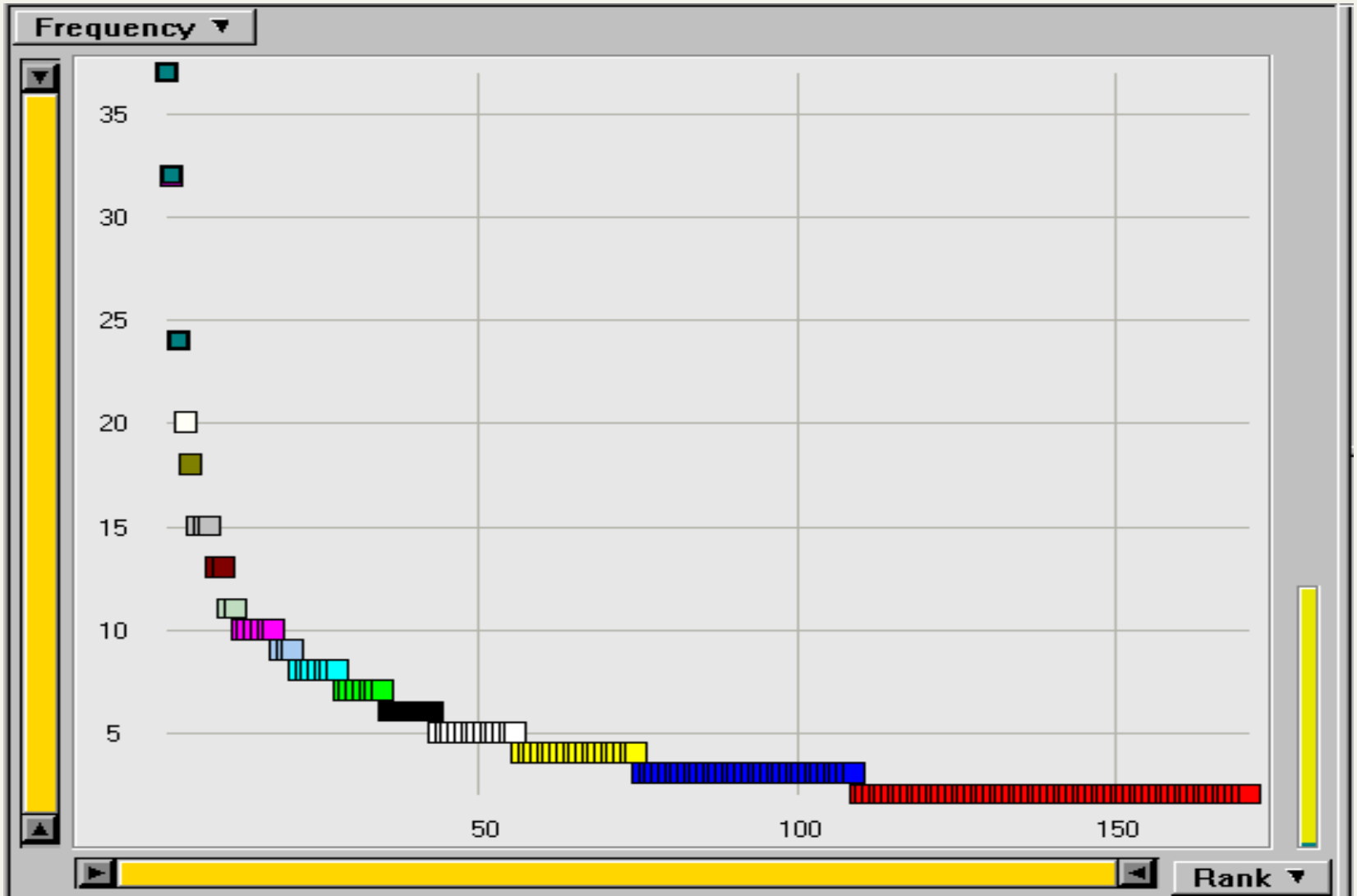
Statistical properties of text

Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

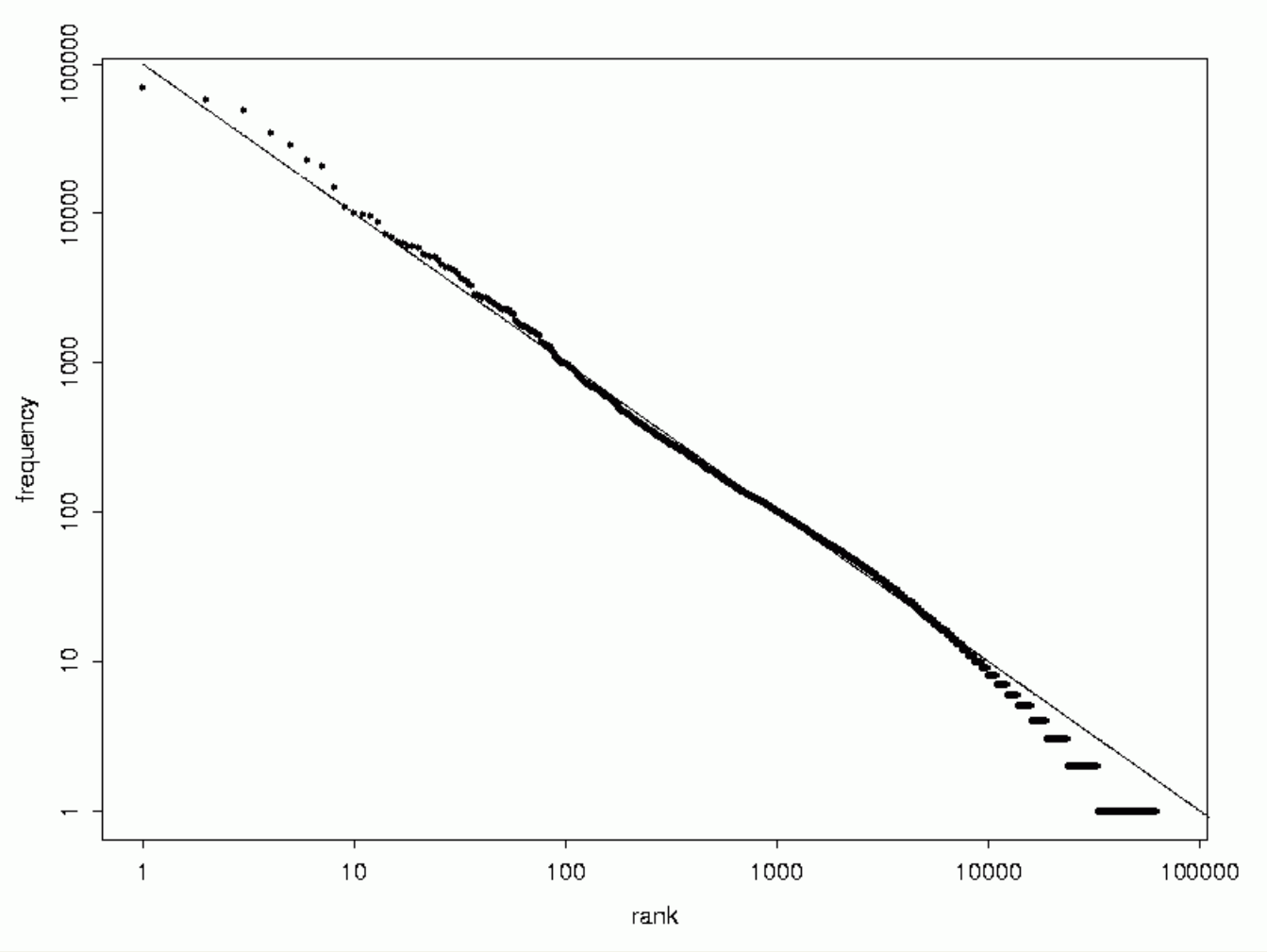
Statistical properties of texts

- Tokens are not distributed uniformly. They follow the so called "Zipf Law"
  - Few tokens are very frequent
  - A middle sized set has medium frequency
  - Many are rare
- The first 100 tokens sum up to 50% of the text, and many of them are *stopwords*

An example of “Zipf curve”



A log-log plot for a Zipf's curve



The Zipf Law, in detail

- k-th most frequent token has frequency f(k) approximately 1/k;
- Equivalently, the product of the frequency f(k) of a token and its rank k is a constant

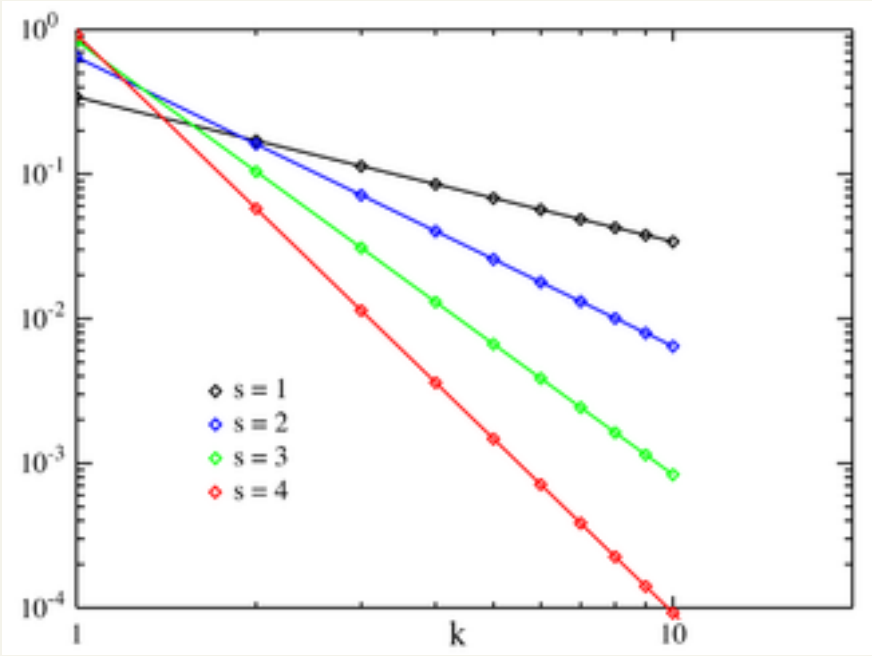
$$k * f(k) = c$$
$$f(k) = c / k$$

$$f(k) = c / k^{s}$$
$$s = 1.5 \div 2.0$$

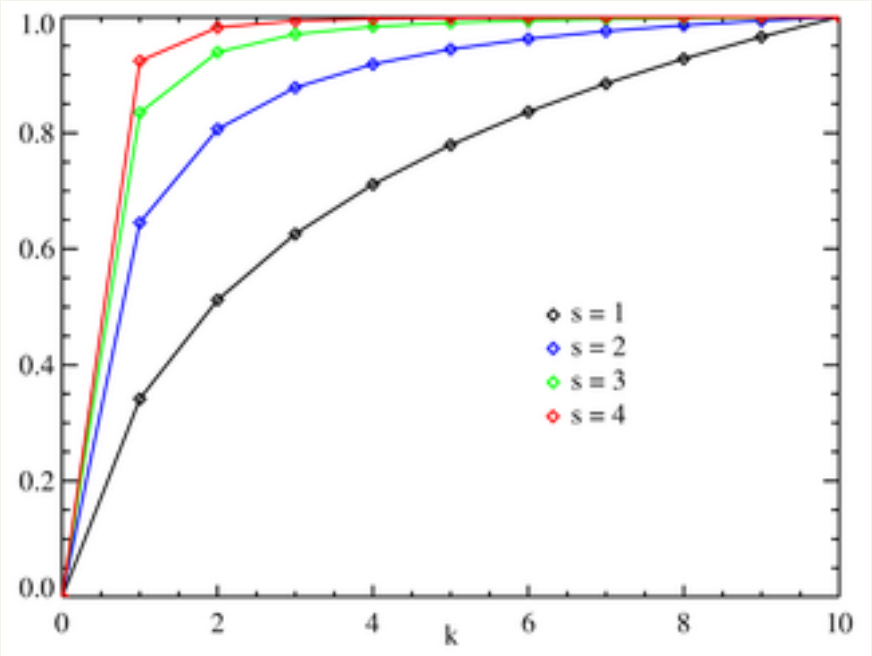
General Law

- Scale invariant: $f(b*k) = b^{-s} * f(k)$

Distribution *vs* Cumulative distr



Log-log plot



Power-law with smaller exponent

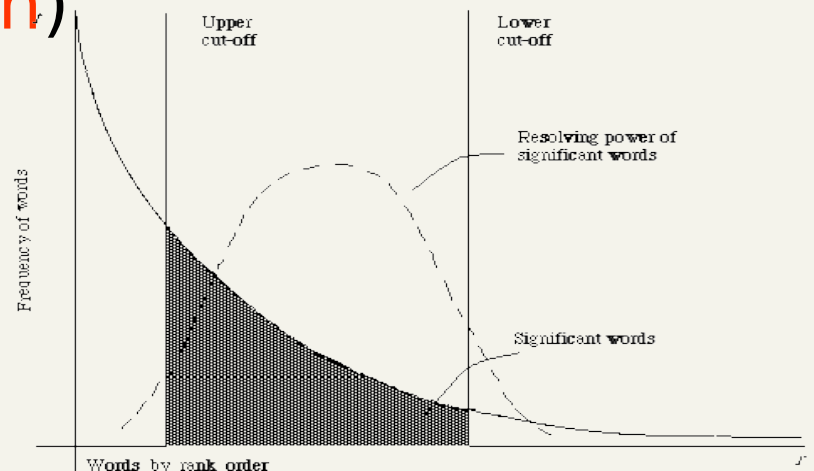
Sum after the k-th element is $\leq$ *f(k) * k/(s-1)*

Sum up to the k-th element is $\geq$ *f(k) * k*

Other statistical properties of texts

- The number of distinct tokens grows as
  - The so called "Heaps Law" (n^{β} where $\beta<1$, tipically 0.5, where n is the total number of tokens)
  - The average token length grows as $\Omega(\log n)$

- Interesting words are the ones with medium frequency (Luhn)



Keywords extraction

- Key step in many IR tasks:
  - Define the token in Inverted Lists
  - Define dictionary terms in bag-of-words
  - Etc. etc.
- Supervised and unsupervised

Keyword extraction

Paolo Ferragina
Dipartimento di Informatica
Università di Pisa

Statistical extraction

Collocation: two or more words that correspond to some conventional way of saying

- **Limited compositionality:** meaning cannot be fully inferred by its constitutent words [*white wine* or *white hair*]
- **Non substituability:** cannot substitute other words and keep same meaning [*yellow wine*]
- **Non modifiability:** cannot add lexical material [*an idiom would be changed*]

Just frequency + PoS tagging

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

$C(w^1\ w^2)$	w^1	w^2	Tag Pa
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N

Mean and Variance

Often the words are not adjacent to each other (*need flexibility*)

a. she knocked on his door

b. they knocked at the door

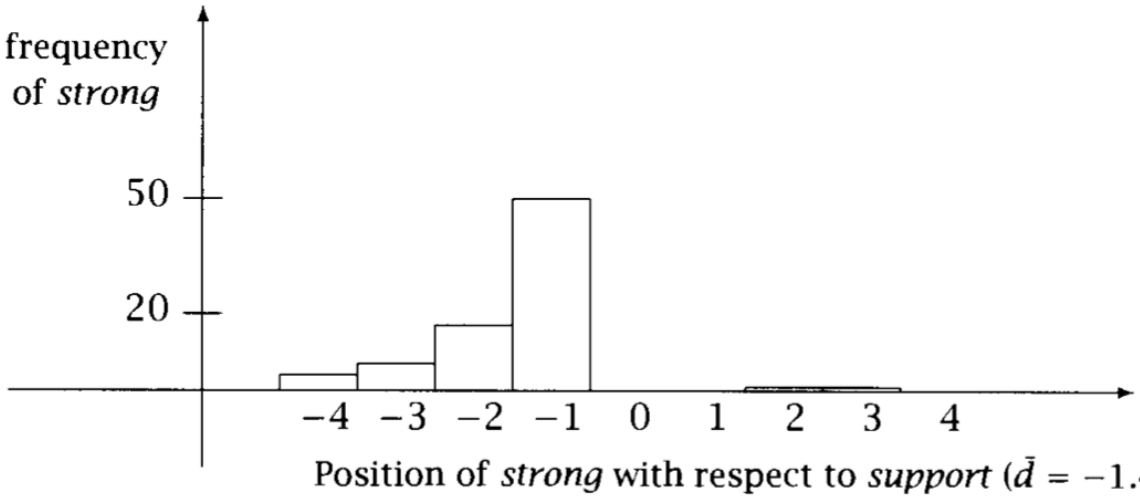
c. 100 women knocked on Donaldson's door

d. a man knocked on the metal front door

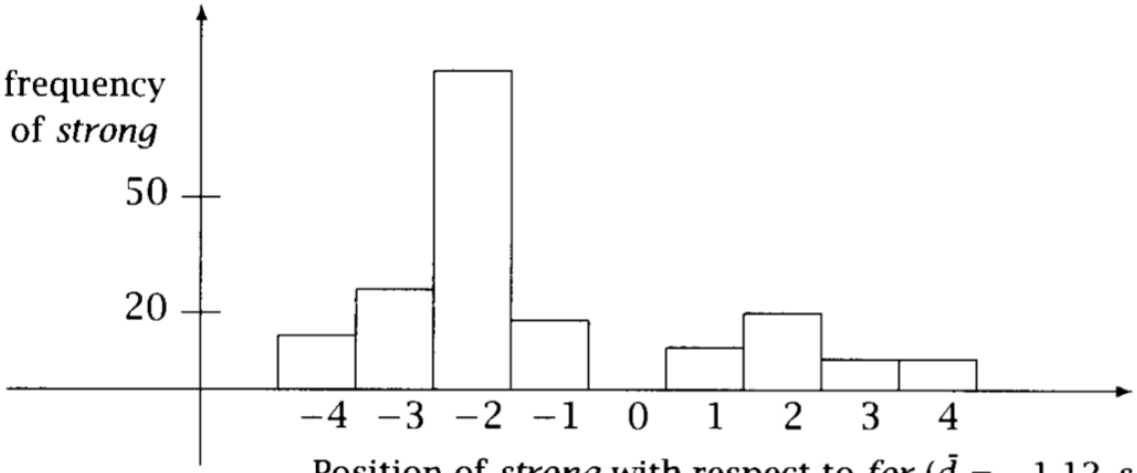
Compute the mean and the variance of the

distance, by restricting within a window.

Example of distance distribution



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

Example of distance distribution

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

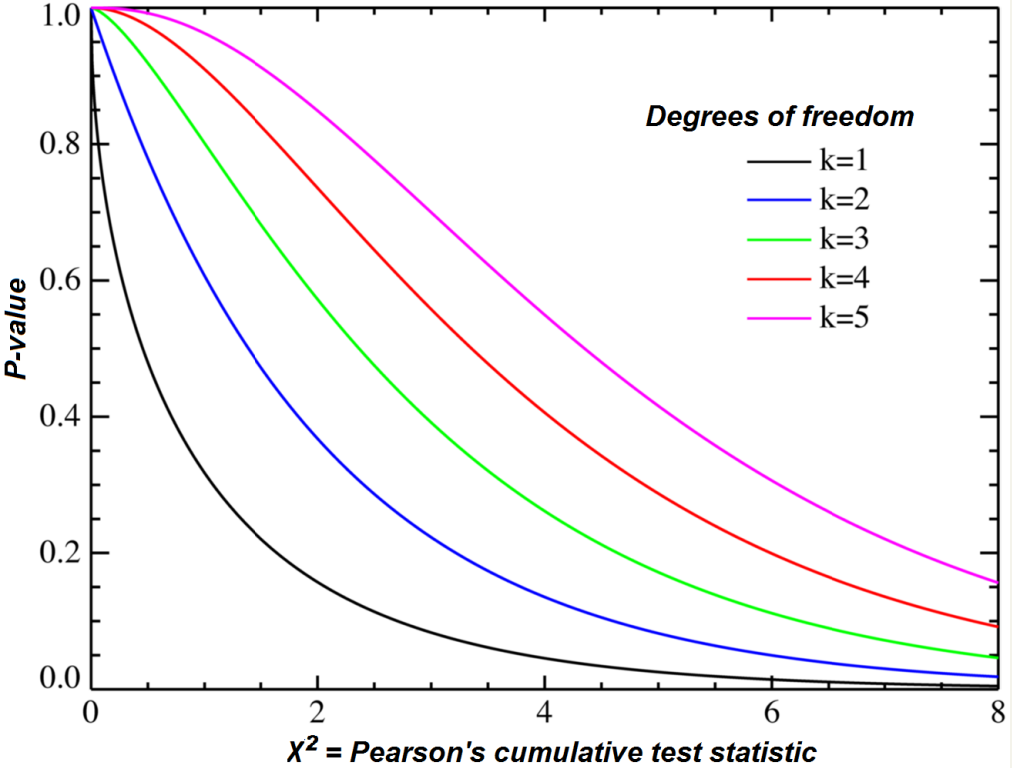
High s no interesting relation

If s *large* → the collocation is not interesting

If d > 0 and s very small → interesting (new)

Pearson's chi-square test (*bigrams*)

Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem.



If it is **improbably large**, then reject the *null* hyp.

Pearson's chi-square test (*bigrams*)

It is good for few occurrences

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

Table 5.8 A 2-by-2 table showing the dependence of occurrences of *new* and *companies*. There are 8 occurrences of *new companies* in the corpus, 4,667 bigrams where the second word is *companies*, but the first word is not *new*, 15,820 bigrams with the first word *new* and a second word different from *companies*, and 14,287,181 bigrams that contain neither word in the appropriate position.

For an r * c table, there are (r - 1)(c - 1) degrees of freedom

Pearson's chi-square test

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

To estabish the significance of a collocation, compute

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where E_ij = N * freq(i) * freq(j)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

→ **1.55**

Pearson's chi-square test

	w_1 = *new*	$w_1 \neq$ *new*
w_2 = *companies*	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq$ *companies*	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

Chi-square → 1.55

There are (r - 1)(c - 1) = 1* 1 = 1 degrees of freedom

Let us consider a P-value = 0.10

	P	0.99	0.95	0.10	0.05	0.01	0.005	0.001
d.f.	1	0.00016	0.0039	2.71	3.84	6.63	7.88	10.83
	2	0.020	0.10	4.60	5.99	9.21	10.60	13.82
	3	0.115	0.35	6.25	7.81	11.34	12.84	16.27
	4	0.297	0.71	7.78	9.49	13.28	14.86	18.47
	100	70.06	77.93	118.5	124.3	135.8	140.2	149.4

The value in this tables is 2.71 so our X^2 is smaller (=1.55), and hence the *null hypothesis* is plausible → this is not a good pair

Rapid Automatic Keyword Extraction

Key properties:

- Works on single (not much long) documents
- Easily applicable to new domains
- Fast
- Unsupervised

Key observation: keywords frequently contain multiple words but rarely contain punctuation or stop words.

RAKE pipeline

The input parameters:

- a set of word delimiters,
- a set of phrase delimiters,
- a list of stop words (or stoplist).

Step #1: Candidate keywords:

- document is split into an array of words by the specified word delimiters.
- This array is split into sequences of contiguous words at phrase delimiters and then stop word.
- Words within a sequence are considered a candidate keyword.

A running example

Compatibility of systems of linear constraints over the set of natural numbers

Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.

Manually assigned keywords:
linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets

- **Step #1**

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

RAKE pipeline

Step #2: Scoring candidate keywords:

- Compute the table of co-occurrences.
- And few metrics: *freq(w)* = total frequency on diagonal, *deg(w)* sum over row

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
algorithms	2						1																		
bounds		1																							1
compatibility			2																						
components				1																					

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

RAKE pipeline

Step #2: Scoring candidate keywords:

- Final score is the sum of *deg(w)/freq(w)* for the constituting words w of a keyword

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

minimal generating sets (8.7), linear diophantine equations (8.5), minimal supporting set (7.7), minimal set (4.7), linear constraints (4.5), natural numbers (4), strict inequations (4), nonstrict inequations (4), upper bounds (4), corresponding algorithms (3.5), set (2), algorithms (1.5), compatibility (1), systems (1), criteria (1), system (1), components (1),constructing (1), solving (1)

RAKE pipeline

Step #3: Adjoining keywords - identifies keywords that contain interior stop words such as *axis of evil.*

- Looks for pairs of keywords that adjoin one another ***at least twice*** in the same document and in the same order
- The score for the new keyword is the ***sum*** of its member keyword scores.

Step #4: Selecting keywords - the top one-third

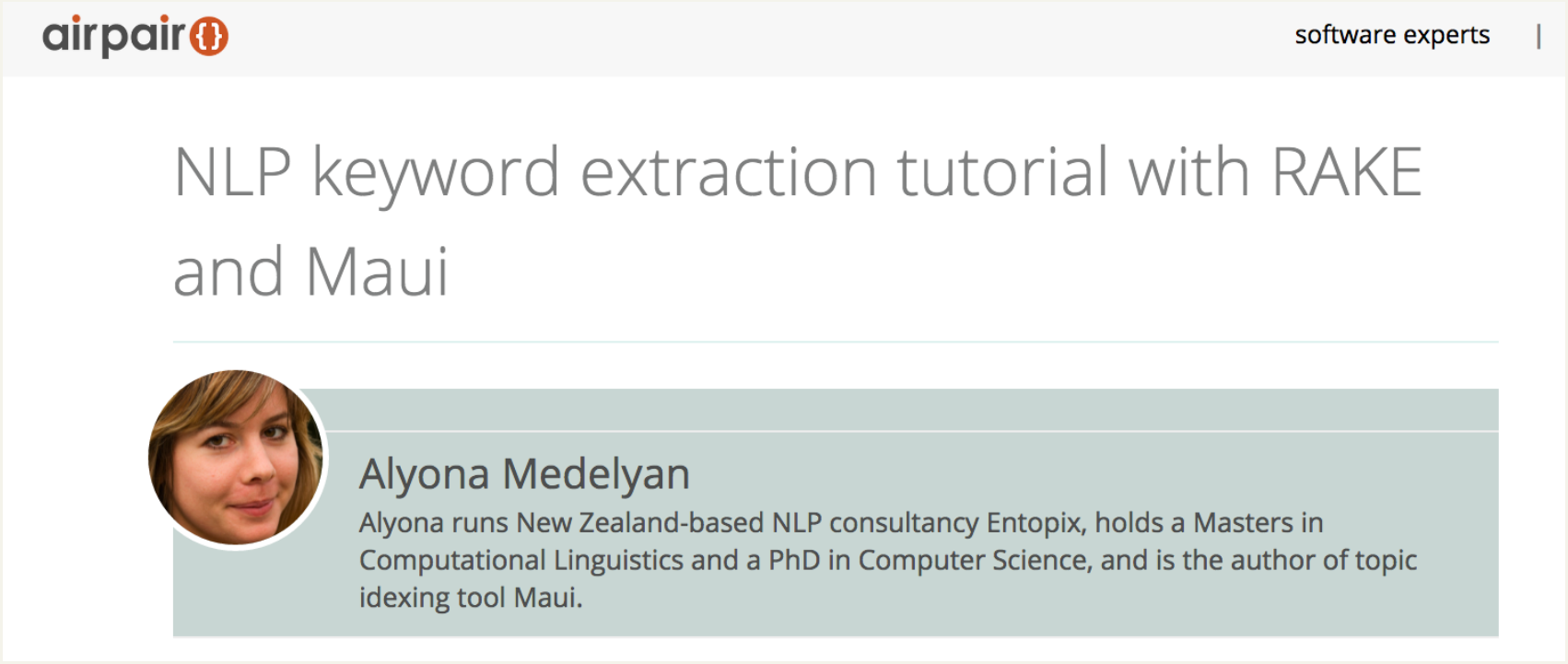
Result

Table 1.1 Comparison of keywords extracted by RAKE to manually assigned keywords for the sample abstract.

Extracted by RAKE	Manually assigned
minimal generating sets	minimal generating sets
linear diophantine equations	linear Diophantine equations
minimal supporting set	
minimal set	
linear constraints	linear constraints
natural numbers	
strict inequations	strict inequations
nonstrict inequations	nonstrict inequations
upper bounds	upper bounds
	set of natural numbers

3 false positives and 1 false negative (but similar..)

Library



Tutorial: www.airpair.com/nlp/keyword-extraction-tutorial

Python: github.com/zelandiya/RAKE-tutorial