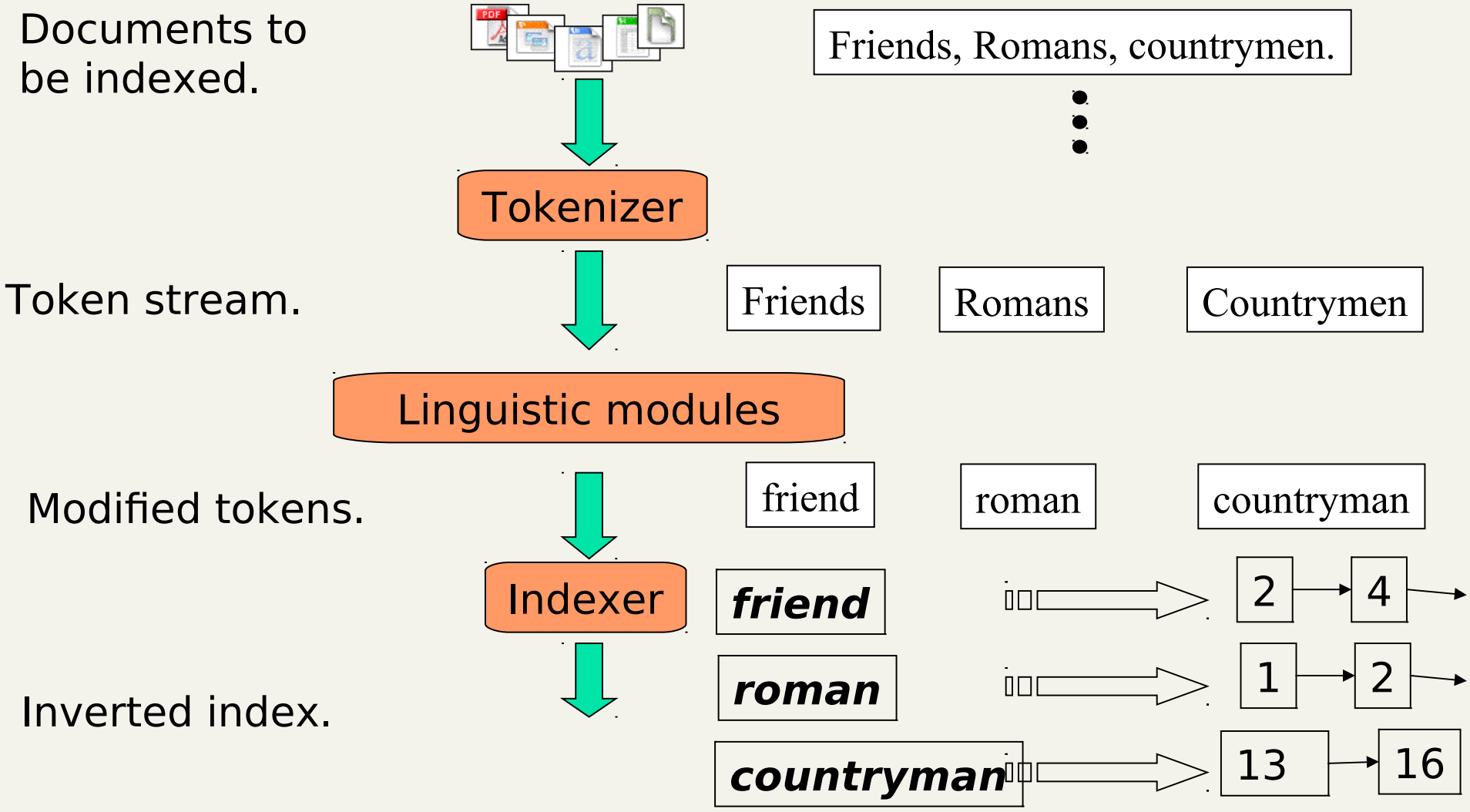
Document Parsing

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.[illegible]/**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 …

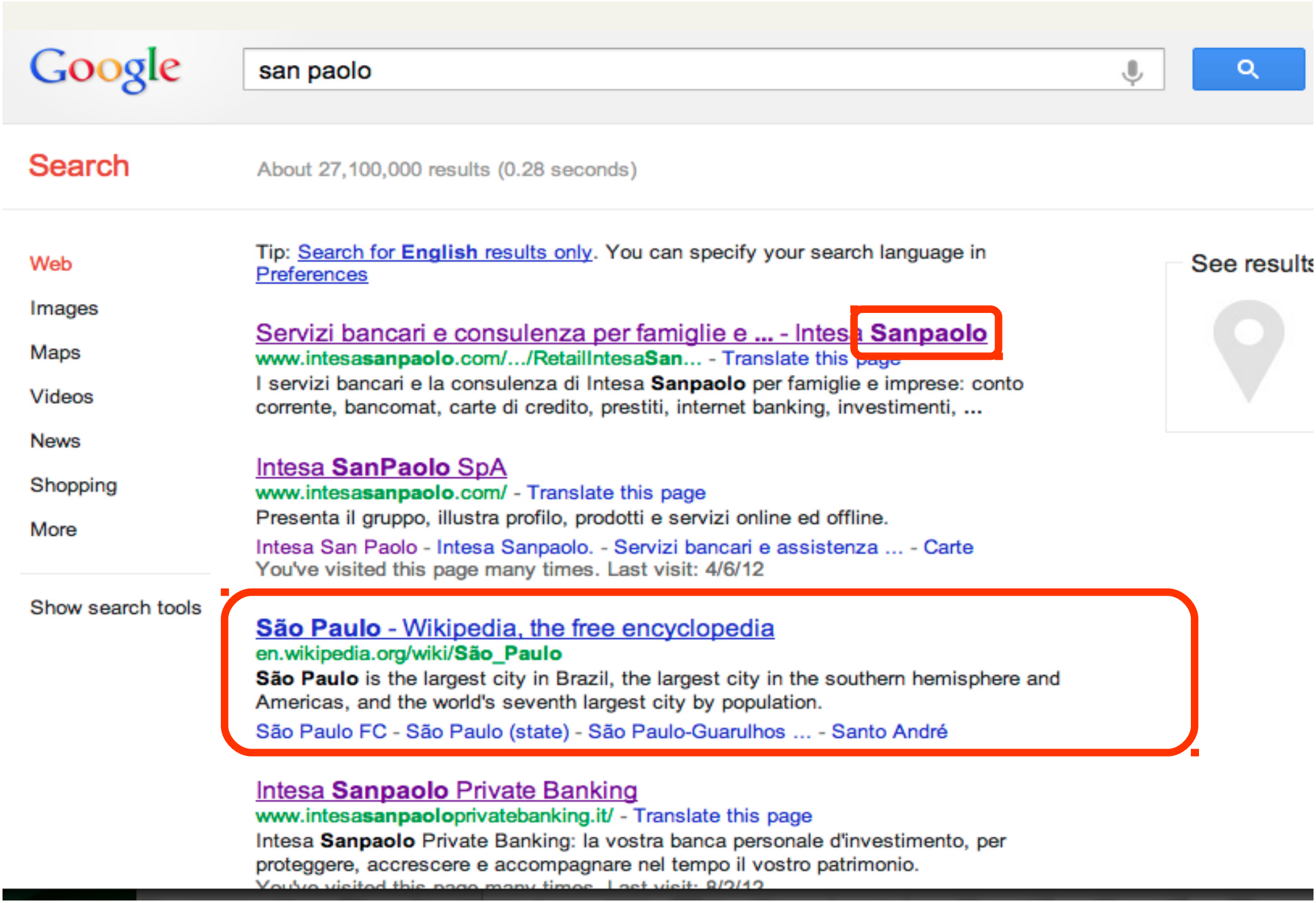
Tokenization

- 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
Servizi bancari e consulenza per famiglie e ... - Intesa Sanpaolo
www.intesasanpaolo.com/.../RetailIntesaSan... - 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.intesasanpaolo.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.intesasanpaoloprivatebanking.it/ - Translate this page
Intesa Sanpaolo Private Banking: la vostra banca personale d'investimento, per proteggere, accrescere e accompagnare nel tempo il vostro patrimonio.
See results

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”

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

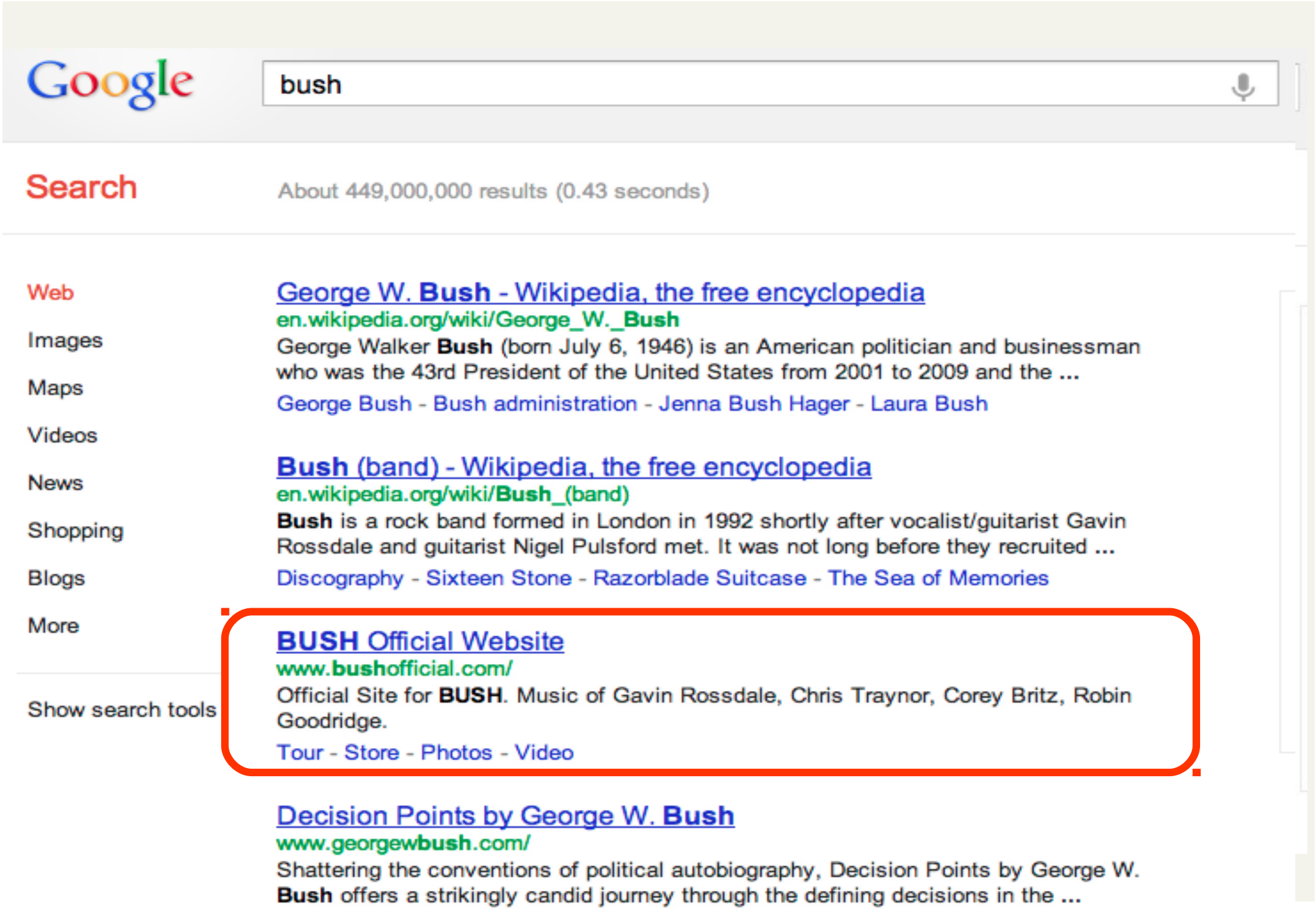
CAT
www.**cat**iim.in/
Registration for **CAT** 2012 is now closed. Registered candidates may log on to https://iim.prometric.com to print a copy of their Admit Card until the end of the ...
Iim cat result - CAT Eligibility - Selection Process of IIMs - CAT 2012 Test Sites

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…


Google
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

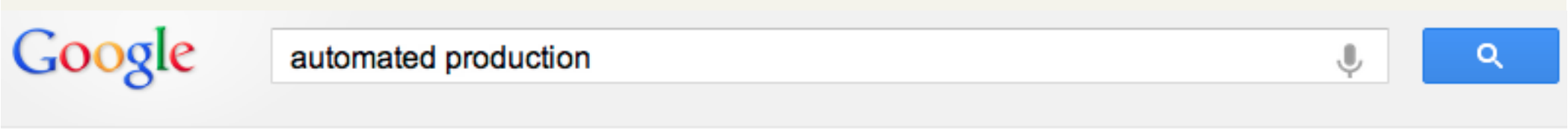
Porter's algorithm

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



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

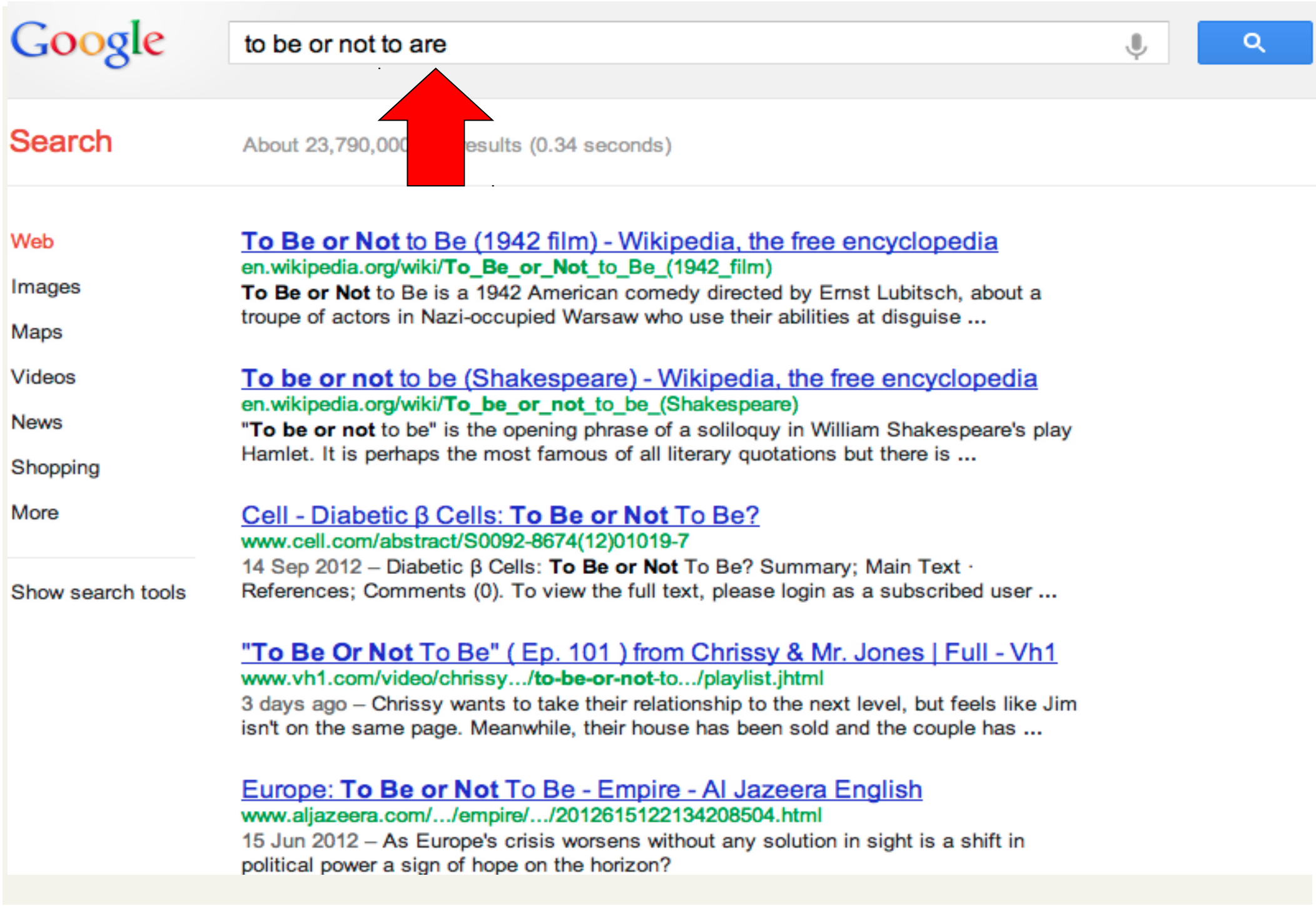
Home | **Automated Production**
automatedproduction.biz/
Automated Production specializes in DNV design and fabrication projects including reels, lift frames, positioning systems, baskets, living quarters, storage ...

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


Google
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/.../20126151221342085O4.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

Does stemming help?

- English: very mixed results. Helps recall for some queries but harms precision on others
  - operative, operational, operations → oper
- Definitely useful for Spanish, German (with compound splitting) Finnish, …
  - 30% performance gains for Finnish!

Index parameters vs. what we index (details *IIR* Table 5.1, p.80)

size of	word types (terms)			non-positional postings			positional postings		
	dictionary			non-positional index			positional index		
	Size (K)	Δ%	cumul %	Size (K)	Δ %	cumul %	Size (K)	Δ %	cumul %
Unfiltered	484			109,971			197,879		
No numbers	474	-2	-2	100,680	-8	-8	179,158	-9	-9
Case folding	392	-17	-19	96,969	-3	-12	179,158	0	-9
30 stopwords	391	-0	-19	83,390	-14	-24	121,858	-31	-38
150 stopwords	391	-0	-19	67,002	-30	-39	94,517	-47	-52
stemming	322	-17	-33	63,812	-4	-42	94,517	0	-52

Exercise: give intuitions for all the '0' entries. Why do some zero entries correspond to big deltas in other columns?

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

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

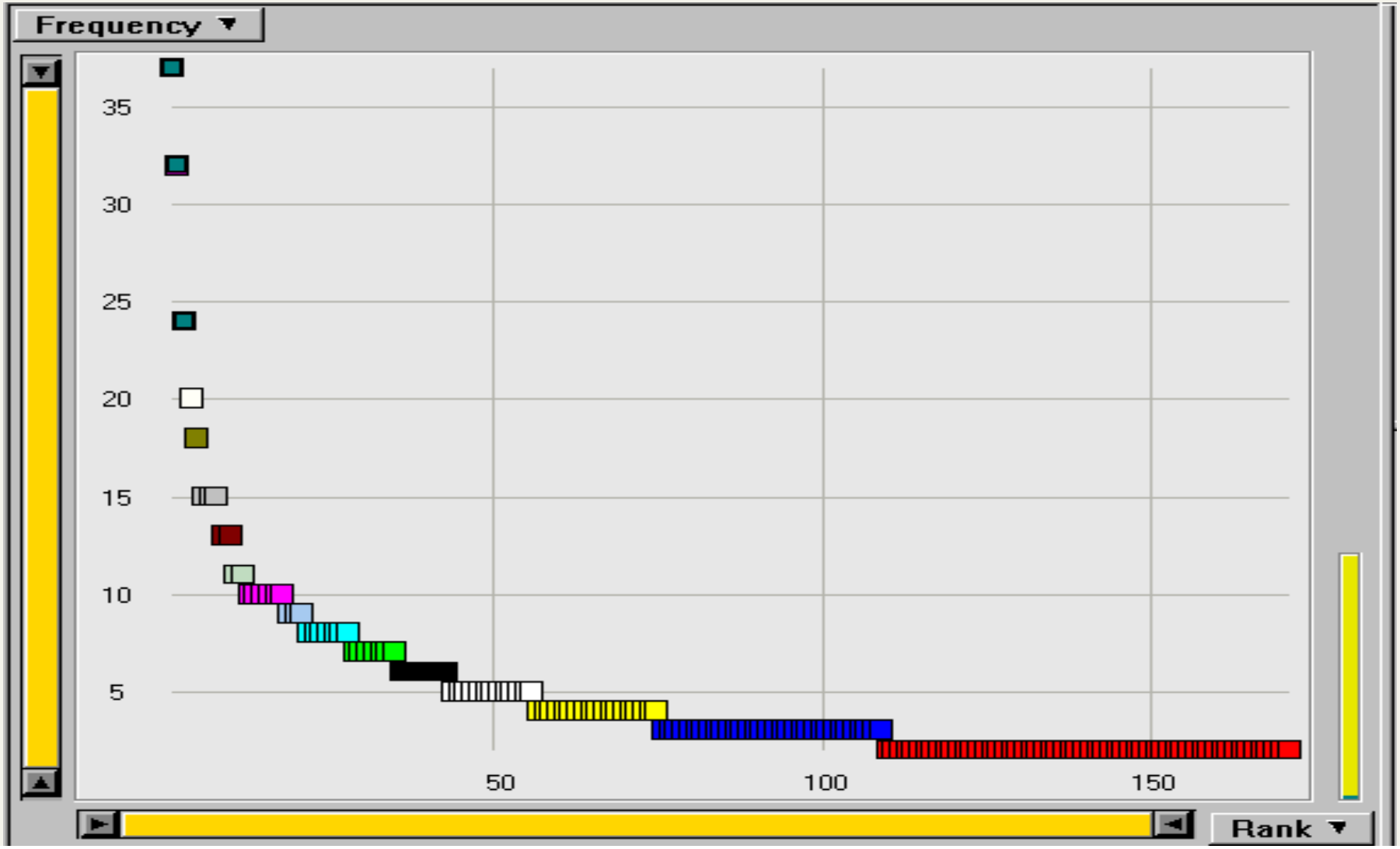
$$k * f(k) = c$$
$$f(k) = c / k$$

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

General Law

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

An example of “Zipf curve”



Some math

- Taking the logarithm $f(k) = c/k^s$ yields

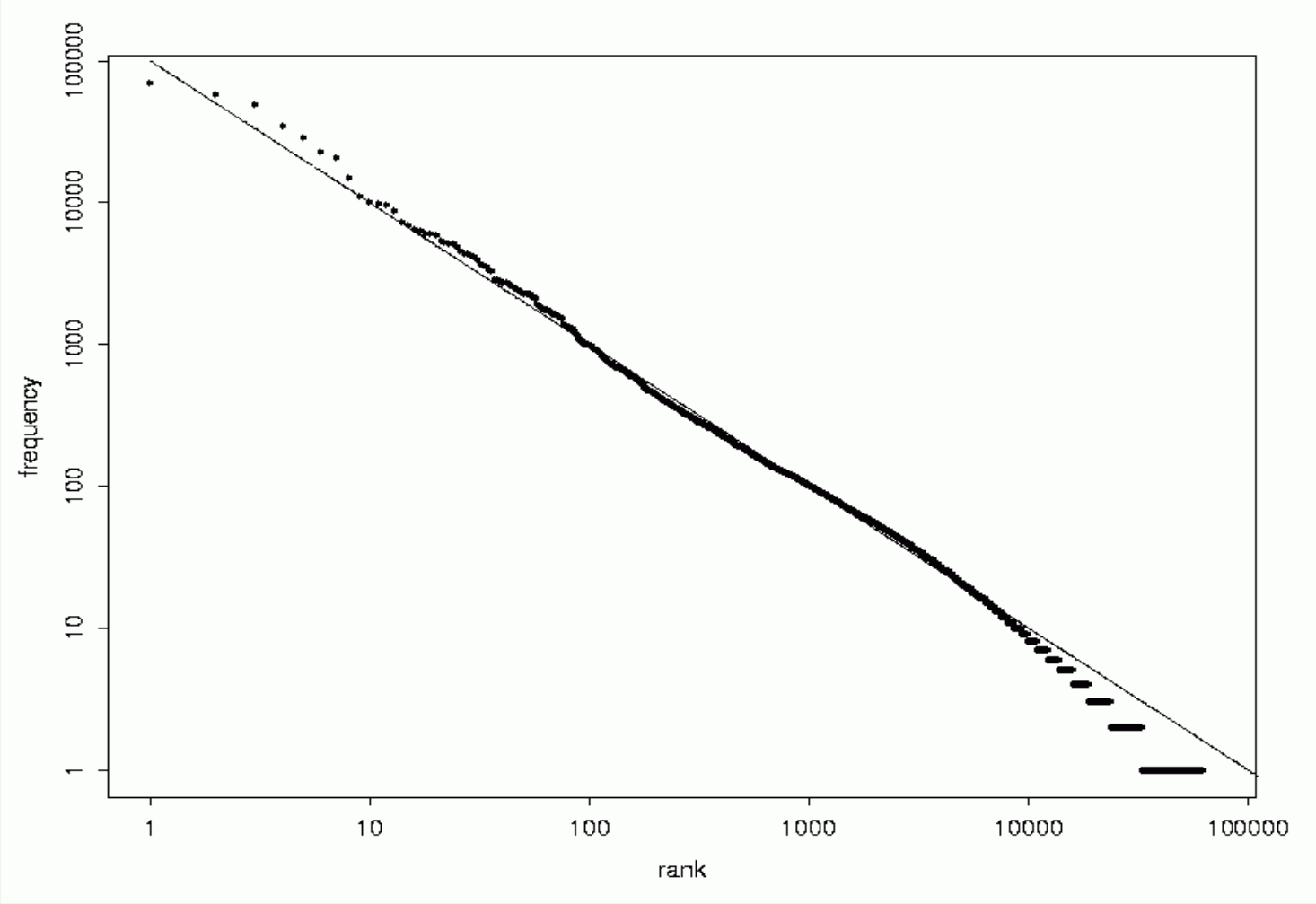
$$\log(f(k)) = \log c - s \log k$$

- If we plot $\log(f(k))$ vs $\log(k)$ we get

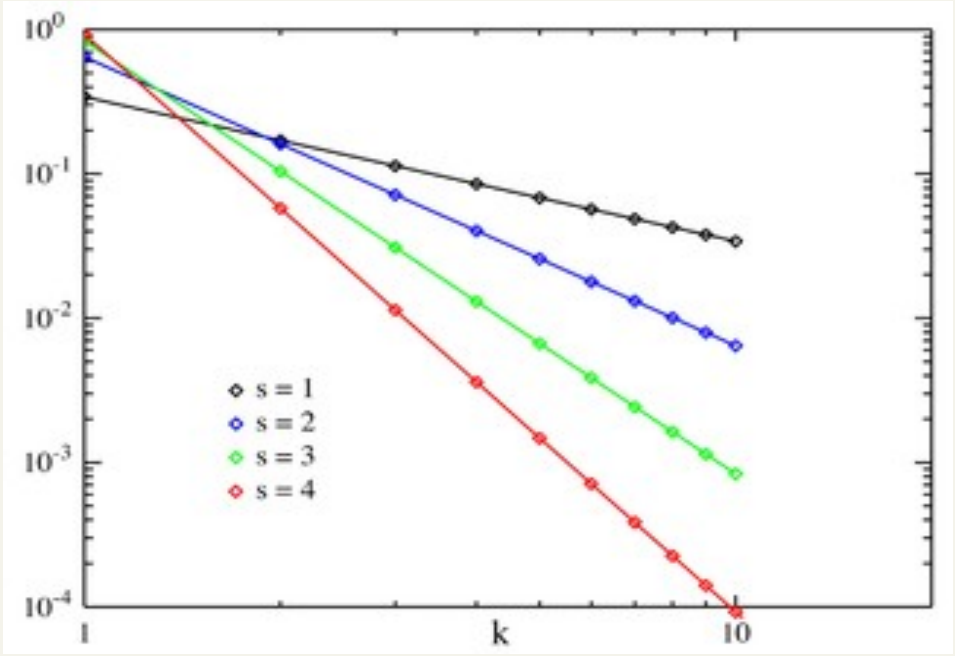
$$y = \log c - s\, x$$

ie a line with slope -s

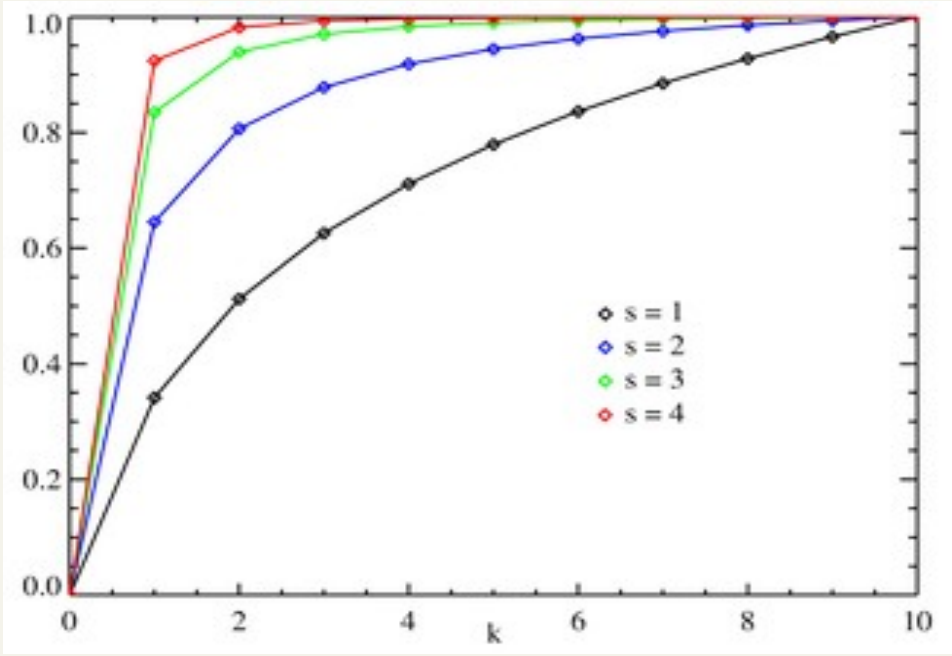
A log-log plot for a Zipf’s curve



Distribution *vs* Cumulative distr



Log-log plot



Power-law with smaller exponent

Sum after the k-th element is ≤ f(k) * k/(s-1)
Sum up to the k-th element is ≥ f(k) * k

Other statistical properties of texts

- The number of distinct tokens grows as
  - The so called "Heaps Law" (n^{β} where $\beta<1$, typically 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)