Recuperació de la Informació (REIN)

Grau en Enginyeria Informàtica

Departament de Ciències de la Computació (CS)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

Índex

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades
  - que sigui **rellevant** a una necessitat d'informació formulada per un usuari (*query*)

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades
  - que sigui **rellevant** a una necessitat d'informació formulada per un usuari (*query*)
  - i que ho faci de forma **eficient**.

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades
  - que sigui **rellevant** a una necessitat d'informació formulada per un usuari (*query*)
  - i que ho faci de forma **eficient**.
- La rellevància de la informació inclou:

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades
  - que sigui **rellevant** a una necessitat d'informació formulada per un usuari (*query*)
  - i que ho faci de forma **eficient**.
- La rellevància de la informació inclou:
  - que correspongui al tema demanat

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades
  - que sigui **rellevant** a una necessitat d'informació formulada per un usuari (*query*)
  - i que ho faci de forma **eficient**.
- La rellevància de la informació inclou:
  - que correspongui al tema demanat
  - que estigui actualitzada

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades
  - que sigui **rellevant** a una necessitat d'informació formulada per un usuari (*query*)
  - i que ho faci de forma **eficient**.
- La rellevància de la informació inclou:
  - que correspongui al tema demanat
  - que estigui actualitzada
  - que sigui fiable (font)

Recuperació de la informació (RI), I

- Podríeu donar alguna definició de **dades**, **informació** i **coneixement**?
- La tasca de RI té a veure amb:
  - Trobar material (sovint documents)
  - en una **gran** col·lecció de dades
  - que sigui **rellevant** a una necessitat d'informació formulada per un usuari (*query*)
  - i que ho faci de forma **eficient**.
- La rellevància de la informació inclou:
  - que correspongui al tema demanat
  - que estigui actualitzada
  - que sigui fiable (font)
  - que satisfaci a l'usuari i a l'ús que vulgui fer-ne

Consultes a BD vs RI, I

Hi ha diferències?

Consultes a BD vs RI, I

Hi ha diferències?

- En una BD
  - la semàntica de cada objecte està ben definida
  - llenguatges de consulta complexos (p.e., SQL)
  - recupera exactament el que has demanat
  - èmfasi en l'eficiència

Consultes a BD vs RI, I

Hi ha diferències?

- En una BD
  - la semàntica de cada objecte està ben definida
  - llenguatges de consulta complexos (p.e., SQL)
  - recupera exactament el que has demanat
  - èmfasi en l'eficiència
- En Recuperació de la Informació (RI)
  - la semàntica de cada objecte és subjectiva, no està ben definida
  - normalment els llenguatges de consulta són simples (p.e., llenguatge natural)
  - hauria de recuperar el que vols, encara que la consulta sigui dolenta
  - èmfasi en l'eficàcia, tot i que l'eficiència és important

Consultes a BD vs RI, II

Hi ha diferències?

Consultes a BD vs RI, II

Hi ha diferències?

En Recuperació de la Informació:

- Podem no saber on està la informació que busquem

Consultes a BD vs RI, II

Hi ha diferències?

En Recuperació de la Informació:

- Podem no saber on està la informació que busquem
- Podem no saber si tenim la informació que busquem

Consultes a BD vs RI, II

Hi ha diferències?

En Recuperació de la Informació:

- Podem no saber on està la informació que busquem
- Podem no saber si tenim la informació que busquem
- Podem, fins i tot, no saber quina informació estem buscant realment

Consultes a BD vs RI, II

Hi ha diferències?

En Recuperació de la Informació:

- Podem no saber on està la informació que busquem
- Podem no saber si tenim la informació que busquem
- Podem, fins i tot, no saber quina informació estem buscant realment
- Per exemple, noteu la gran diferència qualitativa entre:
  - "Troba'm el número de telèfon d'algú" vs
  - "Parla'm de les influències que va rebre Beethoven dels compositors europeus de finals del segle XVI"

- D'un sistema de RI esperem que ens ajudi a trobar allò que busquem (poc o molt concret) en una gran col·lecció de documents.

- D'un sistema de RI esperem que ens ajudi a trobar allò que busquem (poc o molt concret) en una gran col·lecció de documents.
- Esperar trobar alguna cosa que sabem que el sistema no pot recuperar, p.e. predicció sobre un valor al mercat, forma part d'altres àrees: Mineria de dades (DM), Estadística i Aprenentatge automàtic (ML).

Recuperació de la informació, III

Objectiu del curs

- Recuperació de documents de la xarxa

Recuperació de la informació, III

Objectiu del curs

- Recuperació de documents de la xarxa
  - Els documents a la xarxa contenen termes i enllaços.

Recuperació de la informació, III

Objectiu del curs

- Recuperació de documents de la xarxa
  - Els documents a la xarxa contenen termes i enllaços.
  - Els usuaris fan consultes per buscar documents.

Recuperació de la informació, III

Objectiu del curs

- Recuperació de documents de la xarxa
  - Els documents a la xarxa contenen termes i enllaços.
  - Els usuaris fan consultes per buscar documents.
  - Les consultes sovint també estan formades per termes.

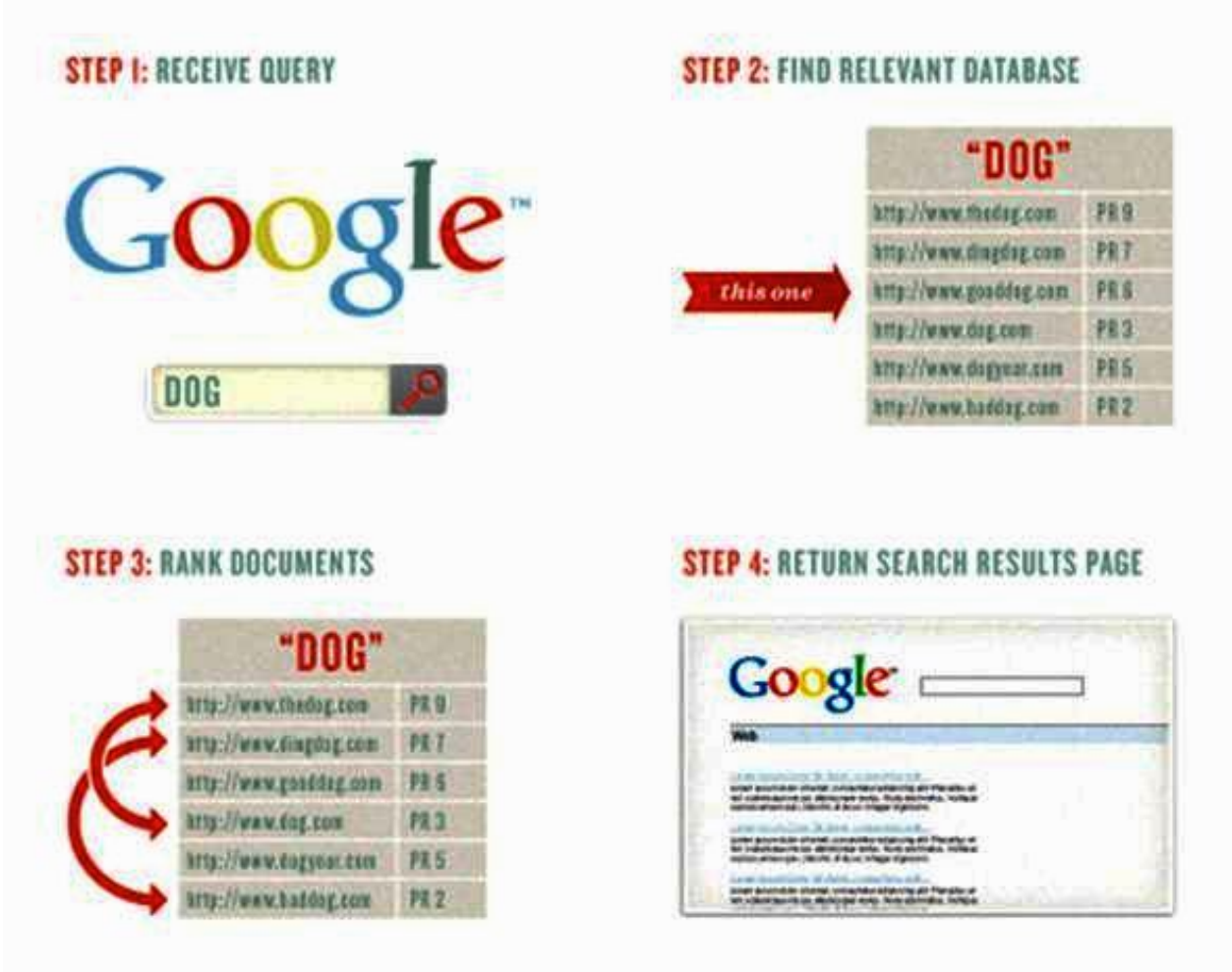
Recuperació de la informació, III

Objectiu del curs

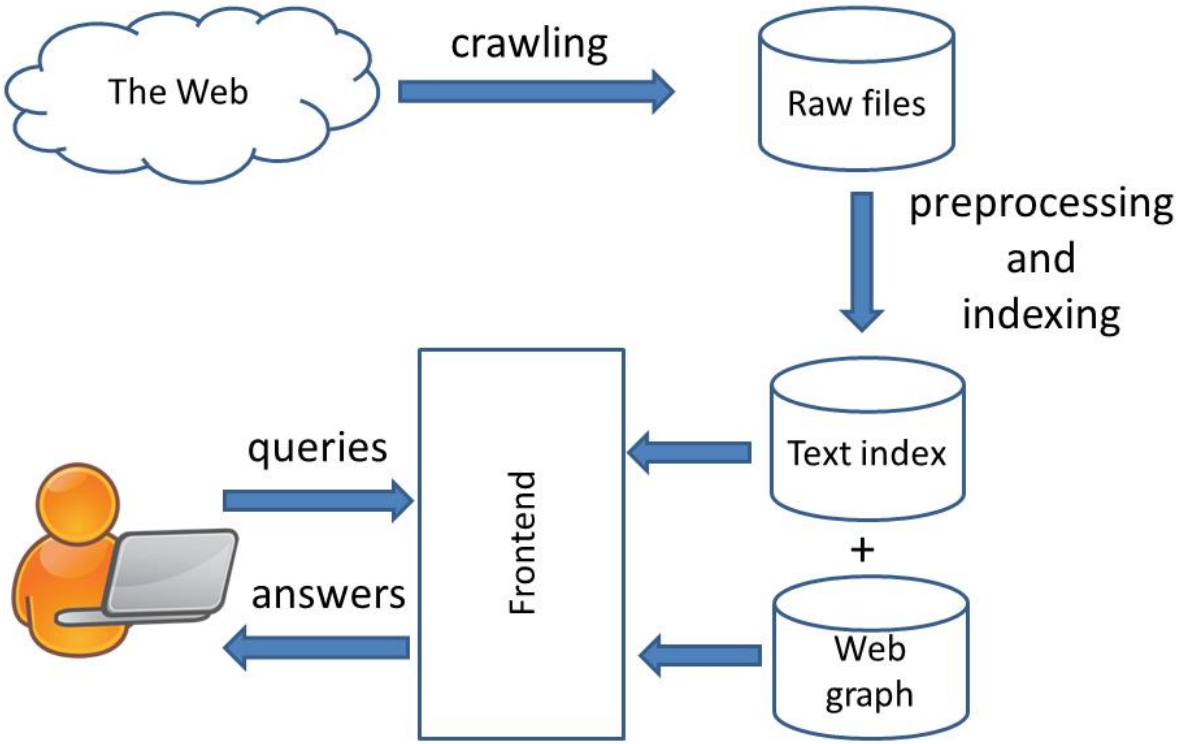
- Recuperació de documents de la xarxa
  - Els documents a la xarxa contenen termes i enllaços.
  - Els usuaris fan consultes per buscar documents.
  - Les consultes sovint també estan formades per termes.
- A REIN no tractarem la recuperació d'àudio, vídeo, imatges, arxius binaris, etc, ni altres tipus de consultes.

Índex

Motors de cerca (*Search engines*)



El procés de la RI, I



El procés de la RI, I

Processos offline:

- Rastreig (*crawling*)
- Preprocessament
- Indexació

El procés de la RI, I

Processos offline:

- Rastreig (*crawling*)
- Preprocessament
- Indexació

Objectiu:

Preparar estructures de dades perquè l'accés online sigui més ràpid.

- Poden abordar càlculs complexos, per exemple, escanejar cada document diverses vegades
- Han de generar una sortida compacta (estructura de dades)

El procés de la RI, II

Processos online:

- Obtenir la consulta
- Recuperar els documents rellevants
- Ordenar els documents (rànquing)
- Formatejar la resposta i tornar-la a l'usuari

El procés de la RI, II

Processos online:

- Obtenir la consulta
- Recuperar els documents rellevants
- Ordenar els documents (rànquing)
- Formatejar la resposta i tornar-la a l'usuari

Objectiu:

Reacció instantània, visualització útil.

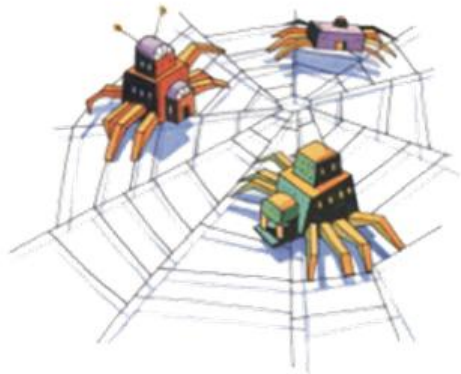
- Pot usar informació addicional: localització de l'usuari, anuncis, ...

Índex

Crawlers (rastrejadors)

Crawlers, rastrejadors, aranyes, robots...

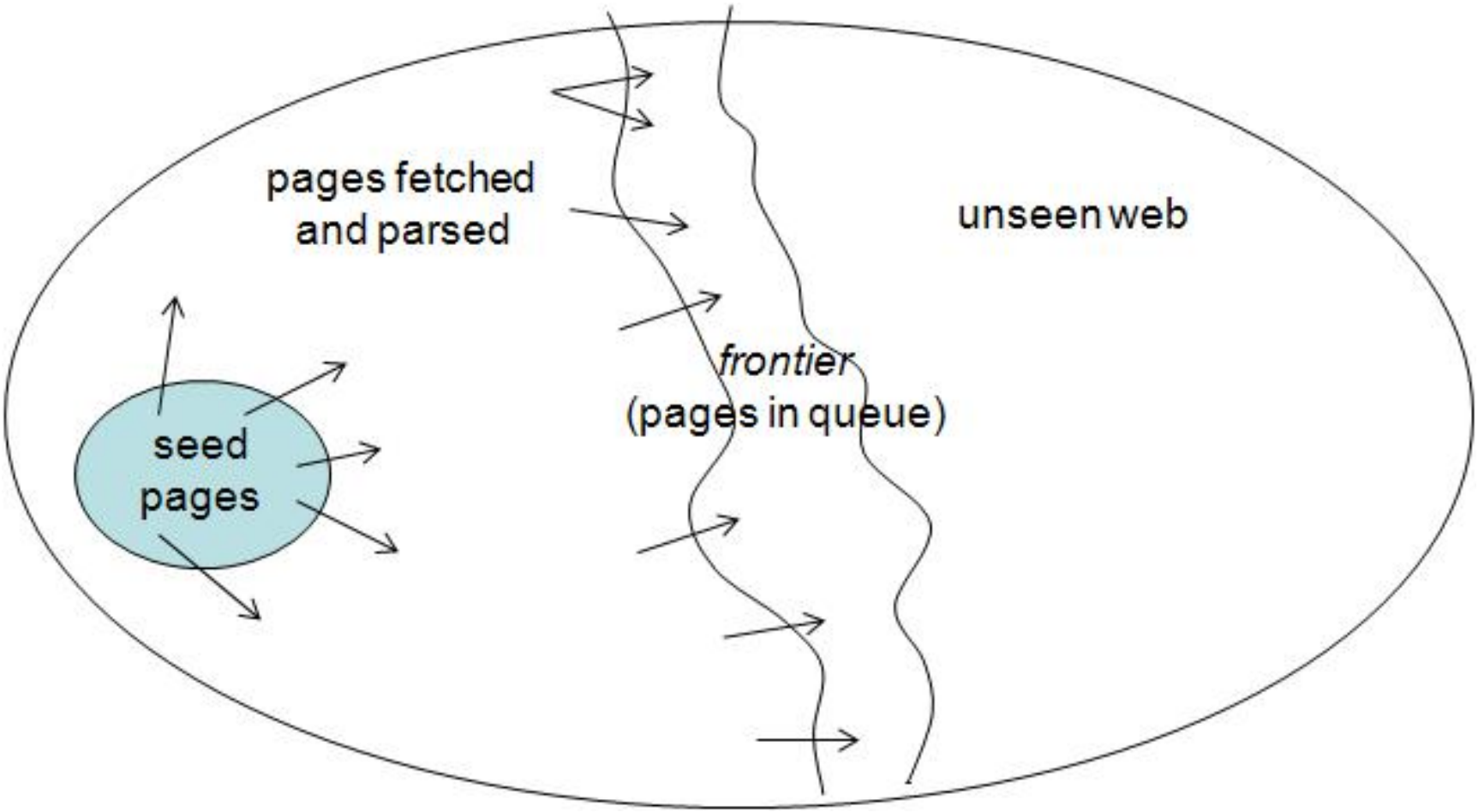
Exploren de forma sistemàtica la xarxa i localitzen els enllaços (interns o externs) dintre de cada pàgina per recuperar altres adreces.



Operació bàsica del rastrejador:

```
inicialitza cua amb les URL de pàgines conegudes
repetir
    escollir una URL de la cua
    obtenir i analitzar la pàgina
    descartar-la o afegir-la a la BD
    afegir a la cua les URL que conté
fi repetir
```

Rastreig com a exploració d'un graf



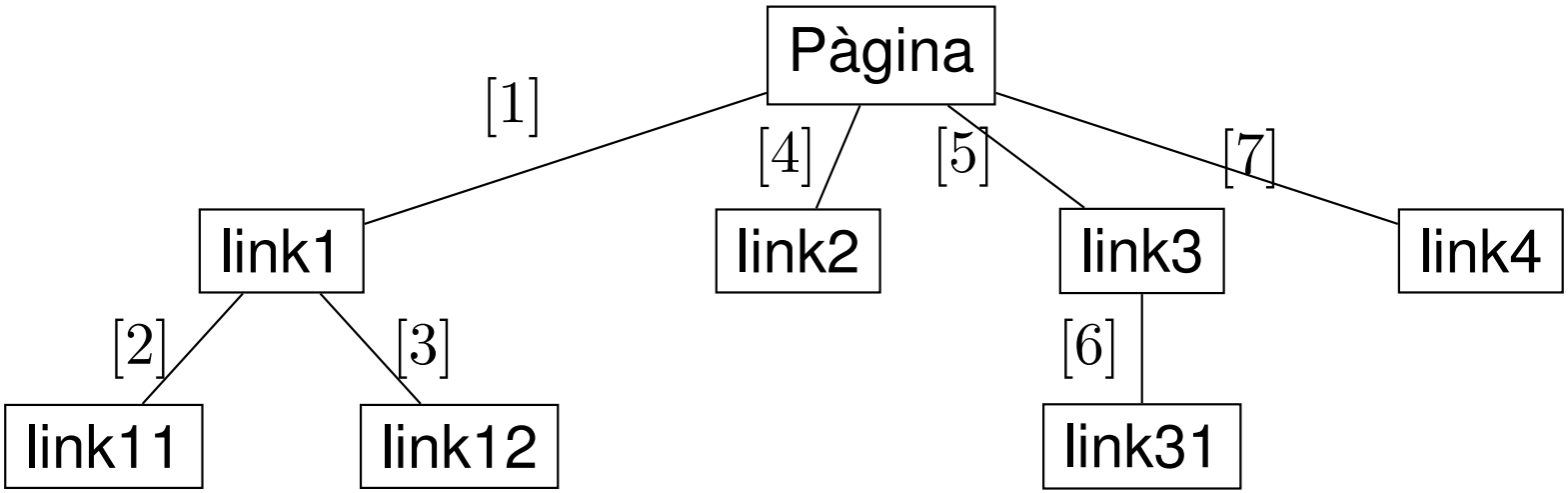
Procés de rastreig, I

L'estratègia d'exploració pot ser:

- Recorregut en profunditat, recorregut en amplada (?)

Procés de rastreig, II

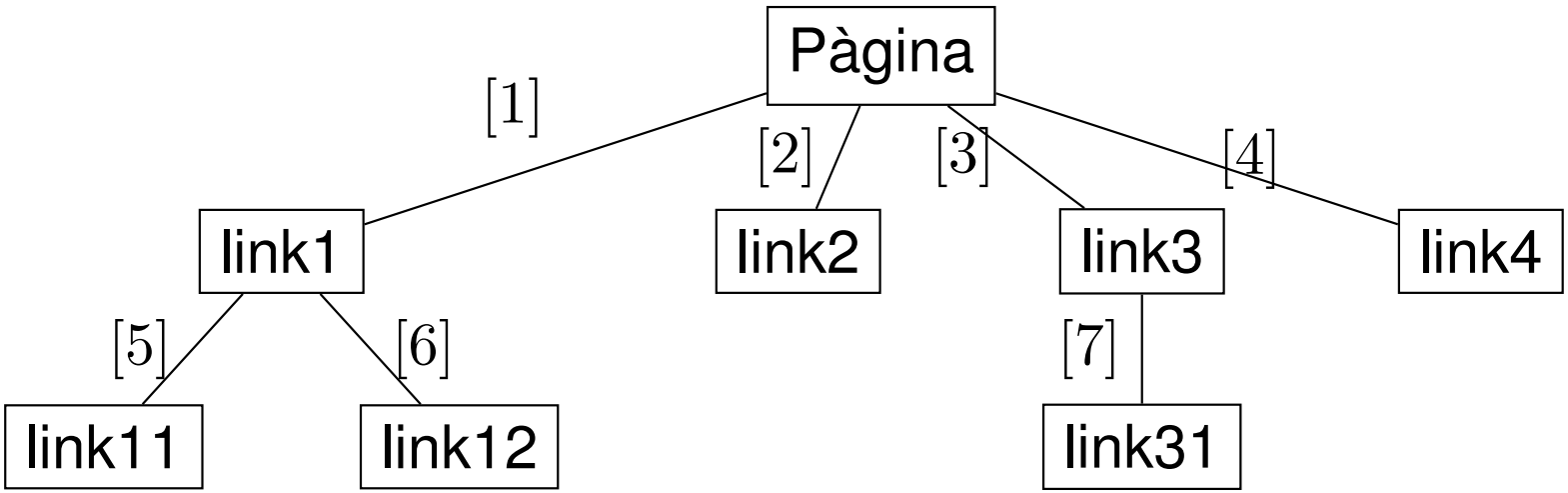
Recorregut en profunditat: Se segueix el primer enllaç d'una cadena de pàgines fins arribar a la més profunda, des de la que es torna recursivament.



Pot perdre's seguint un únic fil d'execució.

Procés de rastreig, III

Recorregut en amplada: S'examinen totes les pàgines a les que s'arriba des de la pàgina actual, i després es visiten les del següent nivell.



Requereix memòria per tots els nodes del nivell previ (exponencial amb la profunditat).

Procés de rastreig, IV

L'estratègia d'exploració pot ser:

- Recorregut en profunditat, recorregut en amplada (?).
- Dirigida –*focused*– (o no): expressar interessos a l'hora de rastrejar.
  - Explorar només les pàgines que pertanyen a un domini.
  - Implícitament, quan es trien les URL inicials.
- Les pàgines de la cua properes a un punt d'interès, s'exploren primer.

Procés de rastreig, IV

L'estratègia d'exploració pot ser:

- Recorregut en profunditat, recorregut en amplada (?).
- Dirigida –*focused*– (o no): expressar interessos a l'hora de rastrejar.
  - Explorar només les pàgines que pertanyen a un domini.
  - Implícitament, quan es trien les URL inicials.
- Les pàgines de la cua properes a un punt d'interès, s'exploren primer.

Cal tenir en compte que:

- Les pàgines s'actualitzen periòdicament $\Rightarrow$ caldrà tornar-les a rastrejar. Totes?

Procés de rastreig, IV

L'estratègia d'exploració pot ser:

- Recorregut en profunditat, recorregut en amplada (?).
- Dirigida –*focused*– (o no): expressar interessos a l'hora de rastrejar.
  - Explorar només les pàgines que pertanyen a un domini.
  - Implícitament, quan es trien les URL inicials.
- Les pàgines de la cua properes a un punt d'interès, s'exploren primer.

Cal tenir en compte que:

- Les pàgines s'actualitzen periòdicament $\Rightarrow$ caldrà tornar-les a rastrejar. Totes?
- Algunes s'actualitzen més sovint que d'altres.

Procés de rastreig, IV

L'estratègia d'exploració pot ser:

- Recorregut en profunditat, recorregut en amplada (?).
- Dirigida –*focused*– (o no): expressar interessos a l'hora de rastrejar.
  - Explorar només les pàgines que pertanyen a un domini.
  - Implícitament, quan es trien les URL inicials.
- Les pàgines de la cua properes a un punt d'interès, s'exploren primer.

Cal tenir en compte que:

- Les pàgines s'actualitzen periòdicament $\Rightarrow$ caldrà tornar-les a rastrejar. Totes?
- Algunes s'actualitzen més sovint que d'altres.
- Començar per les pàgines que s'actualitzen més sovint i que són més populars.

Procés de rastreig, V

Els rastrejadors han de ser:

- eficients
- robustos
- respectuosos

Eficiència en el rastreig, I

- Distribuït: Capacitat d'executar-se de forma distribuïda a través de múltiples màquines.
- Escalable: L'arquitectura ha de permetre afegir més màquines (o ample de banda) per millorar el seu rendiment.

- Les connexions acumulen retards.
- Mantenir moltes (100's?) connexions obertes per màquina.
- Intentar mantenir tots els *threads* ocupats.
- El servidor de DNS sol ser el coll d'ampolla.

Eficiència en el rastreig, II

Cal descartar algunes pàgines:

- Duplicats (25-30% de les pàgines són duplicats!).
  - La detecció ràpida de duplicats és un problema.
  - S'usen *fingerprints* i k-*shingles* (molt semblants als n-grams).
- Irrellevants per l'objectiu del rastrejador (p.e., *focused crawlers*).
- No fiables o *spam*.

Robustesa en el rastreig, I

- URL caigudes: Molt comú. S'usen mecanismes de temps d'espera (*timeout*).
- Pàgines sintàcticament incorrectes.
- Trampes per a aranyes (*spider traps*). Sovint són pàgines que es generen dinàmicament.
- *Web spam* (intents "intencionats" de manipular el rànquing dels motors de cerca usant termes clau)

Cortesia en el rastreig, I

- No visitar el mateix servidor massa sovint, esp. descàrregues.
- Inserir un interval de temps entre peticions successives al mateix servidor.
- Respectar els protocols d'exclusió per a robots:
  - El fitxer `/robots.txt` conté les preferències de l'administrador de la pàgina/lloc web.
  - "Si ets l'agent X, si us plau no exploris el directori Y"

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

Cortesia en el rastreig, II

- Excloure directoris:

```
User-agent: *
Disallow: /tmp/
Disallow: /cgi-bin/
Disallow: /users/paranoid/
```

- Excloure un robot:

```
User-agent: GoogleBot
Disallow: /
```

- Permetre un robot concret:

```
User-agent: GoogleBot
Disallow:

User-agent: *
Disallow: /
```

Índex

El procés de la RI, III

Extracció de termes

Accions potencials:

- *Parsing*: extracció de l'estructura (p.e., HTML)
- *Tokenization*: descompondre les seqüències de caràcters en unitats individuals
- Enriquiment: afegir informació addicional a les unitats
- O bé Lematització o bé *Stemming*: reduir les paraules a la seva arrel o lema

Tokenization, I

Agrupació de caràcters

Ajuntar caràcters consecutius per formar “paraules”: s’usen els espais i els signes de puntuació per determinar els seus límits.
Similar a l’anàlisi lèxica en els compiladors.

Tokenization, I

Agrupació de caràcters

Ajuntar caràcters consecutius per formar “paraules”: s’usen els espais i els signes de puntuació per determinar els seus límits.
Similar a l’anàlisi lèxica en els compiladors.

Dificultats:

De vegades, un terme inclou els seus delimitadors:

- Espais (“David A. Mix Barrington”, “Fahrenheit 451”, “September 11, 1714”),
- puntuació (adreces IP, “I.B.M.”, “753 B.C.”),
- o altres signes (“R+D”, “H&M”, “C#”).

Tokenization, I

Agrupació de caràcters

Ajuntar caràcters consecutius per formar “paraules”: s’usen els espais i els signes de puntuació per determinar els seus límits.
Similar a l’anàlisi lèxica en els compiladors.

Dificultats:

De vegades, un terme inclou els seus delimitadors:

- Espais (“David A. Mix Barrington”, “Fahrenheit 451”, “September 11, 1714”),
- puntuació (adreces IP, “I.B.M.”, “753 B.C.”),
- o altres signes (“R+D”, “H&M”, “C#”).
- Els guionets són especialment complicats:
  - I si s’ometen? Transformar “cultura afro-americana” en “cultura afroamericana ” sembla bé, però

Tokenization, I

Agrupació de caràcters

Ajuntar caràcters consecutius per formar “paraules”: s’usen els espais i els signes de puntuació per determinar els seus límits.
Similar a l’anàlisi lèxica en els compiladors.

Dificultats:

De vegades, un terme inclou els seus delimitadors:

- Espais (“David A. Mix Barrington”, “Fahrenheit 451”, “September 11, 1714”),
- puntuació (adreces IP, “I.B.M.”, “753 B.C.”),
- o altres signes (“R+D”, “H&M”, “C#”).
- Els guionets són especialment complicats:
  - I si s’ometen? Transformar “cultura afro-americana” en “cultura afroamericana ” sembla bé, però
  - “pre-ocupació” no és “preocupació”,
  - i a “vols barats San Francisco-Los Angeles”, “FranciscoLos” seria incorrecte.

Tokenization, II

Reduir la llista de tokens identificant-los

No tots els tokens seran usats com a termes.

Tokenization, II

Reduir la llista de tokens identificant-los

No tots els tokens seran usats com a termes.

Case folding:

Canviar tots els caràcters a minúscula de manera que les cerques siguin independents de majúscules/minúscules.

Tokenization, II

Reduir la llista de tokens identificant-los

No tots els tokens seran usats com a termes.

Case folding:

Canviar tots els caràcters a minúscula de manera que les cerques siguin independents de majúscules/minúscules.
Però cal anar amb compte

Tokenization, II

Reduir la llista de tokens identificant-los

No tots els tokens seran usats com a termes.

Case folding:

Canviar tots els caràcters a minúscula de manera que les cerques siguin independents de majúscules/minúscules.
Però cal anar amb compte

- "USA" no hauria de ser "usa",
- "Windows" potser no hauria de ser "windows"...

Una àrea de recerca molt activa és el reconeixement d'entitats amb nom (*Named Entity Recognition*).

Tokenization, III

Reduir la llista de tokens suprimint-ne

Eliminar les paraules funcionals o buides (*stopwords*): paraules comunes que apareixeran en tots els documents o que no ajudaran a determinar el contingut d'un document

- preposicions
- articles
- paraules com "essencialment", "aleshores"...
- verbs molt comuns com "és", "ha", "haurà"...
- ...però vigileu, "pot" com a nom no és una *stopword*!

Tokenization, III

Reduir la llista de tokens suprimint-ne

Eliminar les paraules funcionals o buides (*stopwords*): paraules comunes que apareixeran en tots els documents o que no ajudaran a determinar el contingut d'un document

- preposicions
- articles
- paraules com "essencialment", "aleshores"...
- verbs molt comuns com "és", "ha", "haurà"...
- ...però vigileu, "pot" com a nom no és una *stopword*!

Aquestes accions poden reduir la mida de l'índex en un 40%.

Tokenization, III

Reduir la llista de tokens suprimint-ne

Eliminar les paraules funcionals o buides (*stopwords*): paraules comunes que apareixeran en tots els documents o que no ajudaran a determinar el contingut d'un document

- preposicions
- articles
- paraules com "essencialment", "aleshores"...
- verbs molt comuns com "és", "ha", "haurà"...
- ...però vigileu, "pot" com a nom no és una *stopword*!

Aquestes accions poden reduir la mida de l'índex en un 40%.

Altres opinions: `La potència de càlcul dels ordinadors actuals podria mantenir l'índex sencer i que sigui la rellevància dels documents trobats la que faci de filtre.`

Tokenization, IV

Resum

- Depèn de l'idioma...

Tokenization, IV

Resum

- Depèn de l'idioma...
- Depèn de l'aplicació...
  - cerca en una biblioteca?
  - cerca en una intranet?
  - cerca a internet?

Tokenization, IV

Resum

- Depèn de l'idioma...
- Depèn de l'aplicació...
  - cerca en una biblioteca?
  - cerca en una intranet?
  - cerca a internet?
- Crucial per l'eficiència en la recuperació.

Tokenization, IV

Resum

- Depèn de l'idioma...
- Depèn de l'aplicació...
  - cerca en una biblioteca?
  - cerca en una intranet?
  - cerca a internet?
- Crucial per l'eficiència en la recuperació.
- Requereix una feina laboriosa implementar en un sistema de RI un conjunt enorme de regles i excepcions.

Enriquiment

Propostes

Enriquiment significa que a cada terme se li associa una informació addicional que pot ser útil per recuperar documents "correctes".

- Sinònims: eina $\rightarrow$ estri
- Paraules o definicions relacionades: portàtil $\rightarrow$ ordinador portàtil
- Categories: golf $\rightarrow$ esports
- Etiquetes sintàctiques (*POS tags*, *part of speech labels*)

Enriquiment

Propostes

Enriquiment significa que a cada terme se li associa una informació addicional que pot ser útil per recuperar documents "correctes".

- Sinònims: eina $\rightarrow$ estri
- Paraules o definicions relacionades: portàtil $\rightarrow$ ordinador portàtil
- Categories: golf $\rightarrow$ esports
- Etiquetes sintàctiques (*POS tags*, *part of speech labels*)

Cal tenir en compte l'ambigüitat i la dependència del context:

escalar: un esport?

escalar: redimensionar?

També hi ha una àrea de recerca molt activa *Word Sense Disambiguation*.

Lematització i *Stemming*, I

Dues opcions

Lematització: reduir les paraules a la seva arrel lingüística.
De vegades només s'analitzen sufixos:

nedar, nedador, nedada, nedo $\rightarrow$ ned

Lematització i *Stemming*, I

Dues opcions

Lematització: reduir les paraules a la seva arrel lingüística.
De vegades només s'analitzen sufixos:

nedar, nedador, nedada, nedo $\rightarrow$ ned

Però de vegades no és tan senzill:

sóc, era, és, serà $\rightarrow$ ser

plans $\rightarrow$ pla

rojos $\rightarrow$ roig

Lematització i *Stemming*, I

Dues opcions

Lematització: reduir les paraules a la seva arrel lingüística.
De vegades només s'analitzen sufixos:

nedar, nedador, nedada, nedo $\rightarrow$ ned

Però de vegades no és tan senzill:

sóc, era, és, serà $\rightarrow$ ser

plans $\rightarrow$ pla

rojos $\rightarrow$ roig

Stemming és un procés alternatiu que elimina sufixos i prefixos, amb canvis de lletres ocasionals. Substitueix la lematització, amb resultats "prou bons" i sent més senzill i ràpid.

Lematització i *Stemming*, II

L'opció *stemming*

Stemmers

es basen en regles que indiquen quan i quin prefix o sufix pot ser eliminat o reescrit.

Lematització i *Stemming*, II

L'opció *stemming*

Stemmers

es basen en regles que indiquen quan i quin prefix o sufix pot ser eliminat o reescrit.

- Funciona bé per l'anglès i les llengües romàniques.
- L'algorisme més famós és el de Porter stemmer.
  - Disponible com a classe a Lucene, com a node a KNIME, i està implementat en molts llenguatges de programació...
  - `http://tartarus.org/martin/PorterStemmer`
- Un algorisme similar, més evolucionat, basat en els mateixos principis és l'Snowball stemmer.

Lematització i *Stemming*, III

Un exemple d'aplicació de l'algorisme de Porter

L'algorisme de Porter

és l'algorisme de *stemming* més utilitzat. En aquest exemple, apliquem primer un preprocés (*case folding* i eliminació de signes de puntuació) i després l'algorisme de Porter:

Lematització i *Stemming*, III

Un exemple d'aplicació de l'algorisme de Porter

L'algorisme de Porter

és l'algorisme de *stemming* més utilitzat. En aquest exemple, apliquem primer un preprocés (*case folding* i eliminació de signes de puntuació) i després l'algorisme de Porter:

> *Stemming is a process that simply removes suffixes and prefixes, with occassional letter changes; it is a replacement for lemmatizing, with good enough results, and much simpler and faster.*

Lematització i *Stemming*, III

Un exemple d'aplicació de l'algorisme de Porter

L'algorisme de Porter

és l'algorisme de *stemming* més utilitzat. En aquest exemple, apliquem primer un preprocés (*case folding* i eliminació de signes de puntuació) i després l'algorisme de Porter:

Stemming is a process that simply removes suffixes and prefixes, with occassional letter changes; it is a replacement for lemmatizing, with good enough results, and much simpler and faster.

stem i a process that simpli remov suffix and prefix with occassion letter chang it i a replac for lemmat with good enough result and much simpler and faster

Lematització i *Stemming*, IV

Analitzant els *stemmers*

Analitzant l'algorisme de Porter:

Està format per unes 60 regles, distribuïdes en 5 fases.

Lematització i *Stemming*, IV

Analitzant els *stemmers*

Analitzant l'algorisme de Porter:

Està format per unes 60 regles, distribuïdes en 5 fases.

- Algunes regles només s'apliquen en determinades condicions (paraules prou llargues, que ja s'hagin aplicat o no d'altres regles...)
- A cada fase, s'escull la regla que es pugui aplicar al sufix més llarg possible. Algunes regles d'exemple:
  - A la fase 1, "sses" substituït per "ss" ("caresses" $\rightarrow$ "caress")
  - A la fase 2, "(>1)ement" eliminat ("replacement" $\rightarrow$ "replac" però no s'aplica a "cement")

Lematització i *Stemming*, V

Pros i Cons de l'*stemming*

Pros:

- millora lleument l'efectivitat de la recuperació
- és gairebé tan eficaç com la lematització
- requereix menys coneixement de l'idioma que la lematització
- és més ràpid, senzill de descriure i d'implementar

Lematització i *Stemming*, V

Pros i Cons de l'*stemming*

Pros:

- millora lleument l'efectivitat de la recuperació
- és gairebé tan eficaç com la lematització
- requereix menys coneixement de l'idioma que la lematització
- és més ràpid, senzill de descriure i d'implementar

Cons:

- el resultat és il·legible (humà), sovint contraintuïtiu
- pot reduir al mateix *stem* paraules molt diferents

Índex

Constitució dels textos

En els textos, alguns termes són **molt** freqüents i d'altres són **molt poc** freqüents.

Qüestions bàsiques:

- Quantes paraules diferents usem freqüentment?
- Com de freqüents són les paraules freqüents?
- Podem formalitzar-ho d'alguna manera?

Hi ha lleis empíriques bastant precises en la majoria d'idiomes.

Estadística dels textos, II

En molts fenòmens, relacionats amb humans o no, la distribució de probabilitat “decreix lentament” comparada amb una gaussiana o una exponencial.

Això significa que els objectes molt poc freqüents tenen un pes notable sobre el total. Alguns casos:

- textos, com va ser mostrat per Zipf
- la distribució dels noms de les persones
- la popularitat d’un lloc web
- la riquesa de les persones, les companyies o els països
- nombre d’enllaços a les pàgines web més populars
- intensitat dels terratrèmols

Estadística dels textos, III

- Un document conté una gran varietat de termes.
- Alguns termes apareixen més freqüentment que d'altres en documents d'un determinat tema.
- La distribució de la freqüència dels termes està molt esbiaixada:

  Les dues paraules més freqüents en anglès ("the", "of") suposen el 10% del total d'ocurrències.
  Les 6 paraules més freqüents suposen el 20% del total.
  Les 50 paraules més freqüents suposen el 40% del total.
  Gairebé la meitat de les paraules només apareixen un cop.
- Poques paraules apareixen molt sovint mentre que moltes paraules no apareixen gairebé mai.

Estadística dels textos, IV

Si analitzem un text (millor llarg), comptem les vegades que surt cada paraula i les ordenem segons la freqüència d'ús, de manera que:

- *r* és el lloc o rang que ocupa certa paraula en aquesta llista
- *f* és la freqüència amb què s'usa aquesta paraula (és a dir, el nombre de vegades que hi apareix)

observem la llei de Zipf: La freqüència d'aparició d'una paraula és inversament proporcional al rang de la paraula elevat a un exponent α.

$$f(r) = k * r^{-\alpha} = \frac{k}{r^{\alpha}}$$

Estadística dels textos, V

Llei de Potència (*power law*)

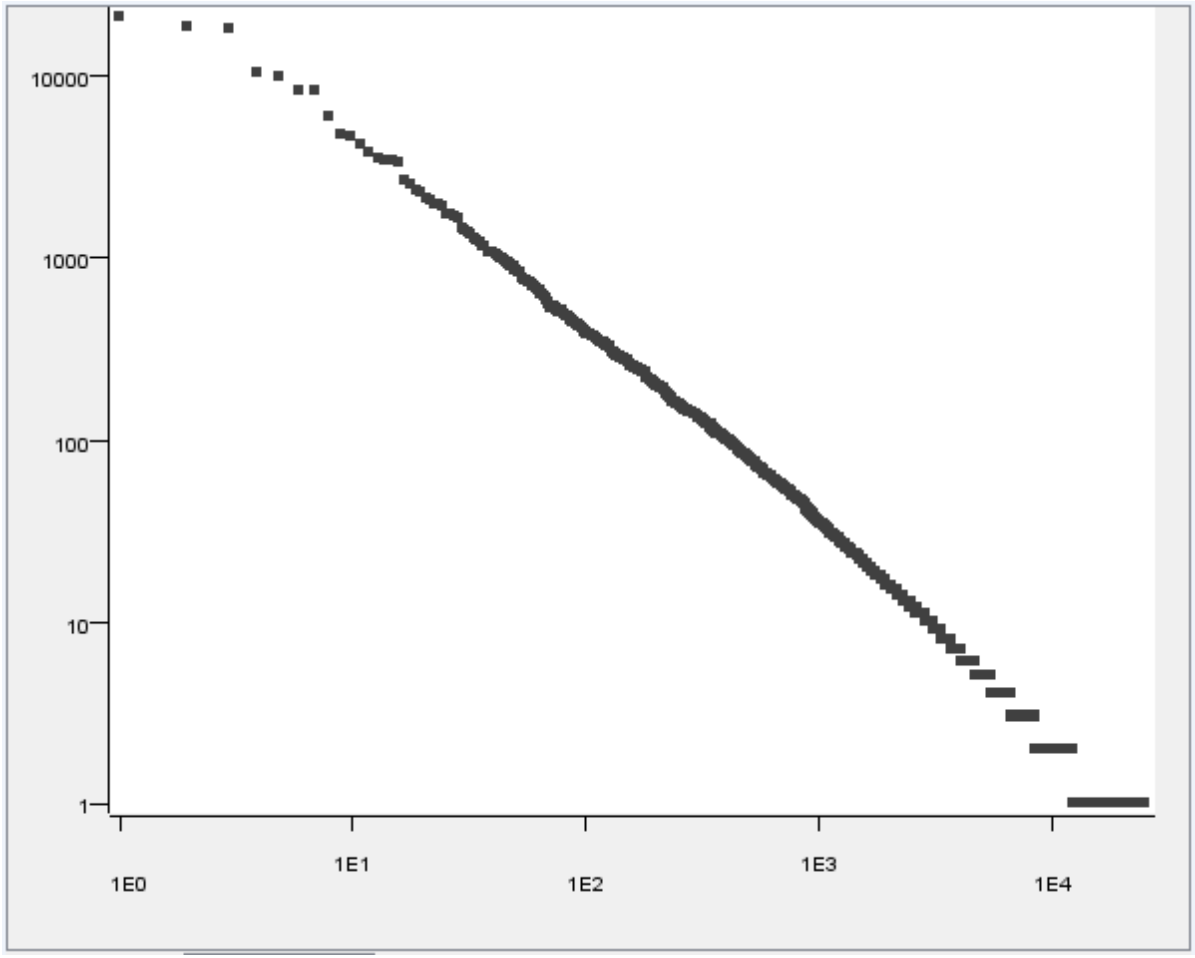
Com detectar *power laws*?

- Ordeneu els elements de forma decreixent segons la seva freqüència.
- Mostreu gràficament les freqüències respecte la seva posició (rang) dins la seqüència.
- Ajusteu-la per obtenir una gràfica log-log:

  És a dir, aplicant a ambdós eixos una escala logarítmica.
- Aleshores veureu alguna cosa semblant a una línia recta.
- Useu aquesta gràfica per identificar l'exponent (pendent de la línia recta).

Estadística dels textos, VI

La llei de Zipf



Freqüències de les paraules al Don Quijote (escala log-log).

Estadística dels textos, VII

Observacions sobre la llei de Zipf

- El paràmetre d'interès és l'α:
  - el valor original proposat per Zipf era $\alpha = 1$
  - pel Don Quijote, després de *case folding* i d'eliminació de la puntuació, la recta entre "caballero" (646, rang 64) i "toscano" (5, rang 5141 en mig de tots els termes de freqüència 5) dóna

$$\alpha = 1.1$$

  - els valors d'$\alpha \in [1.5, 2]$ són comuns
- Una altra variant més flexible de la llei de Zipf, proposada per Mandelbrot, presenta la forma següent

$$f = \frac{k}{(c + r)^{\alpha}}$$

Estadística dels textos, VIII

Quantitat de termes en ús

A mida que el corpus (col·lecció de documents) creix, també creix el vocabulari.

No obstant,

la quantitat de termes nous que van apareixent és menor quan el corpus ja és prou gran.

Estadística dels textos, VIII

Quantitat de termes en ús

A mida que el corpus (col·lecció de documents) creix, també creix el vocabulari.

No obstant,

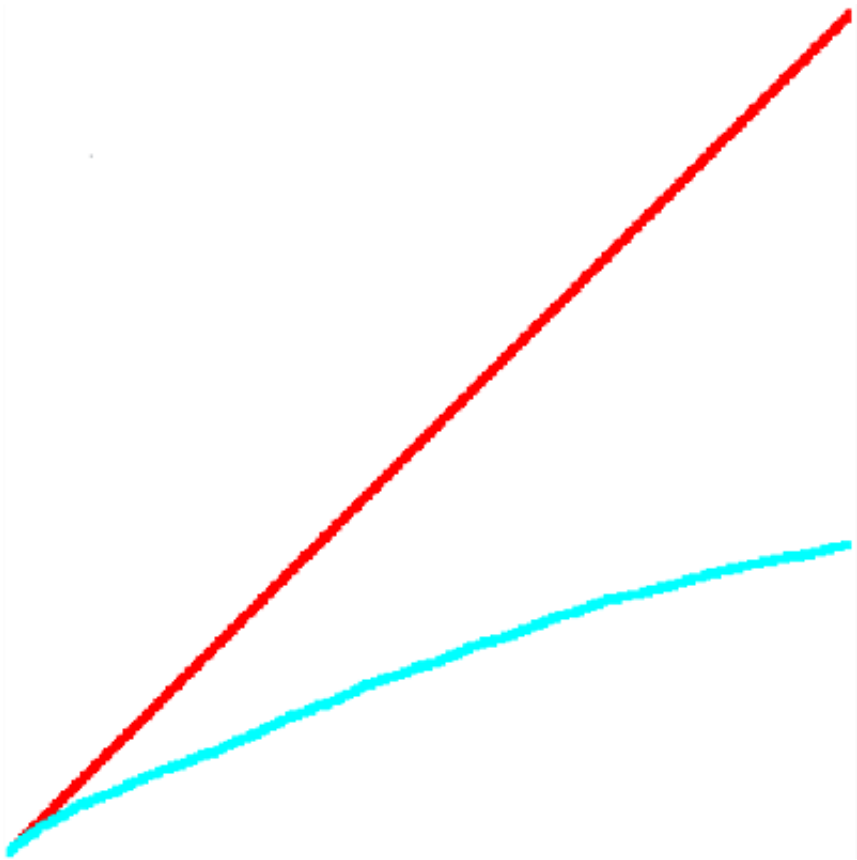
la quantitat de termes nous que van apareixent és menor quan el corpus ja és prou gran.

- Les primeres 2500 paraules del Don Quijote contenen unes 1100 paraules diferents.
- El Don Quijote sencer conté unes 383000 paraules però menys de 40000 són diferents.

Estadística dels textos, IX

Les primeres 2500 paraules de l'article de Vannevar Bush

(La línia blava mostra la quantitat de paraules diferents.)



Estadística dels textos, X

Llei de Heaps (o llei de Herdan)

El nombre de paraules noves

que es troben en una quantitat donada de textos nous, decreix a mida que el corpus creix. Aquesta relació es pot descriure amb una polinòmica de grau menor que 1.

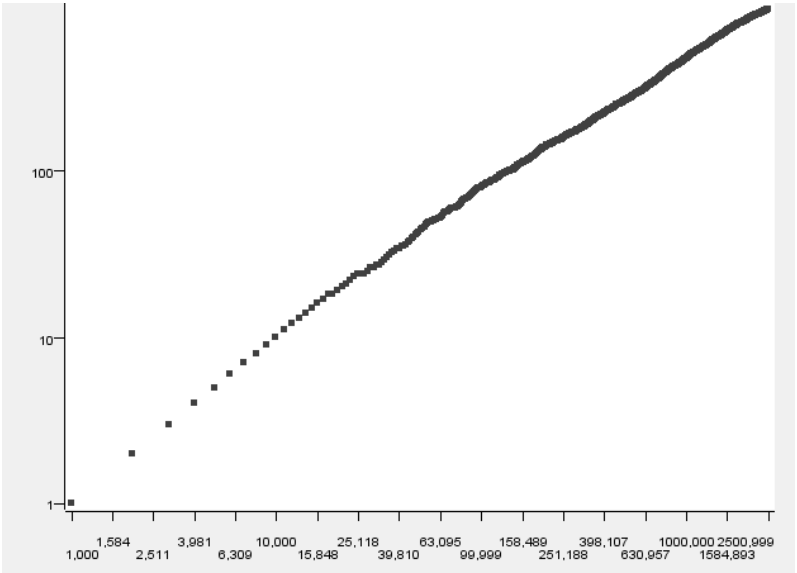
Estadística dels textos, X

Llei de Heaps (o llei de Herdan)

El nombre de paraules noves

que es troben en una quantitat donada de textos nous, decreix a mida que el corpus creix. Aquesta relació es pot descriure amb una polinòmica de grau menor que 1.

També podem veure-ho ajustant la gràfica a escala logarítmica (*log-log plot*). La corba blava es torna "més recta".



Estadística dels textos, XI

Fórmula per la llei de Heaps

En un text de mida N:

Hi trobarem d paraules diferents; quina és la relació entre d i N?

Si mirem la recta obtinguda en l'escala logarítmica, tenim:

$$\log d = k_1 + \beta * \log N, \text{ és a dir, } d = k * N^{\beta} \text{ (on } k_1 = \log k)$$

- β depèn de l'idioma i del tipus de text
- a l'article de Vannevar Bush, trobem $\beta \approx 0.836$; al Don Quijote $\beta \approx 0.806$
- per l'anglès són freqüents els valors de β entre 0.5 i 0.6
- si el corpus és molt gran el vocabulari serà quasi fix

  (però compte: errors tipogràfics, noms propis, paraules foranes...).