

Temu-Kembali Informasi 2019

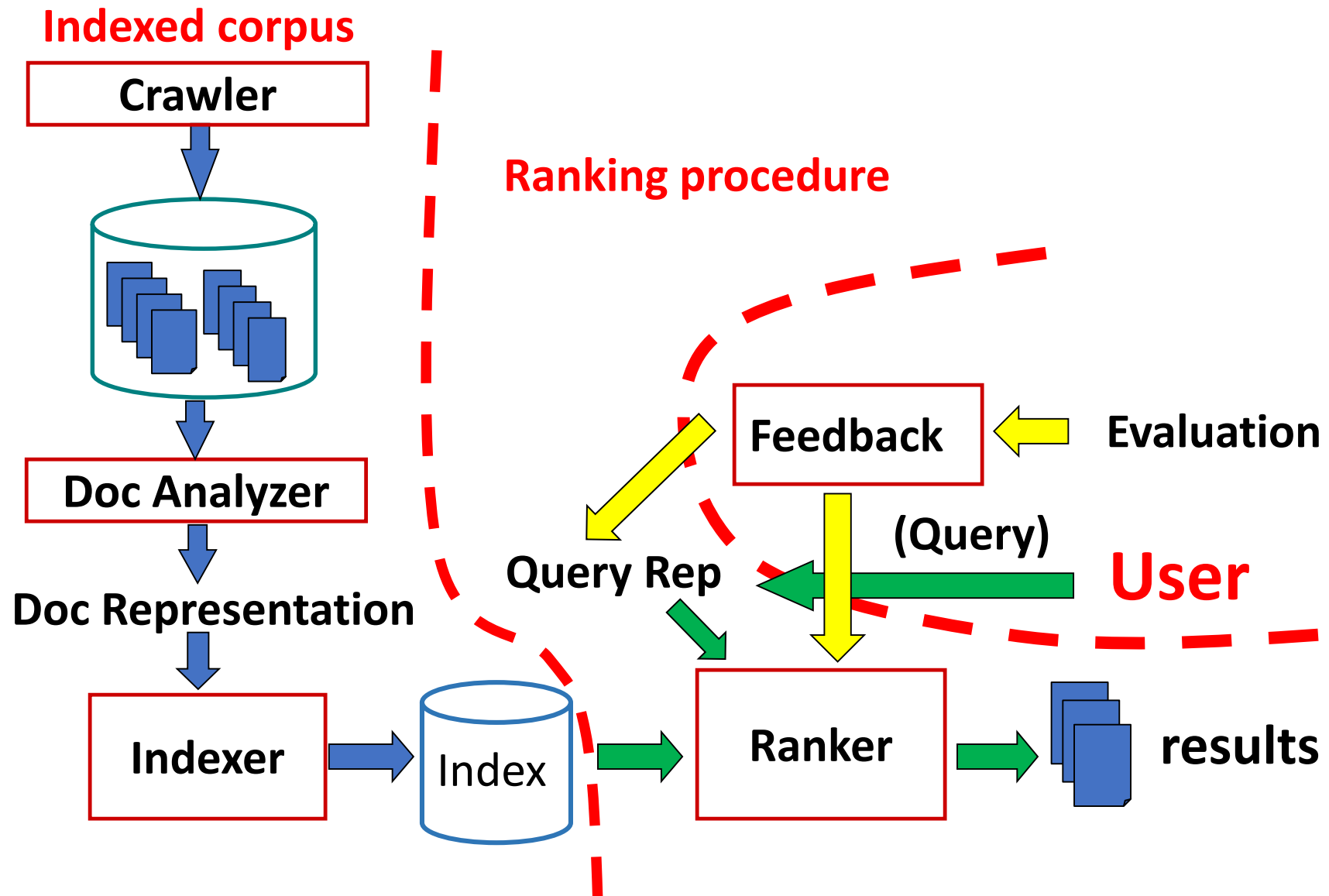
02B: Web Crawling & Analisis Teks

Husni

Husni.trunojoyo.ac.id

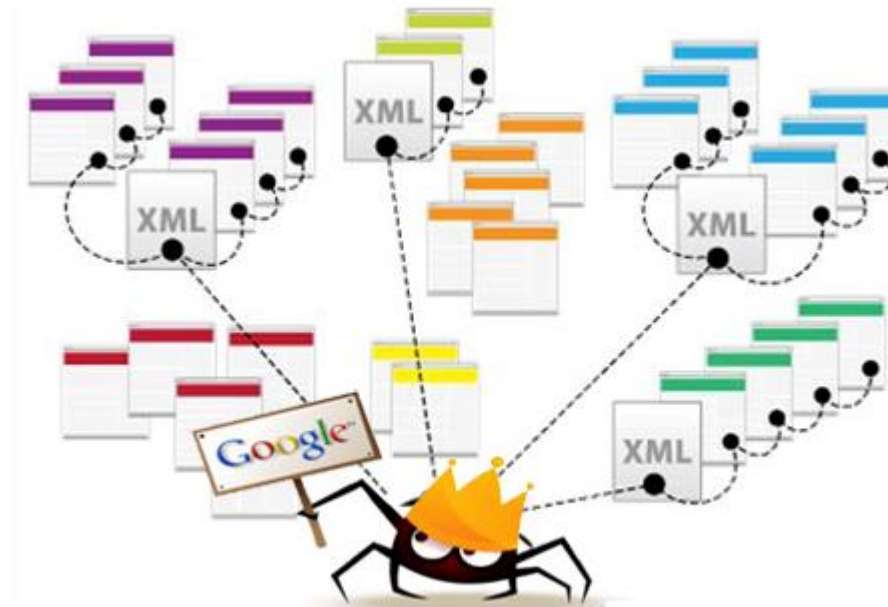
Outline

Abstraksi dari Arsitektur Search Engine



Web Crawler

- Suatu program otomatis yang secara sistematis menjelajah web dengan tujuan mengindeks dan mengupdate isi web
 - Sinonim: spider, robot, bot



Bagaimana Crawler Bekerja

- Dalam bentuk *pseudo code*

```
Def Crawler(entry_point) {  
    URL_list = [entry_point]  
    while (len(URL_list)>0) {  
        URL = URL_list.pop();  
        if (isVisited(URL) or !isLegal(URL) or !checkRobotsTxt(URL))  
            continue;  
        HTML = URL.open();  
        for (anchor in HTML.listOfAnchors()) {  
            URL_list.append(anchor);  
        }  
        setVisited(URL);  
        insertToIndex(HTML);  
    }  
}
```

Halaman mana dikunjungi berikutnya?

*Apakah ini sudah dikunjungi?
Atau haruskah dikunjungi lagi?*

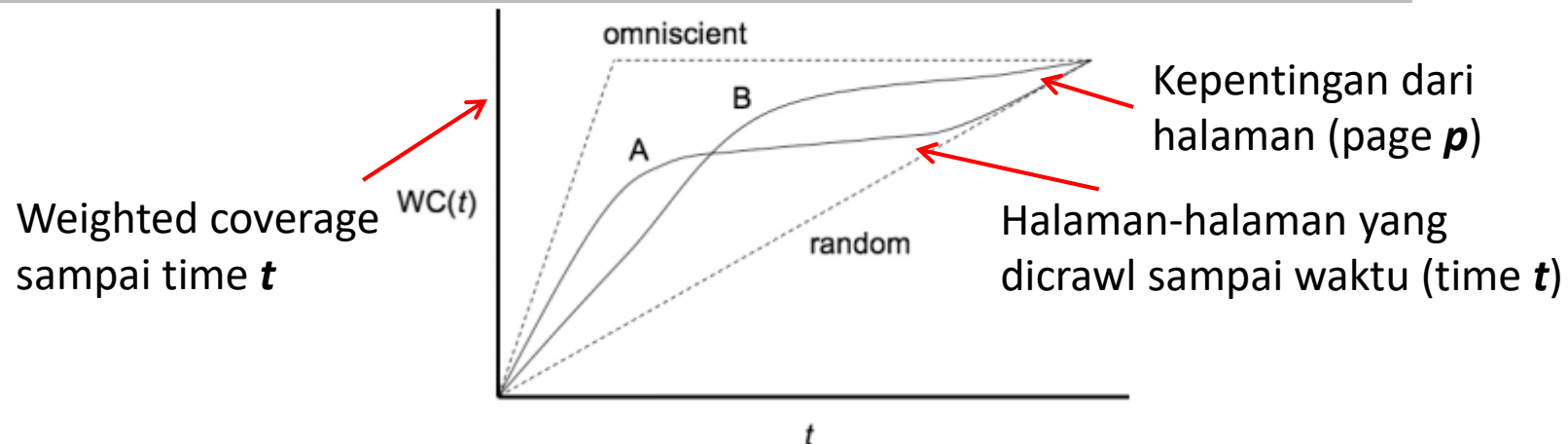
Apakah akses diijinkan?

Strategi Kunjungan

- Melebar dahulu (*breadth first*)
 - Secara seragam menjelajahi dari halaman masuknya
 - Mengingat semua node pada level sebelumnya
 - Sebagaimana ditunjukkan dalam *pseudo code*
- Mendalam lebih dahulu (*depth first*)
 - Menjelajahi web mengikuti cabangnya lebih dahulu
 - Perayapan web sesungguhnya tidaklah membentuk struktur pohon
- Penjelajahan terfokus (*focused crawling*)
 - Prioritas pada link-link baru dengan strategi yang telah terdefinisi.

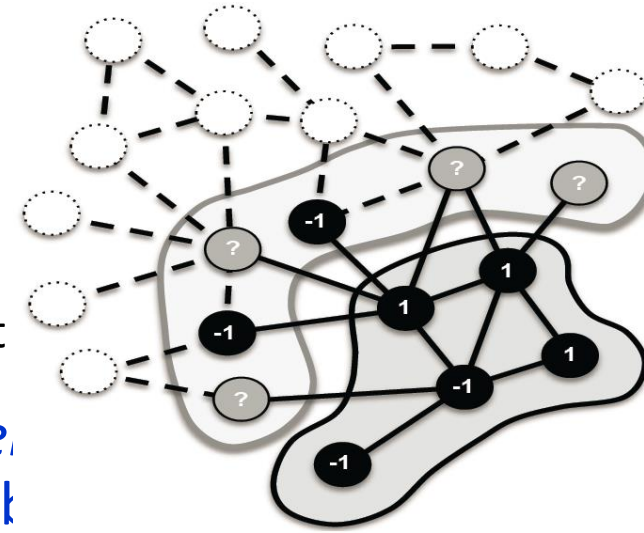
Focused Crawling

- Ada prioritas dalam urutan kunjungan halaman web
 - Ukuran web terlalu besar bagi web crawler untuk ditangani secara lengkap (bahkan oleh google)
 - Pada 1999, tidak ada search engine yang mengindeks lebih dari 16% Web
 - Pada 2005, search engine skala besar mengindeks tidak lebih dari 40-70% dari web yang dapat diindeks.



Focused Crawling

- Prioritas berdasarkan in-degree [Cho et al. WWW'07, Cho and Uri VLDB'07]
 - Halaman dengan jumlah *incoming hyperlinks* telah dicrawl sebelumnya akan dicrawl lebih dahulu
- Prioritas berdasarkan PageRank [Liu et al. SIGIR'09]
 - Breadth-first pada tree PageRank secara bertahap
 - Lebih konsisten dengan PageRank

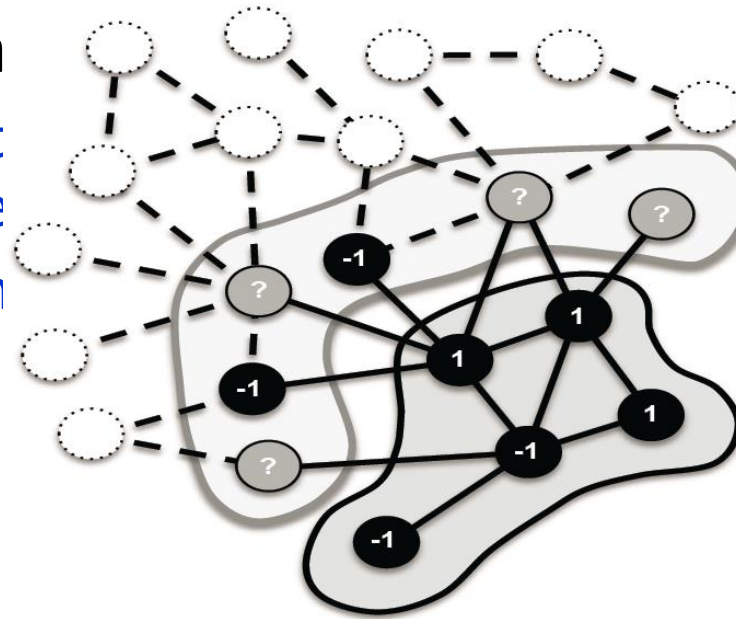


ian yang

WW'07, Cho and Uri VLDB'07]

ghitung/memperkirakan

ly et al. SIGIR'09]



Focused Crawling

- Prioritas berdasarkan relevansi topik
 - Pada vertical search, hanya meng-crawl halaman yang relevan [De et al. WWW'94]
 - Mis. Search engine restoran hanya akan mengcrawl halaman terkait restoran
 - Estimasi kemiripan dengan halaman sekarang (current) berdasarkan pada anchor text atau teks dekat anchor [Hersovici et al. WWW'98]
 - Taksonomi diberikan pengguna atau *topical classifier* [Chakrabarti et al. WWW'98]

Menghindari Kunjungan Duplikat

- Web lebih berbentuk graf daripada tree, menghindari loop dalam crawling adalah penting
- Bagaimana mengetahuinya?
 - Gunakan tree atau hash table!
- Apa yang perlu dicek?
 - URL: harus dinormalisasi, dapat menghindari semua duplikasi
 - <http://dl.acm.org/event.cfm?id=RE160&CFID=516168213&CFTOKEN=99036335>
 - <http://dl.acm.org/event.cfm?id=RE160>
 - Page: perubahan kecil dapat mengakibatkan *misfire* (gagal)
 - Perubahan timestamp, ID data center dalam HTML

Kebijakan Kesopan-santunan

- Crawler dapat meretrieve data jauh lebih cepat dan sangat lebih mendalam daripada pencari manusia
- Biaya (*cost*) menggunakan Web crawlers
 - Sumber daya jaringan (*network*)
 - Server berbeban lebih (*overload*)
- Protokol Eksklusi Berbentuk Robots
 - Contoh: [CNN](#)

Contoh Protokol Eksklusi Robot

- Melarang akses ke direktori tertentu:

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

- Melarang akses suatu robot spesifik:

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

Membolehkan suatu robot spesifik:

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

```
User-agent: *  
Disallow: /
```

Menganalisis Halaman Web yang Dicrawl

- Apa yang diperhatikan dari halaman web yang dirayapi

The screenshot shows the CNN Tech website interface. At the top, there's a red header with navigation links for 'SET EDITION: U.S.', 'INTERNATIONAL', 'MÉXICO', and 'ARABIC'. Below this, there's a search bar and a 'POWERED BY Google' logo. The main navigation bar includes links for 'Home', 'TV & Video', 'U.S.', 'World', 'Politics', 'Justice', 'Entertainment', 'Tech' (highlighted), 'Health', 'Living', 'Travel', 'Opinion', 'iReport', 'Money', and 'Sports'. The main content area features a large article titled 'Apple profits soar while iPads slump' with a photo of people in an Apple store. To the right, there's a 'WHAT MATTERS' section with three smaller articles: '12 real people who appeared in video games', 'Manuel Noriega sues over 'Call of Duty'', and ''530 Fatties' page targets the overweight'. Below the main article, there's a 'TECH PULSE' section with a tweet from @cnn tech. The bottom section is titled 'LATEST NEWS' and features an article about 'Nanomaterial cloaks the sense of touch'. On the right side of the bottom section, there's a 'SUBSCRIBE' button and a 'SPECIAL COVERAGE' section titled 'What will the city of tomorrow look like?'.

SET EDITION: U.S. | INTERNATIONAL | MÉXICO | ARABIC
TV: CNN | CNNi | CNN en Español | HLN
Sign up | Log in
CNN Tech
POWERED BY Google
Home | TV & Video | U.S. | World | Politics | Justice | Entertainment | Tech | Health | Living | Travel | Opinion | iReport | Money | Sports
updated 11:26 AM EDT, Wed July 23, 2014

Apple profits soar while iPads slump

WHAT MATTERS

- 12 real people who appeared in video games
f 202 73
- Manuel Noriega sues over 'Call of Duty'
f 3406 1567
- '530 Fatties' page targets the overweight
f 407 0

TECH PULSE
(+) FOLLOW ON TWITTER

CNN Tech
@cnn tech Meet Dark Mail: Email that hides from the NSA
http://t.co/4yGTHKVvuy via @Jose_Pagliery http://t.co/XntyWtxwOX
53 minutes ago via TweetDeck | RETWEET

CNN Tech
@cnn tech Apple profit as iPads slump! http://t.co/...
20 hours ago via TweetDeck

LATEST NEWS
FILTER: ☒ ALL ☐ SOCIAL MEDIA ☐ MOBILE ☐ WEB ☐ GAMING & GADGETS ☐ INNOVATION ☐ TECH BIZ
updated 1 hour, 14 minutes ago
Now you feel it, now you don't: Nanomaterial cloaks the sense of touch
A real invisibility cloak may still be the stuff of fantasy, but scientists have figured out a way to hide objects from touch.
FULL STORY

SUBSCRIBE

SPECIAL COVERAGE
What will the city of tomorrow look like?
Imagine turning the ocean into drinking water, ending traffic jams, making policing smarter and creating buildings with zero carbon footprints. This city of tomorrow may already be here.

Menganalisis Halaman Web yang Dicrawl

- Apa yang didapat mesin dari halaman web yang dirayapi

```
<!DOCTYPE HTML>
<html lang="en-US">
<head>
<title>Technology News - Computers, Internet, Invention and Innovation Tech from CNN.com</title>
<meta http-equiv="content-type" content="text/html; charset=utf-8"/>
<meta http-equiv="last-modified" content="2014-07-23T15:25:56Z"/>
<meta name="robots" content="index, follow"/>
<meta name="googlebot" content="noarchive"/>
<meta name="viewport" content="width=1024"/>
<meta name="title" content="Technology News - Computers, Internet, Invention and Innovation Tech from CNN.com"/>
<meta name="description" content="Find information about the latest advances in technology at CNN. CNN Technology news and video covers the internet, business and personal tech, video games, and more."/>
<meta name="keywords" content="CNN, CNN news, CNN.com, CNN TV, news, news online, breaking news, U.S. news, world news, weather, business, CNN Money, sports, politics, law, technology, entertainment, education, travel, health, special reports, autos, developing story, news video, CNN Intl"/>
<link rel="canonical" href="http://www.cnn.com/TECH/">
<link type="image/png" rel="apple-touch-icon" href="http://i.cdn.turner.com/cnn/.e/img/3.0/global/misc/apple-touch-icon.png"/>
<link type="application/rss+xml" rel="alternate" href="http://rss.cnn.com/rss/cnn_tech.rss" title="CNN - Tech [RSS]"/>
<link type="application/rss+xml" rel="alternate" href="http://rss.cnn.com/rss/cnn_topstories.rss" title="CNN - Top Stories [RSS]"/>
<link type="application/rss+xml" rel="alternate" href="http://rss.cnn.com/rss/cnn_latest.rss" title="CNN - Recent Stories [RSS]"/>
<link type="application/opensearchdescription+xml" rel="search" href="/tools/search/cnncom.xml" title="CNN.com"/>
<link type="application/opensearchdescription+xml" rel="search" href="/tools/search/cnncomvideo.xml" title="CNN.com Video"/>
<link href="https://plus.google.com/u/0/b/117515799321987910349/117515799321987910349/posts" rel="publisher"/>
<link type="text/css" rel="stylesheet" href="http://z.cdn.turner.com/cnn/tmpl_asset/static/www_section/2695/css/techlib-min.css"/>
<script>
var cnnCVPAdSection='cnn.com_technology_section_homepage',
cnnIsSectionPage=true,
cnnSectionName='Tech',
cnnSectionFront='Tech',
sectionName='tech';
</script>
<script src="http://z.cdn.turner.com/cnn/tmpl_asset/static/www_section/2695/js/techlib-min.js"></script>
<script>
var cnnPageType="Section";
if(typeof(cnn_metadata)=='undefined'){var cnn_metadata={};}
var cnn_edtnswtchver='www';
cnn_metadata.section=['tech','tch : frontpage'];
cnn_metadata.friendly_name='Tech Home Page';
cnn_metadata.template_type='section front';
var CNN_gallery_0_ad_0="/cnn_adspaces/3.0/technology/main/bot1.120x90.ad";
var CNN_gallery_0_ad_1="/cnn_adspaces/3.0/technology/main/bot2.120x90.ad";
var CNN_gallery_0_ad_2="/cnn_adspaces/3.0/technology/main/bot3.120x90.ad";
</script>
```

Teknik Analisis Teks Dasar

- Perlu menganalisis dan mengindeks halaman web yang dirayapi
 - Ekstrak isi informatif dari HTML
 - Membangun representasi data yang dapat diakses mesin

Penguraian HTML (*parsing*)

- Umumnya sulit karena gaya bebas dari HTML
- Solusi
 - Penguraian dangkal (*shallow parsing*)
 - Hapus semua tag HTML
 - Hanya simpan teks di antara <title></title> dan <p></p>
 - Pembuatan bungkus otomatis (*wrapper*) [Crescenzi et al. VLDB'01]
 - Wrapper: ekspresi reguler untuk kombinasi tag HTML
 - Penalaran induktif dari contoh
 - Penguraian visual (*visual parsing*) [Yang and Zhang DAR'01]
 - Penambahan pola yang sering dari blok HTML yang mirip secara visual

Penguraian HTML (*parsing*)

- jsoup

- Java-based HTML

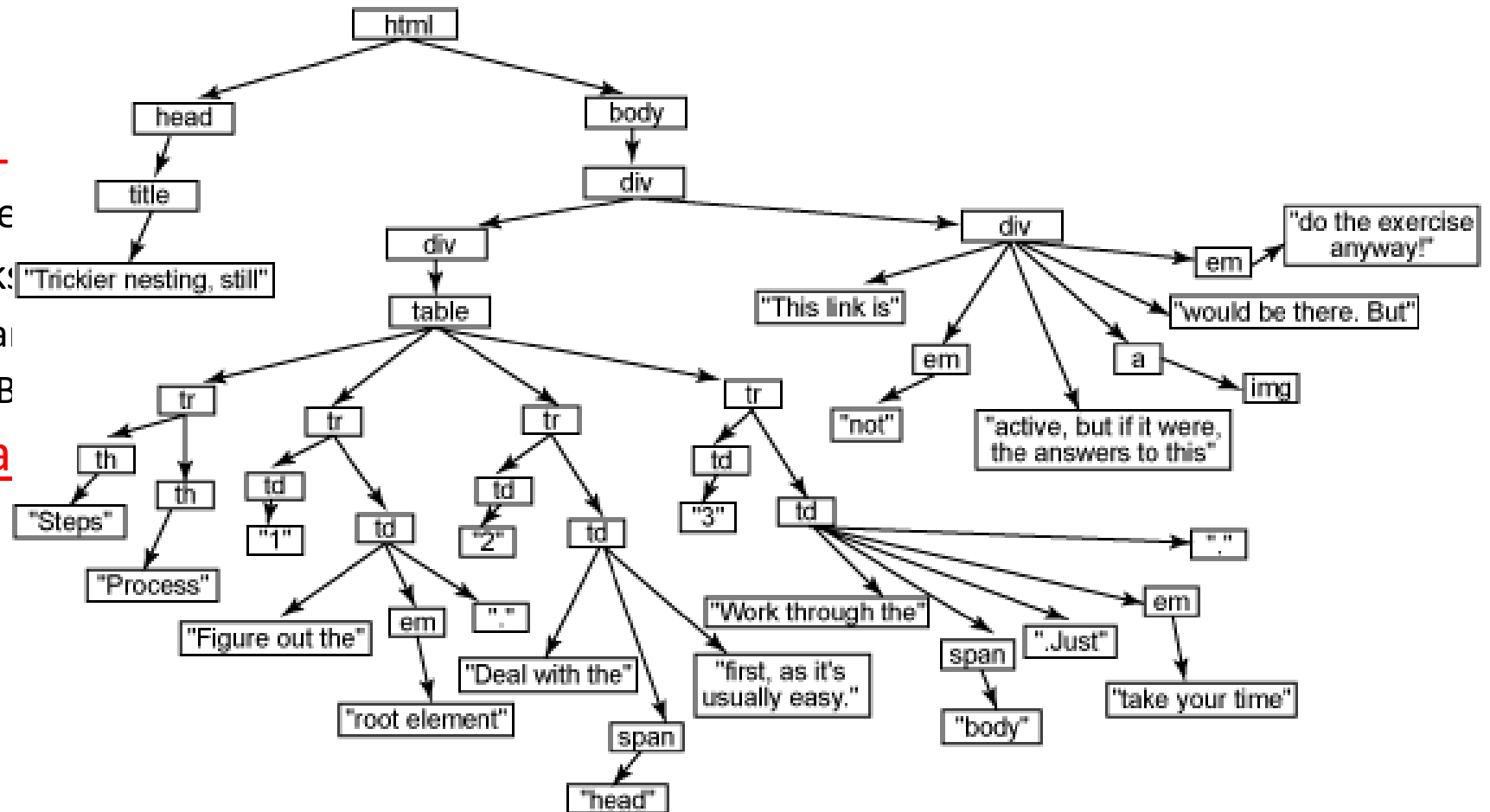
- scrape dan parse

- Temukan dan eks

- `children()`, `parent()`

- `getElementB`

- Versi Python: Bea



Rekap: Focused Crawling

- Ada prioritas urutan kunjungan terhadap halaman web
 - Ukuran web terlalu besar bagi suatu web crawler (bahkan bagi Google)
 - Tidak semua dokumen mempunyai kepentingan yang sama
 - Lebih ditekankan pada dokumen berkualitas tinggi
 - Memaksimalkan *weighted coverage*

$$\text{WC}(t) = \sum_{p \in \mathcal{C}(t)} w(p)$$

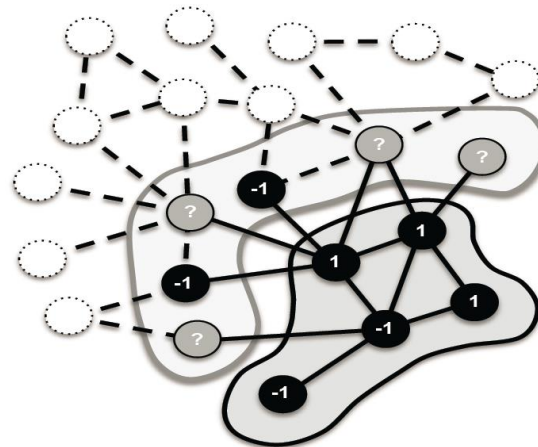
Weighted coverage sampai waktu t

Kepentingan halaman p

Halaman yang dicrawl sampai waktu t

Rekap: Focused Crawling

- Prioritas berdasarkan PageRank [Abiteboul et al. WWW'07, Cho and Uri VLDB'07]
 - Pada tahapan awal menggunakan *breadth-first*, kemudian menghitung PageRank secara periodik
 - Lebih konsisten dengan relevansi pencarian [Fetterly et al. SIGIR'09]



Bagaimana Merepresentasikan Dokumen

- ```
<HEAD>Crowds in Liverpool to Mark 10th Anniversary of John Lennon's
Death</HEAD>
<DATELINE>LIVERPOOL, England(AP) </DATELINE>
<TEXT>
 Dozens of fans of rock legend and former Beatle John Lennon gathered in the snow on a windy Saturday for a ceremony
 marking the 10th anniversary of his death. Liverpool's mayor, Dorothy Gavin, led Lennon devotees who laid wreaths at the
 foot of a bronze statue of The Beatles in the city's Cavern Walks shopping center. The center was built on the original site of
 the Cavern Club, made famous when The Beatles played there in the 1960s, and has become a place of pilgrimage. ``Give
 peace a chance," the title of one of singer-songwriter Lennon's greatest hits, was the theme for the day.
 ...
 Lennon and his wife, Yoko Ono, were returning to their apartment in New York's Dakota apartment building after a recording
 session on Dec. 8, 1980, when Lennon was shot to death by Mark David Chapman, a deranged fan to whom Lennon had
 given his autograph only hours before. Lennon was 40. A spokesman for the Lennon family said Ms. Ono and the couple's
 son, Sean, were in Europe and would spend the anniversary privately.
 ...
 Peebles said late in 1980 that Lennon had just recovered from a period when he had ``gone off the rails" and his relationship
 with Ms. Ono had suffered. ``But (when I saw him) they'd had the baby, Sean had been born, and everything was great."
</TEXT>
```

# Tokenisasi

- Pecahkan sederetan teks ke dalam unit-unit berarti
  - Token: kata, frase, simbol
    - **Input:** It's not straight-forward to perform so-called "tokenization."
    - **Output(1):** 'It's', 'not', 'straight-forward', 'to', 'perform', 'so-called', '"tokenization."'
    - **Output(2):** 'It', "'", 's', 'not', 'straight', '-', 'forward', 'to', 'perform', 'so', '-', 'called', '"', 'tokenization', '.', '"""
- Definisi tergantung pada bahasa, corpus, atau konteks.

# Tokenisasi

- Solusi
  - Ekspresi regular
    - `[\w]+`: so-called -> 'so', 'called'
    - `[\S]+`: It's -> 'It's' instead of 'It', 's'
  - Metode statistik
    - Jelajahi fitur yang kaya untuk memutuskan di mana batas kata
      - Apache OpenNLP (<http://opennlp.apache.org/>)
      - Stanford NLP Parser (<http://nlp.stanford.edu/software/lex-parser.shtml>)
    - Online Demo
      - Stanford (<http://nlp.stanford.edu:8080/parser/index.jsp>)
      - UIUC (<http://cogcomp.cs.illinois.edu/curator/demo/index.html>)

# Pengindeksan Teks Lengkap

## *full text indexing*

- Representasi Bag-of-Words
  - Doc1: Information retrieval is helpful for everyone.
  - Doc2: Helpful information is retrieved for you.

	information	retrieval	retrieved	is	helpful	for	you	everyone
Doc1	1	1	0	1	1	1	0	1
Doc2	1	0	1	1	1	1	1	0



Matriks kedekatan dokumen-kata

# Full Text Indexing

- Representasi *Bag-of-Words*
  - Asumsi: kata saling bebas satu sama lain
  - Pros: simpel
  - Cons: tata bahasa dan urutan kata hilang
  - Representasi dokumen yang paling sering digunakan untuk
    - Image, speech, gene sequence

# Full Text Indexing

- Representasi *Bag-of-Words* yang ditingkatkan
  - N-grams: urutan yang berdekatan dari n item dari urutan teks yang diberikan
    - Mis. Information retrieval is helpful for everyone
    - Bigrams: 'information\_retrieval', 'retrieval\_is', 'is\_helpful', 'helpful\_for', 'for\_everyone'
  - Pros: menangkap ketergantungan dan urutan lokal
  - Cons: murni tampilan statistik, menambah ukuran kosa kata  $O(V^N)$

# Full Text Indexing

- Indeks dokumen dengan semua kata yang muncul
  - Pro:
    - Menyimpan semua informasi dalam teks (semoga)
    - Sepenuhnya otomatis
  - Kontra:
    - Kesenjangan kosakata: cars v.s., car
    - Penyimpanan besar: mis. Dalam N-grams  $O(V^N)$
  - Solusi
    - Bangun kosakata terkontrol

# Properti Statistik dari Bahasa

- Hukum Zipf

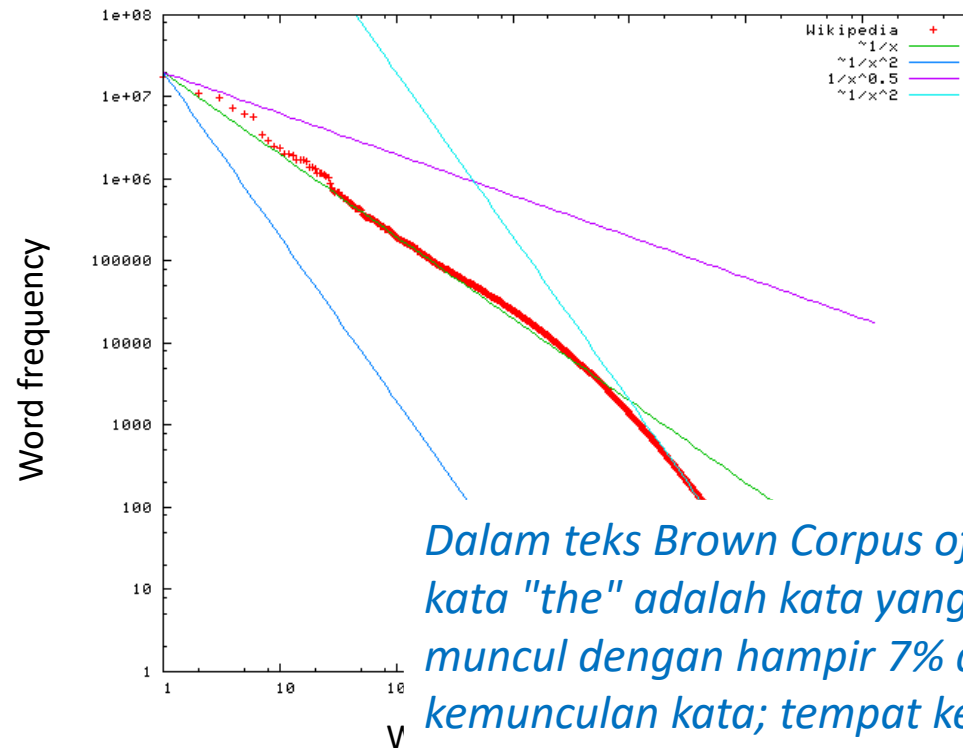
- Frekuensi kata apa pun berbanding terbalik dengan peringkatnya di tabel frekuensi

- Secara formal:

- $$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^N 1/n^s}$$

di mana  $k$  adalah peringkat kata;  $N$  adalah ukuran kosakata;  $s$  adalah parameter khusus bahasa

*discrete version of power law*



Plot frekuensi kata

*Dalam teks Brown Corpus of American English, kata "the" adalah kata yang paling sering muncul dengan hampir 7% dari semua kemunculan kata; tempat kedua adalah kata "dari" menyumbang sedikit di atas 3,5% kata.*

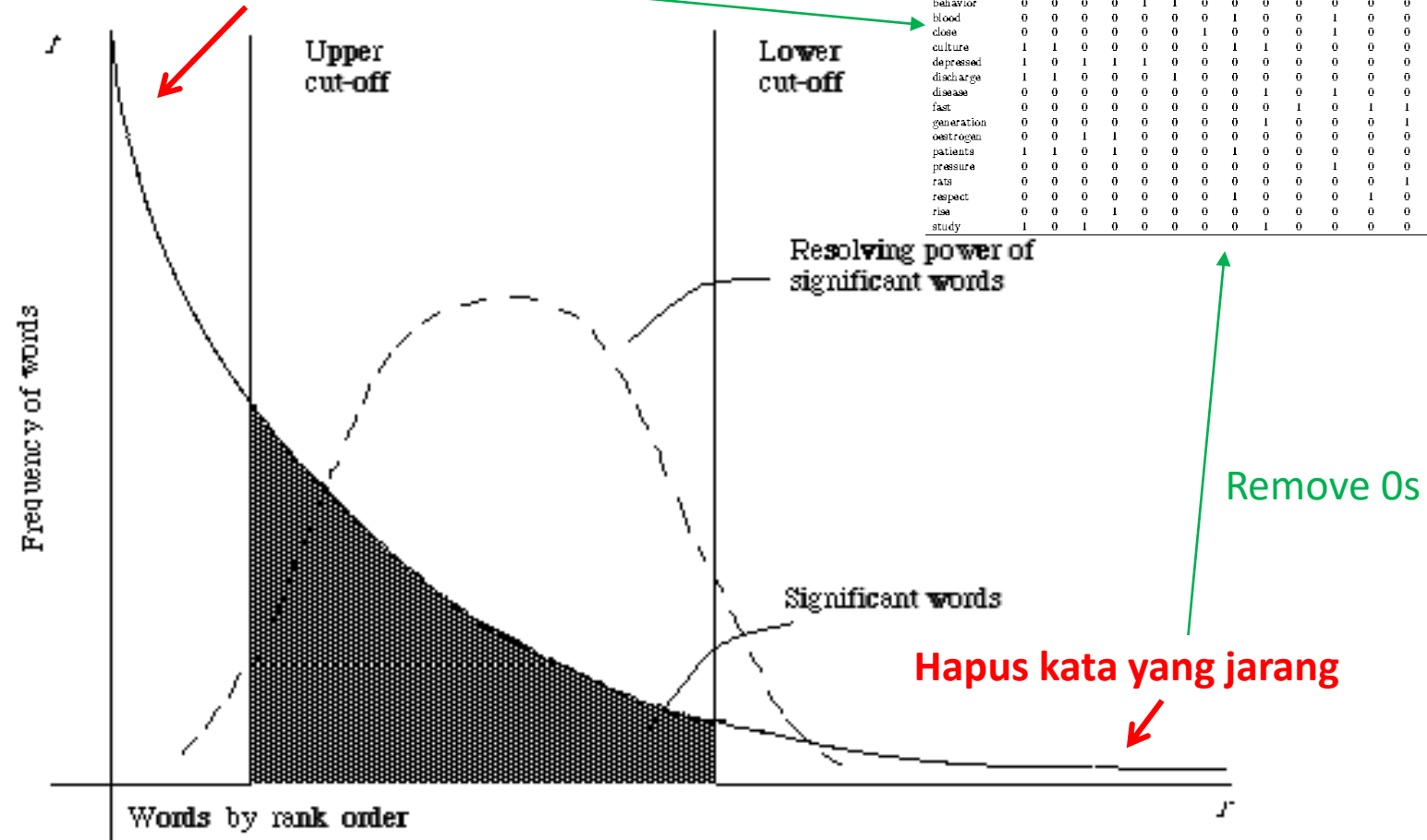
# Hukum Zipf Memberitahukan...

- Kata-kata utama mungkin lebih sering muncul, tetapi secara semantik tidak berarti
  - Mis. the, a, an, we, do, to
- Kata-kata ekor mengambil bagian utama dari kosa kata, tetapi jarang hadir di dalam dokumen
  - Mis. dextrosinistral
- Sisanya paling representatif
  - Untuk dimasukkan dalam kosakata terkontrol

# Indexing Teks Otomatis

Hapus kata-kata *non-informative*

Remove 1s



Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Figure 2.1. A plot of the hyperbolic curve relating  $f$ , the frequency of occurrence and  $r$ , the rank order (Adapted from Schultz<sup>44</sup> page 120)

# Stopwords

- Kata-kata tidak
  - Tidak semua kata
  - Hapus kata-kata
  - Tidak ada definisi
  - Resiko: merusa
    - Mis., *this is not*  
*to be or not to be*

## Nouns

1. time
2. person
3. year
4. way
5. day
6. thing
7. man
8. world
9. life
10. hand
11. part
12. child
13. eye
14. woman
15. place
16. work
17. week
18. case
19. point
20. government
21. company
22. number
23. group
24. problem
25. fact

## Verbs

1. be
2. have
3. do
4. say
5. get
6. make
7. go
8. know
9. take
10. see
11. come
12. think
13. look
14. want
15. give
16. use
17. find
18. tell
19. ask
20. work
21. seem
22. feel
23. try
24. leave
25. call

## Adjectives

1. good
2. new
3. first
4. last
5. long
6. great
7. little
8. own
9. other
10. old
11. right
12. big
13. high
14. different
15. small
16. large
17. next
18. early
19. young
20. important
21. few
22. public
23. bad
24. same
25. able

## Prepositions

1. to
2. of
3. in
4. for
5. on
6. with
7. at
8. by
9. from
10. up
11. about
12. into
13. over
14. after
15. beneath
16. under
17. above

## Others

1. the
2. and
3. a
4. that
5. I
6. it
7. not
8. he
9. as
10. you
11. this
12. but
13. his
14. they
15. her
16. she
17. or
18. an
19. will
20. my
21. one
22. all
23. would
24. there
25. their

**OEC: Fakta tentang bahasa**

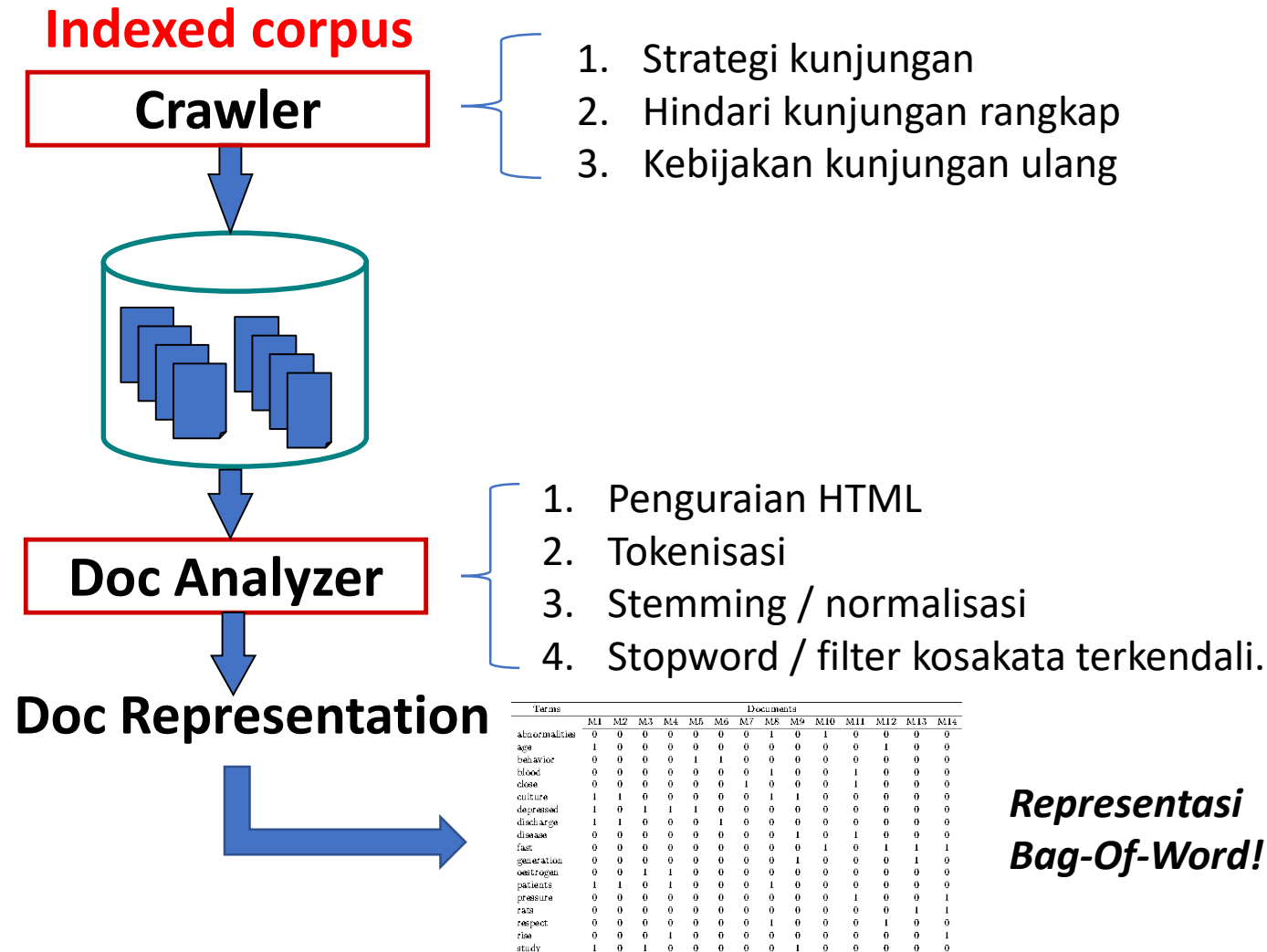
# Normalisasi

- Ubah berbagai bentuk kata menjadi bentuk normal dalam kosakata
  - U.S.A -> USA, St. Louis -> Saint Louis
- Solusi
  - *Rule-based*
    - Hapus titik dan strip ("-")
    - Semua dalam huruf kecil (lower case)
  - *Dictionary-based*
    - Bangun kelas ekuivaken
      - Car -> "automobile, vehicle"
      - Mobile phone -> "cellphone"

# Stemming

- Mengurangi kata-kata berimbuhan atau turunan ke bentuk asalnya
  - Bentuk jamak, kata keterangan, dan bentuk kata yang berubah
    - Mis. ladies -> ladi, referring -> refer, forgotten -> forget
  - Jembatani kesenjangan kosa kata
  - Risiko: kehilangan makna kata yang tepat
    - Mis. lay -> lie (penyataan yang salah? Atau berada dalam posisi horizontal?)
  - Solusi (untuk English)
    - Porter stemmer: pola urutan vokal-konsonan
    - Krovetz Stemmer: aturan morfologis

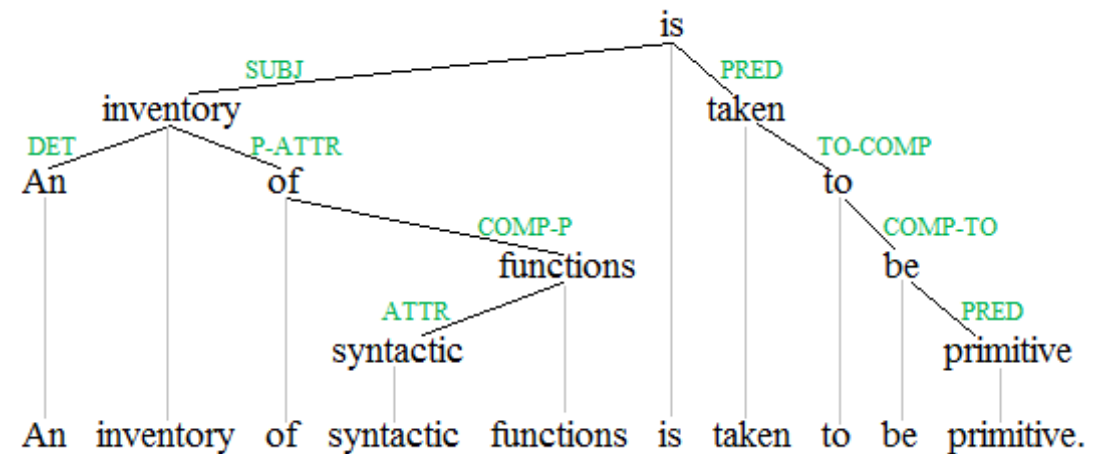
# Abstraksi dari Arsitektur Search Engine



# Indexing Teks Otomatis

*Query: “to be or not to be”*

- Pada search engine modern
  - Tidak ada *stemming* atau *stopword removal*, karena komputasi dan *storage* sudah tidak lagi menjadi perhatian utama
  - Teknik NLP yang lebih maju diterapkan
    - *Named entity recognition*
      - Mis., orang-orang, lokasi dan organisasi
    - *Dependency parsing*



# Apa yang Sebaiknya Diketahui

- Teknik dasar untuk *crawling*
- Hukum Zipf
- Prosedur untuk *indexing* teks otomatis
- Representasi dokumen *Bag-of-Words*

# Bacaan hari ini

- Introduction to Information Retrieval
  - Bab 20: Web crawling and indexes
    - Bagian 20.1, Overview
    - Bagian 20.2, Crawling
  - Bab 2: The term vocabulary and postings lists
    - Bagian 2.2, Determining the vocabulary of terms
  - Bab 5: Index compression
    - Bagian 5.1, Statistical properties of terms in information retrieval

# Referensi I

- Cho, Junghoo, Hector Garcia-Molina, and Lawrence Page. "[Efficient crawling through URL ordering.](#)" *Computer Networks and ISDN Systems* 30.1 (1998): 161-172.
- Abiteboul, Serge, Mihai Preda, and Gregory Cobena. "[Adaptive on-line page importance computation.](#)" *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.
- Cho, Junghoo, and Uri Schonfeld. "[RankMass crawler: a crawler with high personalized pagerank coverage guarantee.](#)" *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007.
- Fetterly, Dennis, Nick Craswell, and Vishwa Vinay. "[The impact of crawl policy on web search effectiveness.](#)" *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009.
- De Bra, Paul ME, and R. D. J. Post. "[Information retrieval in the World-Wide Web: making client-based searching feasible.](#)" *Computer Networks and ISDN Systems* 27.2 (1994): 183-192.

# Referensi II

- Chakrabarti, Soumen, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. "[Automatic resource compilation by analyzing hyperlink structure and associated text.](#)" *Computer Networks and ISDN Systems* 30, no. 1 (1998): 65-74.
- Crescenzi, Valter, Giansalvatore Mecca, and Paolo Merialdo. "[Roadrunner: Towards automatic data extraction from large web sites.](#)" *VLDB*. Vol. 1. 2001.
- Hersovici, Michael, et al. "[The shark-search algorithm. An application: tailored Web site mapping.](#)" *Computer Networks and ISDN Systems* 30.1 (1998): 317-326.
- Yang, Yudong, and HongJiang Zhang. "[HTML page analysis based on visual cues.](#)" *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*. IEEE, 2001.

# Pertanyaan?

Terimakasih!