Data Collection & Data Preprocessing

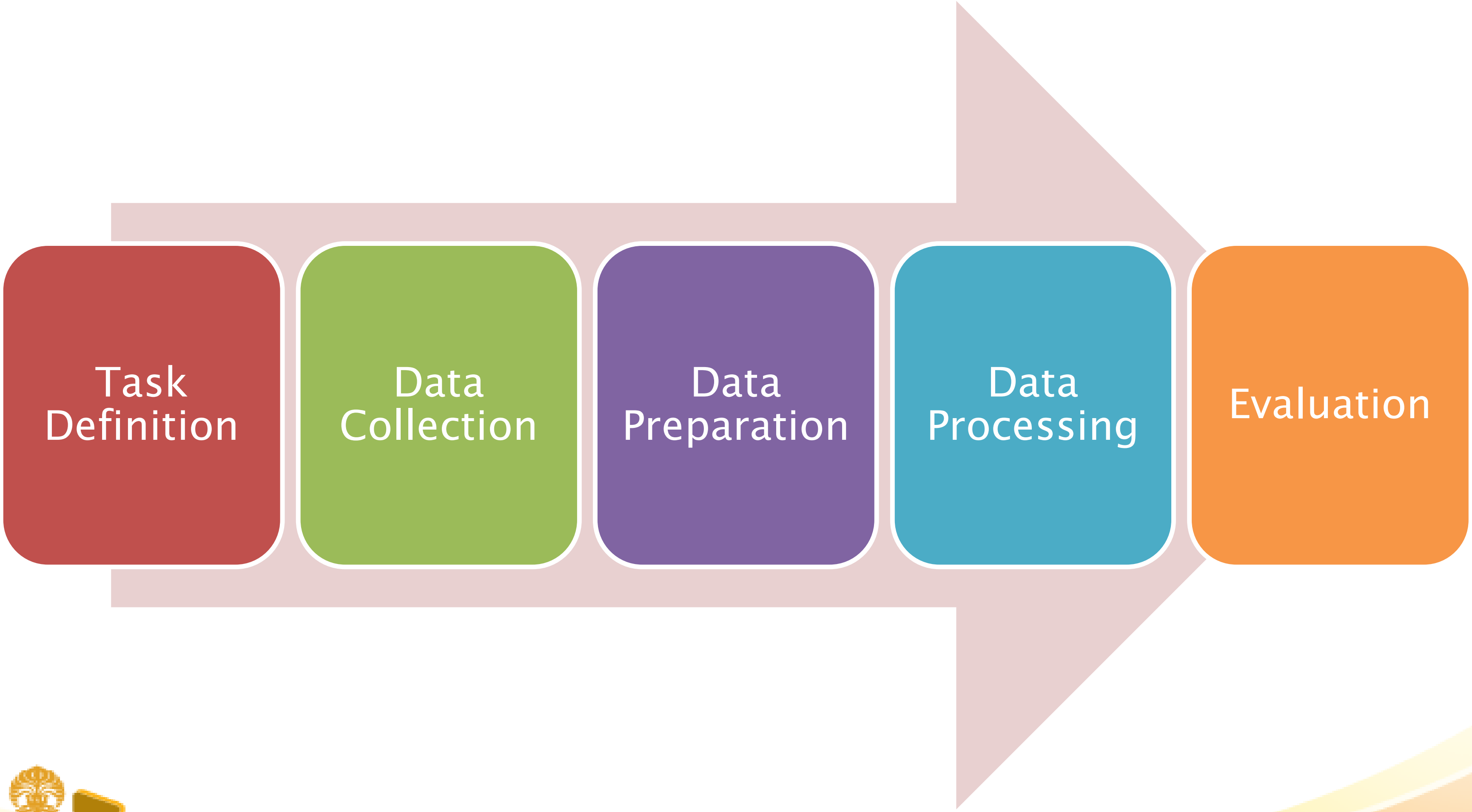
Bayu Distiawan

Natural Language Processing & Text Mining
Short Course

Pusat Ilmu Komputer UI
22 – 26 Agustus 2016

DATA COLLECTION

Text Mining Process



Task Definition

Data Collection

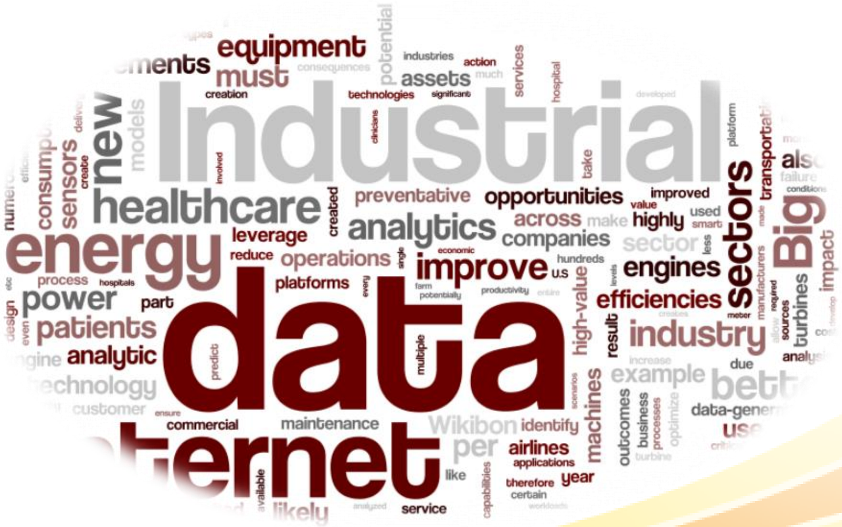
Data Preparation

Data Processing

Evaluation

Data Collection

- Collect from internal source.
- Collaboration with partner
- Pay for the data
- Collect public data



Collecting public data

- Available corpus
  - http://qwone.com/~jason/20Newsgroups/
  - http://www.daviddlewis.com/resources/testcollections/reuters21578/
  - https://dumps.wikimedia.org/
  - http://schwa.org/projects/resources/wiki/Wikiner
- Available Data
  - The Internet!

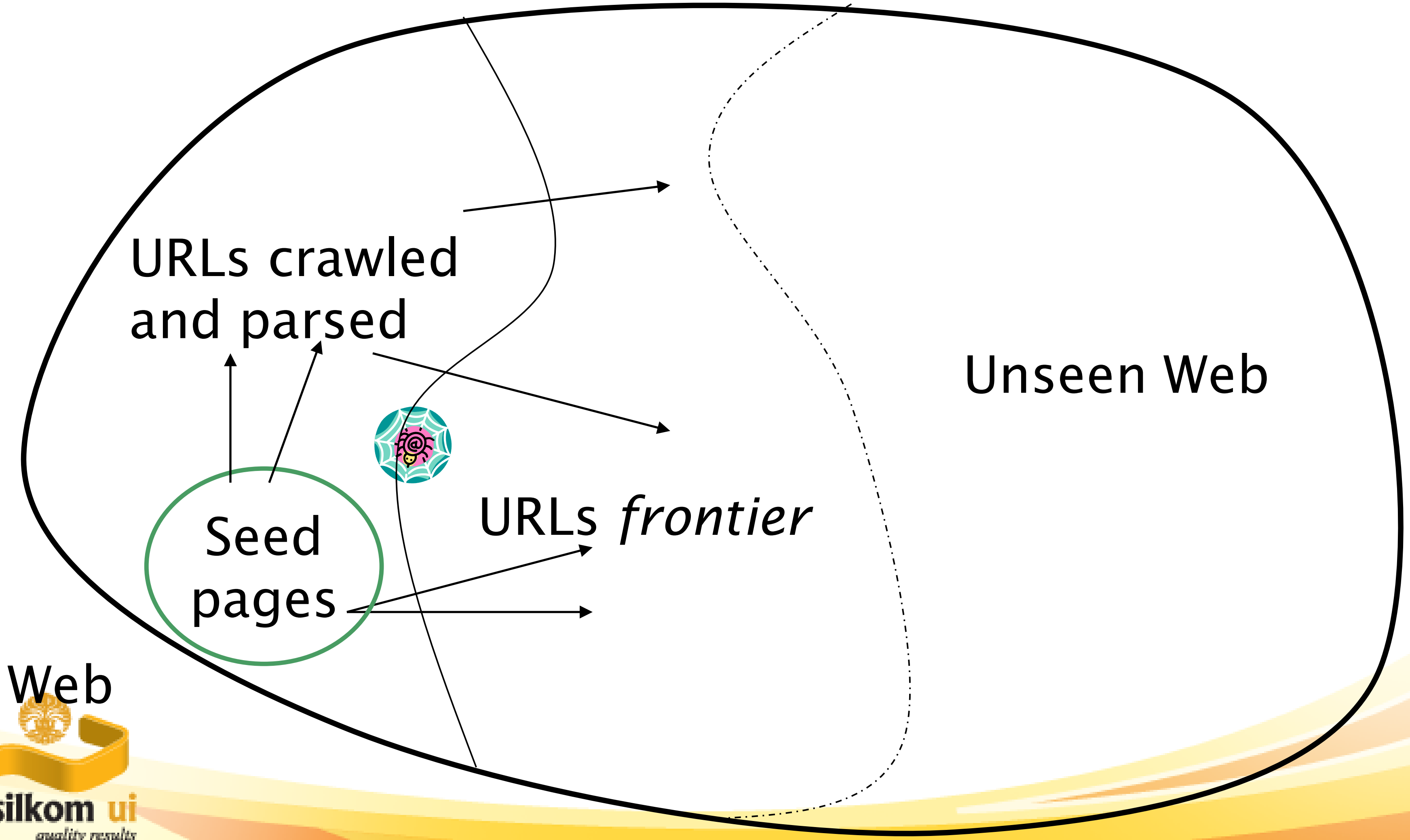
Collecting Data From Internet

- Crawler
- Spider
- Robot (or bot)
- Web agent
- Wanderer, worm, …
- And famous instances: googlebot, scooter, slurp, msnbot, …

Basic crawler operation

- Begin with known “seed” URLs
- Fetch and parse them
  - Extract URLs they point to
  - Place the extracted URLs on a queue
- Fetch each URL on the queue and repeat

Crawling picture



URLs crawled and parsed

Unseen Web

Seed pages

URLs *frontier*

Web

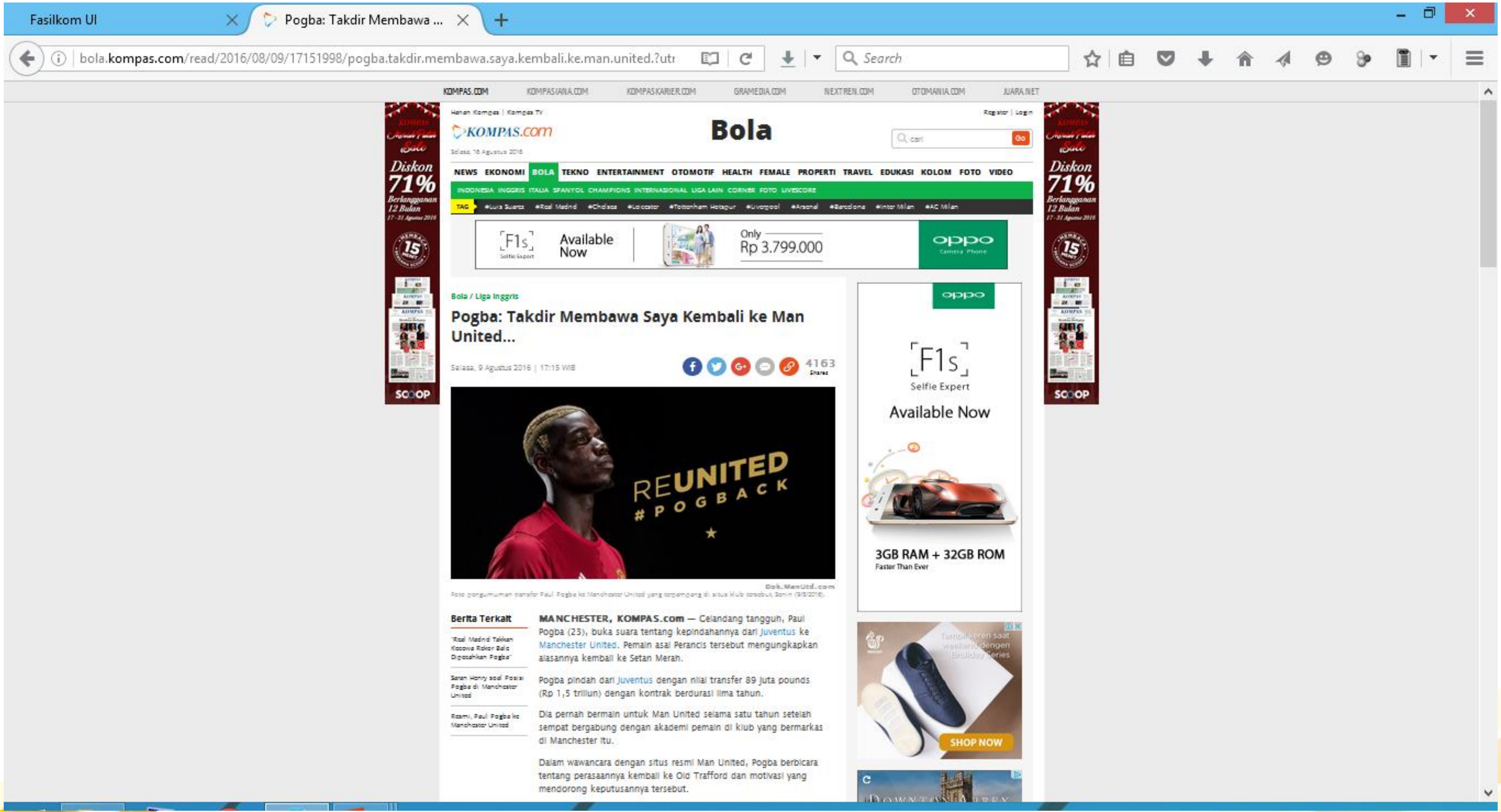
Simple picture – complications

- Web crawling isn't feasible with one machine
  - All of the above steps distributed
- Malicious pages
  - Spam pages
  - Spider traps – incl dynamically generated
- Even non-malicious pages pose challenges
  - Latency/bandwidth to remote servers vary
  - Webmasters' stipulations
    - How "deep" should you crawl a site's URL hierarchy?
  - Site mirrors and duplicate pages
- Politeness – don't hit a server too often

What any crawler *must* do

- Be Polite: Respect implicit and explicit politeness considerations
  - Only crawl allowed pages
  - Respect *robots.txt* (more on this shortly)
- Be Robust: Be immune to spider traps and other malicious behavior from web servers

After Crawling???


Fasilkom UI
Pogba: Takdir Membawa ...
bola.kompas.com/read/2016/08/09/17151998/pogba.takdir.membawa.saya.kembali.ke.man.united.?utr
Search
KOMPAS.COM
KOMPASIANA.COM
KOMPASKARIER.COM
GRAMEDIA.COM
NEXTREN.COM
OTOMANIA.COM
JUARA.NET
Harian Kompas | Kompas TV
Register | Login
KOMPAS.com
Bola
cari
Go
Selasa, 16 Agustus 2016
NEWS EKONOMI BOLA TEKNO ENTERTAINMENT OTOMOTIF HEALTH FEMALE PROPERTI TRAVEL EDUKASI KOLOM FOTO VIDEO
INDONESIA INGGRIS ITALIA SPANYOL CHAMPIONS INTERNASIONAL LIGA LAIN CORNER FOTO LIVESCORE
TAG #Luis Suarez #Real Madrid #Chelsea #Leicester #Tottenham Hotspur #Liverpool #Arsenal #Barcelona #Inter Milan #AC Milan
F1s Selfie Expert Available Now Only Rp 3.799.000 oppo Camera Phone
3GB RAM + 32GB ROM
Faster Than Ever
SHOP NOW
REUNITED #POGBACK
4163 Shares

Bola / Liga Inggris

Pogba: Takdir Membawa Saya Kembali ke Man United...

Selasa, 9 Agustus 2016 | 17:15 WIB

Dok.ManUtd.com

Foto pengumuman transfer Paul Pogba ke Manchester United yang terpampang di situs klub tersebut, Senin (8/8/2016).

Berita Terkait

- "Real Madrid Takkan Kecewa Rekor Bale Dipecahkan Pogba"
- Saran Henry soal Posisi Pogba di Manchester United
- Resmi, Paul Pogba ke Manchester United

MANCHESTER, KOMPAS.com — Gelandang tangguh, Paul Pogba (23), buka suara tentang kepindahannya dari Juventus ke Manchester United. Pemain asal Perancis tersebut mengungkapkan alasannya kembali ke Setan Merah.

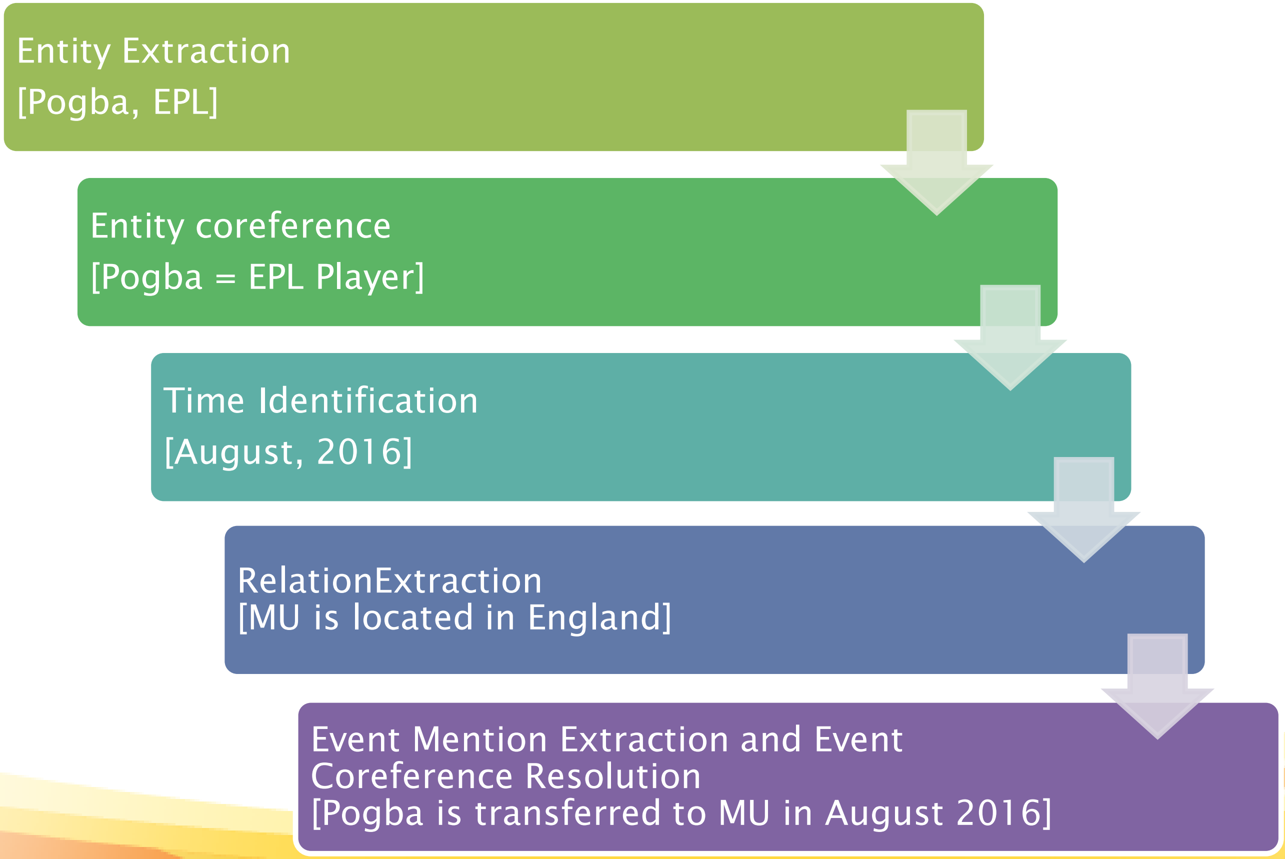
Pogba pindah dari Juventus dengan nilai transfer 89 juta pounds (Rp 1,5 triliun) dengan kontrak berdurasi lima tahun.

Dia pernah bermain untuk Man United selama satu tahun setelah sempat bergabung dengan akademi pemain di klub yang bermarkas di Manchester itu.

Dalam wawancara dengan situs resmi Man United, Pogba berbicara tentang perasaannya kembali ke Old Trafford dan motivasi yang mendorong keputusannya tersebut.

Extracting Information

- Gather information from unstructured data
  - Creating “relational” like data



Entity Extraction
[Pogba, EPL]

Entity coreference
[Pogba = EPL Player]

Time Identification
[August, 2016]

RelationExtraction
[MU is located in England]

Event Mention Extraction and Event Coreference Resolution
[Pogba is transferred to MU in August 2016]

DATA PREPARATION/PREPROCESSING

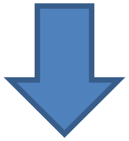
Data Preprocessing

- Depends on the task
- Some preprocessing:
  - Sentence Splitting
  - Filtering
  - Stemming
  - Normalization
  - POS Tagging
  - NP Chunking
  - Parsing
  - Etc.

Sentence Splitting

- Split paragraph/article into sentences

Manchester United have agreed a world record deal to sign Paul Pogba for €110 million, **Goal** understands. Officials from the Premier League club met with their Juventus counterparts earlier on Wednesday to discuss a deal to bring Pogba back to Old Trafford. It is now understood that United have settled on a fee of €110m for Pogba, which eclipses the previous record set when Real Madrid paid €100m for Gareth Bale in 2013.



Manchester United have agreed a world record deal to sign Paul Pogba for €110 million, **Goal** understands.

Officials from the Premier League club met with their Juventus counterparts earlier on Wednesday to discuss a deal to bring Pogba back to Old Trafford.

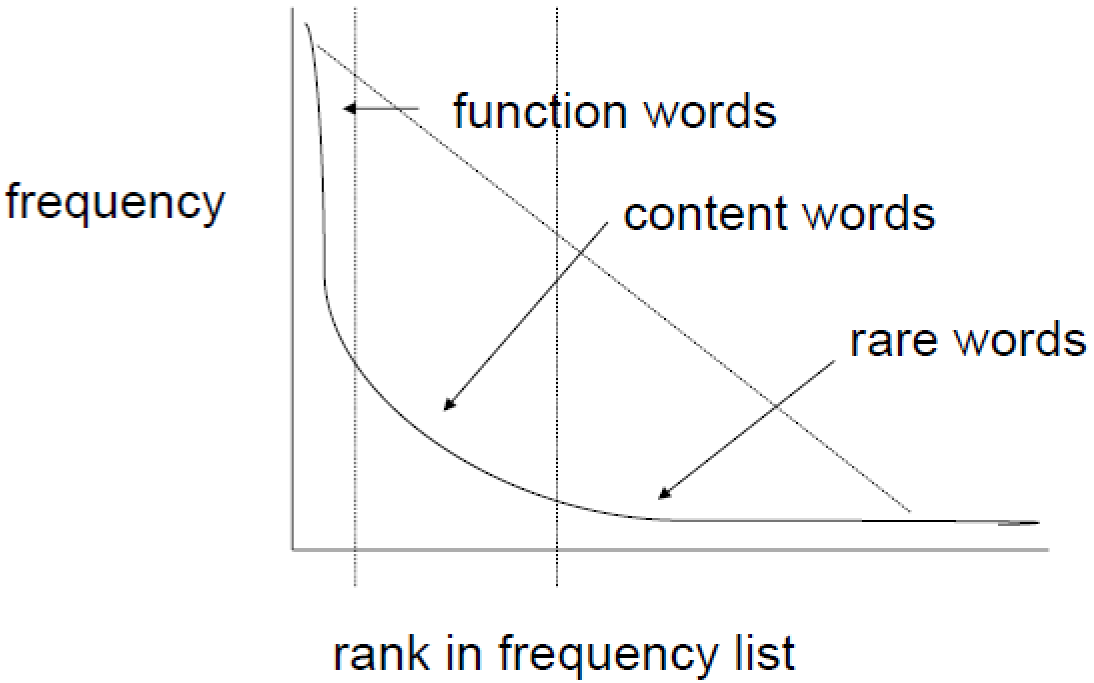
It is now understood that United have settled on a fee of €110m for Pogba, which eclipses the previous record set when Real Madrid paid €100m for Gareth Bale in 2013.

Filtering/Stop Word Removal

- Many of the most frequently used words in English are useless in IR and text mining – these words are called *stop words*.
  - the, of, and, to, ….
  - Typically about 400 to 500 such words
  - For an application, an additional domain specific stopwords list may be constructed
- Why do we need to remove stopwords?
  - Reduce indexing (or data) file size
    - stopwords accounts 20-30% of total word counts.
  - Improve efficiency and effectiveness
    - stopwords are not useful for searching or text mining
    - they may also confuse the retrieval system.

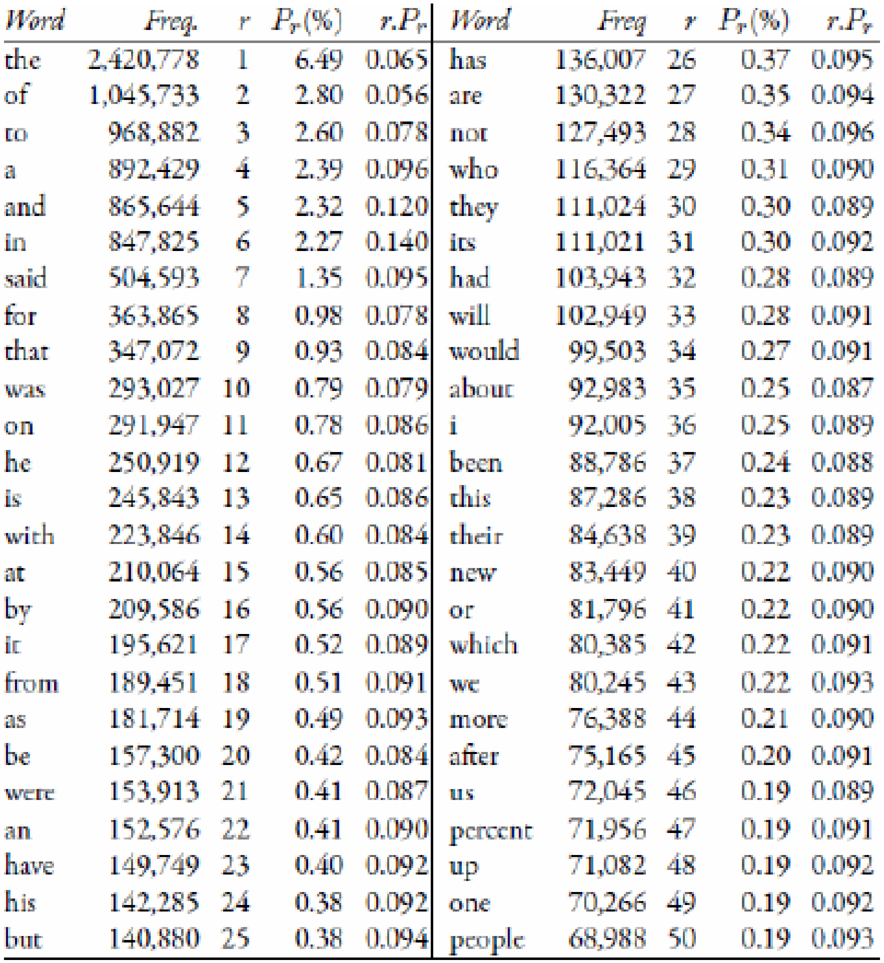
Filtering/Stop Word Removal

Word distribution frequency



Filtering/Stop Word Removal

Top 50 words from AP89 corpus



Word	Freq.	r	$P_r(\%)$	$r.P_r$	Word	Freq	r	$P_r(\%)$	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

Stemming

- Techniques used to find out the root/stem of a word. E.g.,
  - user
  - users
  - used
  - using

  engineering
  engineered
  engineer
- stem: use engineer

Usefulness:

- improving effectiveness of IR and text mining
  - matching similar words
  - Mainly improve recall
- reducing indexing size
  - combing words with same roots may reduce indexing size as much as 40-50%.

Basic stemming methods

Using a set of rules. E.g.,

- remove ending
  - if a word ends with a consonant other than s,

    followed by an s, then delete s.
  - if a word ends in es, drop the s.
  - if a word ends in ing, delete the ing unless the remaining word consists only of one letter or of th.
  - If a word ends with ed, preceded by a consonant, delete the ed unless this leaves only a single letter.
  - ......
- transform words
  - if a word ends with “ies” but not “eies” or “aies” then “ies --> y.”

Normalization

Token normalization is *the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens* (Stanford IR Book).

U.S.A, USA → **usa**

windows,Windows, window, Window → **windows**

Lexical Normalization in Social Media

User creativity on social media creates a problem for NLP Processing.

I love u -> **i love you**

tmrw -> **tomorrow**

4eva -> **forever**

Lexical Normalization in Social Media

Technique: Using dictionary

Other technique **(Han & Baldwin, 2011)**:

Machine learning, features:

- Edit distance value
- Prefix substring
- Suffix substring
- Longest common subsequence (LCS)

Linguistic Pre-processing

- Advanced preprocessing task

- POS Tagging
  - Budi/NN eats/VB bakso/NN
- NP Chunking
  - **[NP The most expensive footballer]** wearing **[NP a red shirt]**
- Named Entity Recognition
  - **<PERS>**President Joko Widodo**</PERS>** meets Ahok in **<LOC>**Istana Negara**</LOC>**
- Parsing

Thank you

pusilkom ui

quality results