

IF5181 Pengenalan Pola

Mining Text

Masayu Leylia Khodra

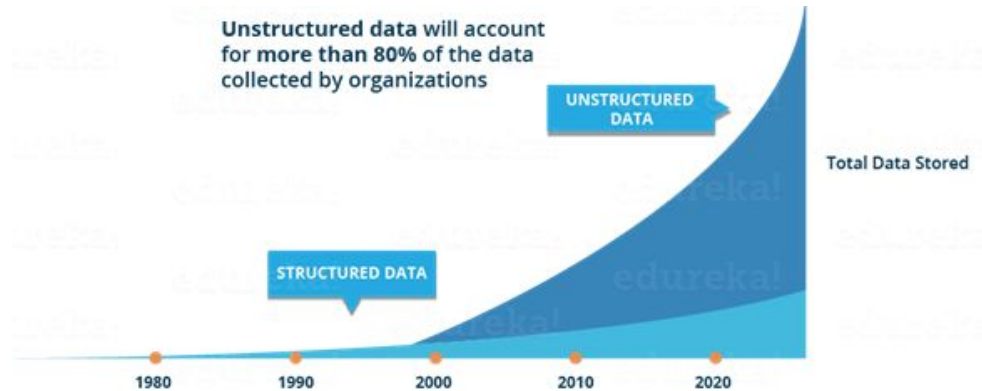
Magister Informatika ITB

Referensi

- Ayu Purwarianti. Natural Language Processing & NLTK. DTS ITB. 2019.
- Yudi Wibisono. Text Mining
- Masayu Leylia Khodra. Text Mining: Peringkasan Teks. 2016

Teks

- Today, data is mostly unstructured or semi-structured.



Source: <https://www.edureka.co/blog/wp-content/uploads/2017/01/Flow-of-unstructured-data.png>

2019 This Is What Happens In An Internet Minute



Tantangan dalam Text Mining

- Data berbentuk “free text”
 - Tidak atau kurang memiliki struktur
- Ambiguitas
 - “Dia melihat orang dengan teropong”
 - “Film-nya sangat bagus sampai membuat saya tertidur”
- Data memerlukan anotasi manual
 - Contoh: memberi label polaritas pada twitter

Text Mining

Proses menemukan pola yang berharga, dan tersembunyi dari korpus (kumpulan teks)

- Kategorisasi teks: melabeli teks dengan set kategori terdefinisi
- *Clustering* teks: mengelompokkan teks
- Ekstraksi informasi: menstrukturkan dokumen yang tidak terstruktur
- *Opinion mining*: menentukan sentimen [dan menstrukturkan opini by aspek]
- Question answering: memberikan jawaban atas pertanyaan pengguna
- Textual entailment: menentukan apakah suatu teks H diturunkan dari teks T
- Peringkasan teks: mentransformasi teks menjadi lebih pendek dan berisi konten penting

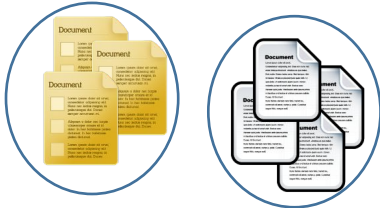
Kategorisasi/Klasifikasi Teks

Memberikan label/kelas/kategori kepada dokumen secara otomatis



spam filtering

Label: spam vs not spam



Document clustering

Label: depends on document keywords



Plagiarism detection

Label: plagiat vs not plagiat



Sentiment analysis

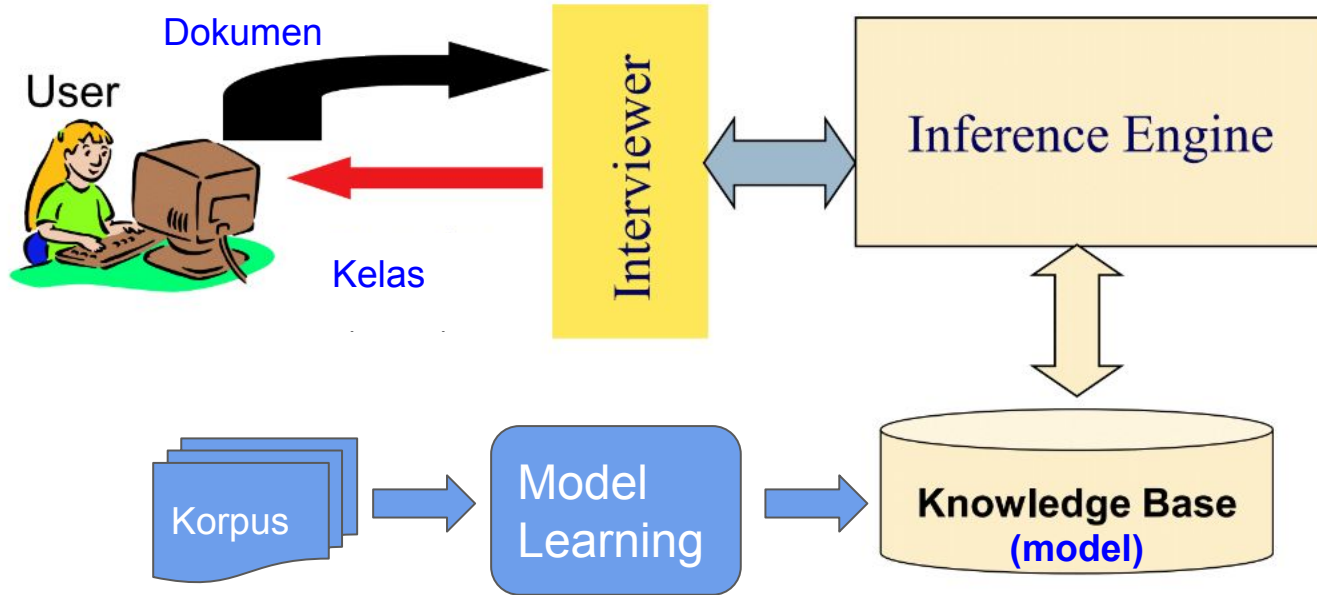
Label: positive, negative, neutral



Essay scoring

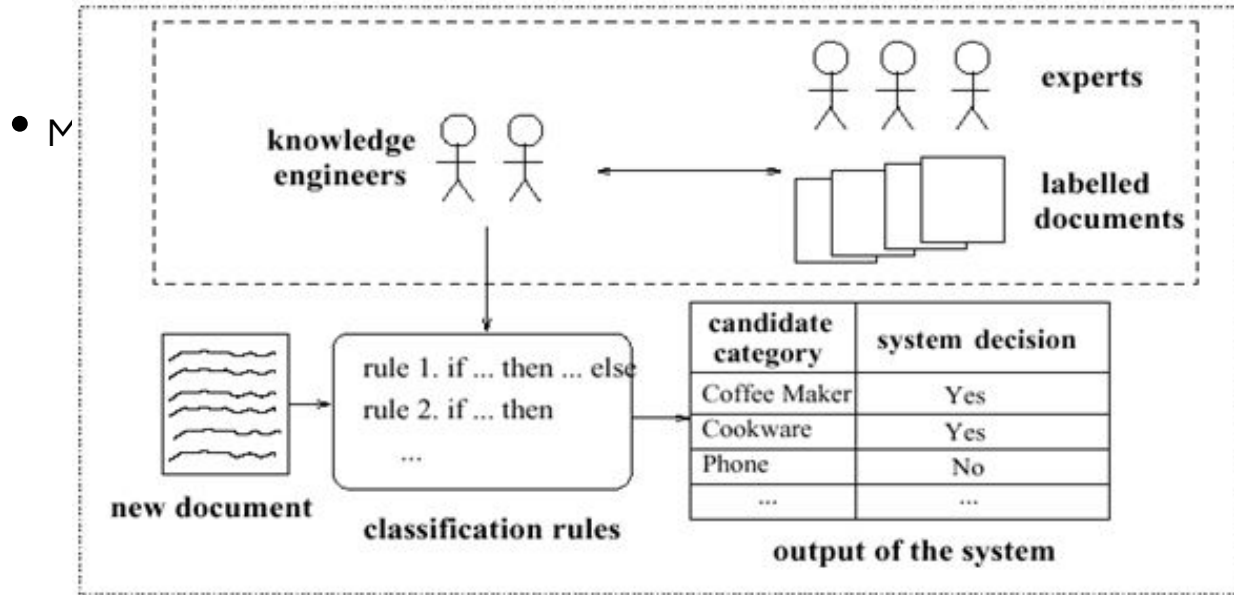
Label: score

Kategorisasi/Klasifikasi Teks



- Klasifikasi teks: dokumen \rightarrow class $\in C$

Pendekatan Rekayasa Pengetahuan

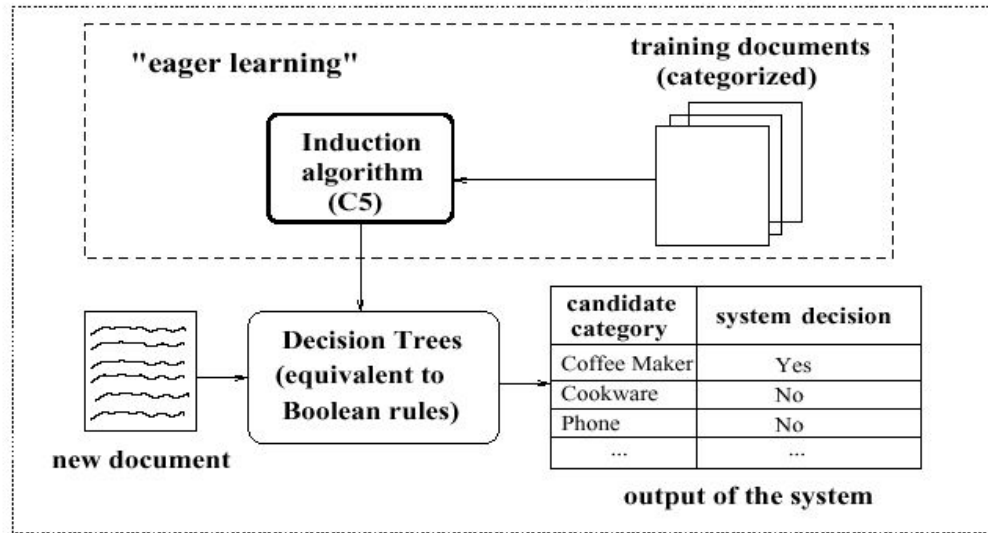


Development time: 192 person-months (2 people, 8 years)

Sumber: Bennett, 2002

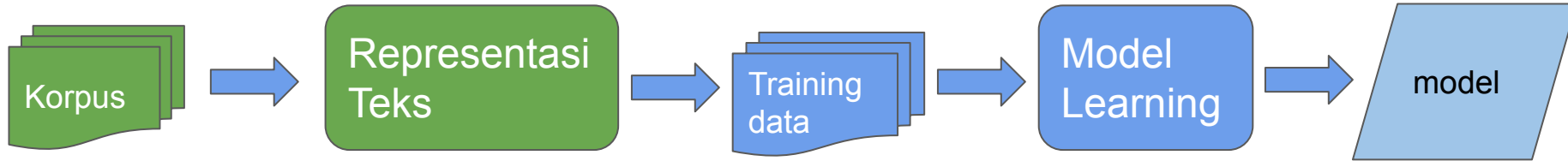
Pendekatan Pembelajaran Mesin

DTree induction for text categorization (since 1994)



Development time: 4 person-months

Konstruksi Model Learning: Contoh



Decision tree
learning

Naive Bayes
learning

atr1	...	atr _n	kelas

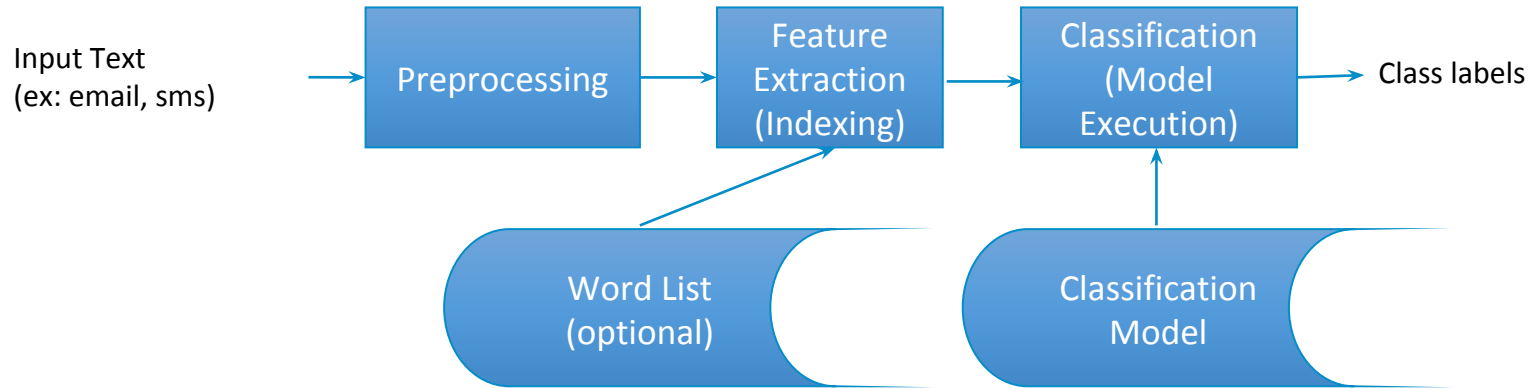
Contoh Representasi Teks: Bag of Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Flow of Text Classification Task



PreProcessing

Input Text
(ex: email, sms)



Preprocessing



List of token

- Tokenization: Text is tokenized into tokens such as words, bi-grams, tri-grams, n-grams
- Lemmatization: Word is lemmatized into its lemma form
- Stemming: Word is stemmed into its stemmed form
- Lowercase: all words are lowercased
- Stopword elimination: words are filtered by a stop word list
- POS Tagger and stopword elimination: conduct POS Tagger and filter words based on its POS Tag
- Spelling correction: incorrect words (including informal) are corrected
- Entity masking: words fulfilling certain patterns are masked

Feature Extraction

- Change **word** feature into **number** □ called **indexing**
- Example on unigram word feature for spam filtering:
 - INPUT: “Ini mama ... tolong kirim pulsa ke nomor hp ini” (English: This is mom.. Please send balance to this phone number”
 - Word List: mama (mom), tolong (please), kirim (send), pulsa (balance), nomor (number), hp (phone)
 - Training data
 - Feature: unigram words

Mama (mom)	Tolong (please)	Kirim (send)	Pulsa (balance)	Nomor (number)	Hp (phone)	...	Class
1	1	1	1	1	1	...	Spam
1	0	0	0	0	0	...	Not spam

Representasi Teks: Bag of Words

Iman mengatakan dengan aturan capping bunga deposito ini akan mempercepat transmisi penurunan bunga DPK yang akhirnya bisa menurunkan suku bunga kredit.

Destry mengungkapkan, pihaknya menyambut baik adanya BI 7-days Reverse Repo Rate.

Sektor perbankan mengalami tekanan aksi jual setelah Bank Indonesia mengumumkan perubahan instrumen suku bunga acuan yang diganti menjadi 7-days reverse repo rate.

Fitur	Dok1	Dok2	Dok3
7-Days	1	0	1
adanya	1	0	0
aksi	0	0	1
aturan	0	1	0
BI	1	0	0
baik	0	1	0
bunga	0	1	1
capping	0	1	0
deposito	0	1	0
Destry	1	0	0
DPK	0	1	0
Iman	0	1	0
kredit	0	1	0
mempercepat	0	1	0
mengalami	0	0	1
mengatakan	0	1	0
mengungkapkan	1	0	0
menurunkan	0	1	0

stoplist
dengan
ini
akan
yang
akhirnya
bisa
saat
lanjut
dan
dalam

Naive Bayes Text Classifier

Fitur	Dok1	Dok2	Dok3
7-Days	1	0	1
adanya	1	0	0
aksi	0	0	1
aturan	0	1	0
BI	1	0	0
baik	0	1	0

bunga	7-Days		adanya		aksi		aturan		BI		baik	
	+	-	+	-	+	-	+	-	+	-	+	-
0	0.5	0	0.5	1	1	0	0.5	1	0.5	1	0.5	1
1	0.5	1	0.5	0	0	1	0.5	0	0.5	0	0.5	0

$P(7\text{-days}=1 \mid +)$

$P(\text{aturan}=0 \mid -)$

Kategorisasi dengan Naive Bayes

Dokumen baru:

Dia mengatakan instrumen
7-days Repo Rate bersifat
jangka pendek



$$label = \arg \max_{label_j \in \{+, -\}} P(label_j) \prod_i P(attribute_i | label_j)$$

$P(+).P(dia=1 | +).P(mengatakan | +) \dots P(7\text{-days} | +) \dots P(pendek | +) = A$

$P(-).P(dia=1 | -).P(mengatakan | -) \dots P(7\text{-days} | -) \dots P(pendek | -) = B$

IF $A > B$ THEN label = + ELSE label = -

Representasi dengan N-Gram

'Text mining is to identify useful information.'

Bigrams (2-Grams): *'text_mining', 'mining_is', 'is_to', 'to_identify', 'identify_useful', 'useful_information', 'information_.'*

3-Grams: *'text_mining_is', 'mining_is_to', 'is_to_identify'*
dst.

Pembobotan TF-IDF

- Ide: kata yang penting adalah yang banyak muncul di suatu dokumen, tapi jarang di dokumen lain.
- Inverse jumlah dokumen yang mengandung kata tersebut (IDF). Semakin jarang (semakin eksklusif) semakin tinggi nilainya.
- Kombinasi: $TF\text{-}IDF = TF * IDF$

– Formula

$$IDF(t) = 1 + \log\left(\frac{N}{df(t)}\right)$$

Non-linear scaling

Total number of docs in collection

Number of docs containing term t

Ekstraksi Informasi

- Task menstrukturkan informasi dari dokumen tidak terstruktur secara otomatis.
- Task: text \rightarrow frame



Transformasi Teks ke Frame Informasi

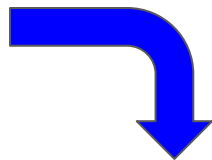
Named Entity

5W1H (What, Who..)

Dependency Grammar

Keterangan warna: ORANG ORGANISASI LOKASI LAIN-LAIN

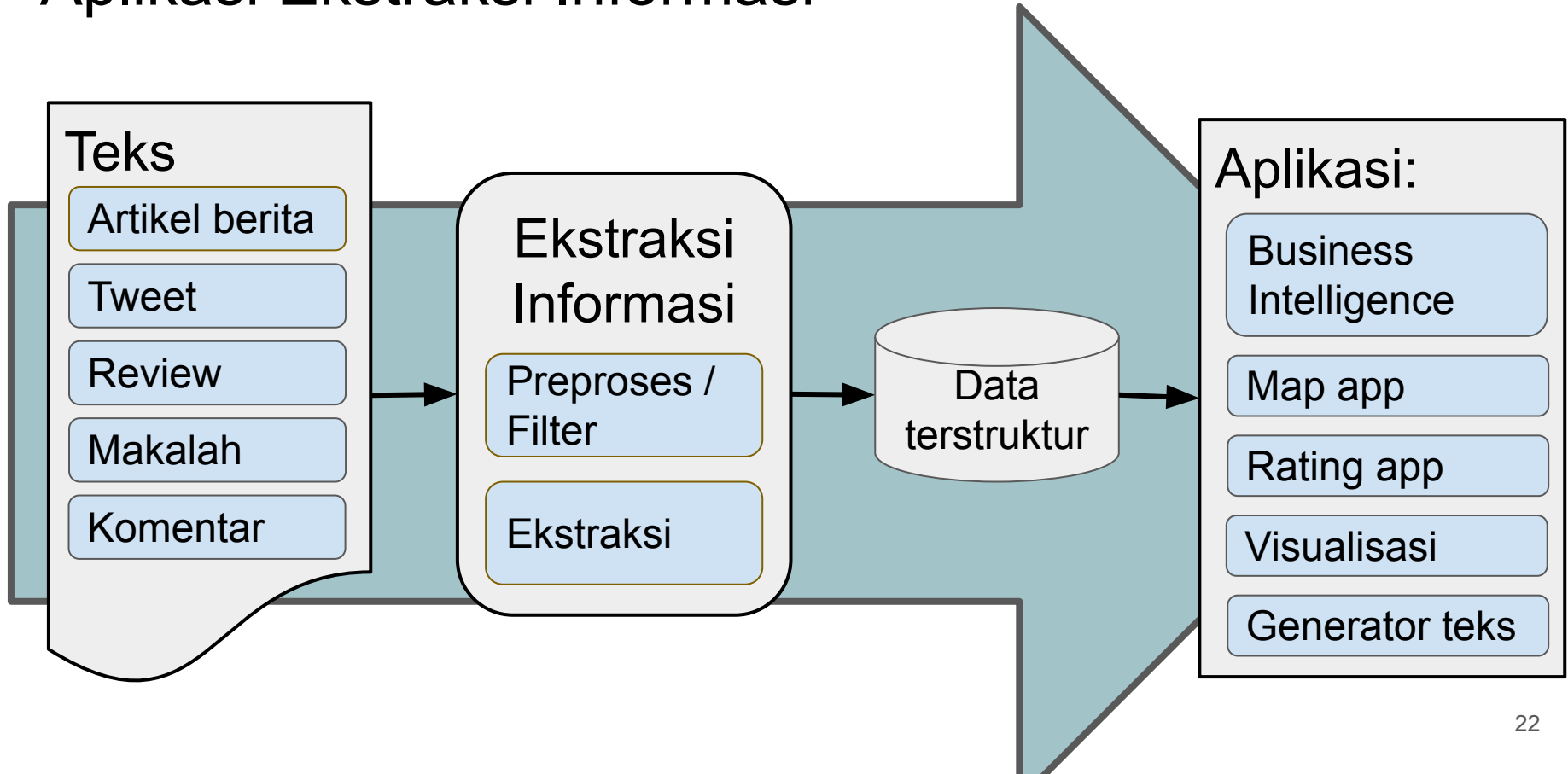
Jokowi beserta ketua PDIP , Megawati Soekarnoputri mendiskusikan Pemilu Serentak di Jakarta Pusat pada hari Kamis 29 Maret agar elektabilitas meningkat .



Tools: <http://nlp.yuliadi.pro/>

Orang:	Jokowi, Megawati Soekarnoputri
Organisasi:	PDIP
Lokasi:	Jakarta Pusat
Lain-lain:	Pemilu Serentak

Aplikasi Ekstraksi Informasi



Ekstraksi Informasi pada Tweet: Kondisi Lalu Lintas (Hasby & Khodra, 2013)

'#LalinBDG 20.10 – Sukajadi arah IP macet banget. Gara2 lampu merah lama dan lampu hijau yg sebentar bgt'

Named-Entity Recognition

Location: 'Sukajadi', 'IP'
Time: '20.10'
Condition: 'macet'

Relation Recognition

arah(Sukajadi, Jalan Pasir Kaliki),
kondisi(Jalan Pasir kaliki, macet),
kondisi(Sukajadi, macet).



Ekstraksi Informasi pada Artikel Berita: 5W1H

Masukkan kalimat atau langsung tekan tombol analisis.

Jokowi beserta ketua PDIP, Megawati Soekarnoputri mendiskusikan Pemilu Serentak di Jakarta Pusat pada hari Kamis 29 Maret agar elektabilitas meningkat.

Analisis!

Named Entity

5W1H (What, Who..)

Dependency Grammar

Keterangan warna: Who What Why When Where How

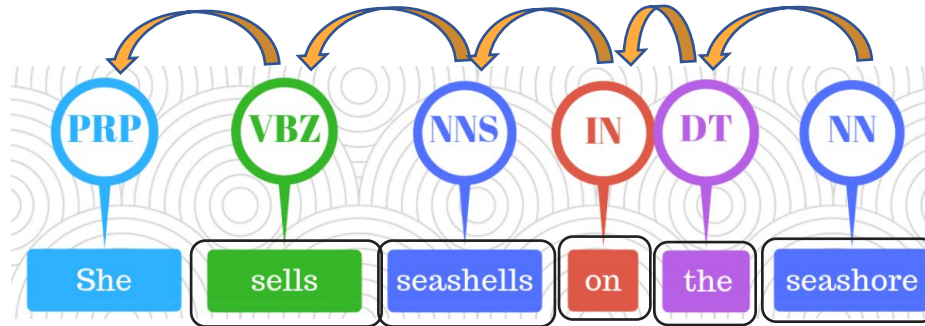
Jokowi beserta ketua PDIP , Megawati Soekarnoputri
mendiskusikan Pemilu Serentak di Jakarta Pusat pada hari
Kamis 29 Maret agar elektabilitas meningkat .

Who:	Jokowi beserta ketua PDIP , Megawati Soekarnoputri
What:	mendiskusikan Pemilu Serentak
When:	hari Kamis 29 Maret
Where:	Jakarta Pusat
Why:	agar elektabilitas meningkat
How	-

<http://nlp.yuliadi.pro/>

Ekstraksi Informasi dgn Sequence Labeling

- Classify an instance into its label as a sequence of labels
- Using the label of previous instance
- Example: POS Tagger



Sequence Labeling

- POS Tagger
 - Label: POS Tag, example: NN (Noun), VB (Verb), ADJ (Adjective), ADV (Adverb)
- Named Entity Tagger
 - Label: NE Tag, example: Person, Organization, Location
- Phrase Chunker
 - Label: Phrase Tag, example: NP, VP, PP
- Keyword Extraction
 - Label: Keyword tag, example: yes/no
- Answer Finder
 - Label: Answer, example: answer or no

Method for Sequence Labeling

- Rule based
 - top down method
 - Rules are defined manually by human
- Statistical (Machine learning) based
 - Bottom up method
 - corpus is used as training data to define the best tag probabilistically for a given word
 - Corpus is used to learn rules to select the predefined tag

Rule-based Tagging

- Steps:
 - Using dictionary to define tag to every input word
 - Using rules to eliminate the incorrect POS Tag □ rule concerns about the sentence context (neighbor words of a given word)
- Example from ENGTWOL
 - Pavlov □ *Pavlov N Nom Sg Proper*
 - Had □ *Have V Past ; Have Pcp2*
 - Shown □ *show pcp2*
 - That □ *Adv; Pron Dem Sg; Det; Cs*

Machine learning based Tagger

- Using machine learning to build Tagger
- Data is taken from a tagged corpus

Brown Corpus Sample (untagged)

A01 0010 The Fulton County Grand Jury said Friday an investigation
A01 0020 of Atlanta's recent primary election produced "no evidence" that
A01 0030 any irregularities took place. The jury further said in term-end
A01 0040 presentments that the City Executive Committee, which had over-all
A01 0050 charge of the election, "deserves the praise and thanks of the
A01 0060 City of Atlanta" for the manner in which the election was conducted.

Brown Corpus Sample (tagged)

A01_FO 0010_MC The_AT Fulton_NP1 County_NN1 Grand_JJ Jury_NN1 said_VVD Friday_NPD1 an_AT1 investigation_NN1
A01_FO 0020_MC of_IO Atlanta_93 's_03 recent_JJ primary_JJ election_NN1 produced_VVD " _ " no_AT evidence_NN1 " _ " that_CST
A01_FO 0030_MC any_DD irregularities_NN2 took_VVD place_NN1 .
The_AT jury_NN1 further_RRR said_VVD in_IL term-end_NN1
A01_FO 0040_MC presentments_NN2 that_CST the_AT City_NN1 Executive_NN1 Committee_NN1 , which_DDQ had_VHD over-all_RR
A01_FO 0050_MC charge_NN1 of_IO the_AT election_NN1 , " _ " deserves_VVZ the_AT praise_NN1 and_CC thanks_NN2 of_IO the_AT
A01_FO 0060_MC City_NN1 of_IO Atlanta_NP1 " _ " for_IF the_AT manner_NN1 in_IL which_DDQ the_AT election_NN1 was_VBDZ conducted_VVN .

Preprocessing



Feature extraction

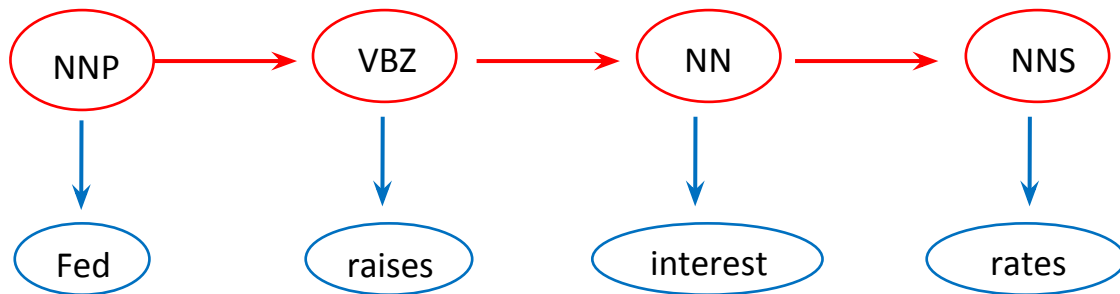


Model Training



Accuracy

Data



- Feature (X):

- Important:

- Current word
 - Preceding POS tag □ SEQUENCE!!!

- Optional:

- Word window: preceding words, succeeding words
 - Orthography information: lowercase/uppercase/number/symbol, etc

- Class (Y):

- POS Tags: NN, ADJ, ADV, VB, etc

```
import nltk
nltk.download('treebank')
```

```
tagged_sentences = nltk.corpus.treebank.tagged_sents()
print(tagged_sentences[0])
```

```
def features(sentence, index):
    """ sentence: [w1, w2, ...], index: the index of the word """
    return {
        'word': sentence[index],
        'is_first': index == 0,
        'is_last': index == len(sentence) - 1,
        'is_capitalized': sentence[index][0].upper() == sentence[index][0],
        'is_all_caps': sentence[index].upper() == sentence[index],
        'is_all_lower': sentence[index].lower() == sentence[index],
        'prefix-1': sentence[index][0],
        'prefix-2': sentence[index][:2],
        'prefix-3': sentence[index][:3],
        'suffix-1': sentence[index][-1],
        'suffix-2': sentence[index][-2:],
        'suffix-3': sentence[index][-3:],
        'prev_word': ' ' if index == 0 else sentence[index - 1],
        'next_word': ' ' if index == len(sentence) - 1 else sentence[index + 1],
        'has_hyphen': '-' in sentence[index],
        'is_numeric': sentence[index].isdigit(),
        'capitals_inside': sentence[index][1:].lower() != sentence[index][1:]
    }
```

Preprocessing &
Feature
Extraction

```
from nltk.tag.util import untag

# Split the dataset for training and testing
cutoff = int(.75 * len(tagged_sentences))
training_sentences = tagged_sentences[:cutoff]
test_sentences = tagged_sentences[cutoff:]

def transform_to_dataset(tagged_sentences):
    X, y = [], []

    for tagged in tagged_sentences:
        X.append([features(untag(tagged), index) for index in range(len(tagged))])
        y.append([tag for _, tag in tagged])

    return X, y
```

```
X_train, y_train = transform_to_dataset(training_sentences)
X_test, y_test = transform_to_dataset(test_sentences)
```

```
print(len(X_train))
print(len(X_test))
print(X_train[0])
print(y_train[0])
```

Data:

- Preprocessing
- Feature Extraction

Text Summarization

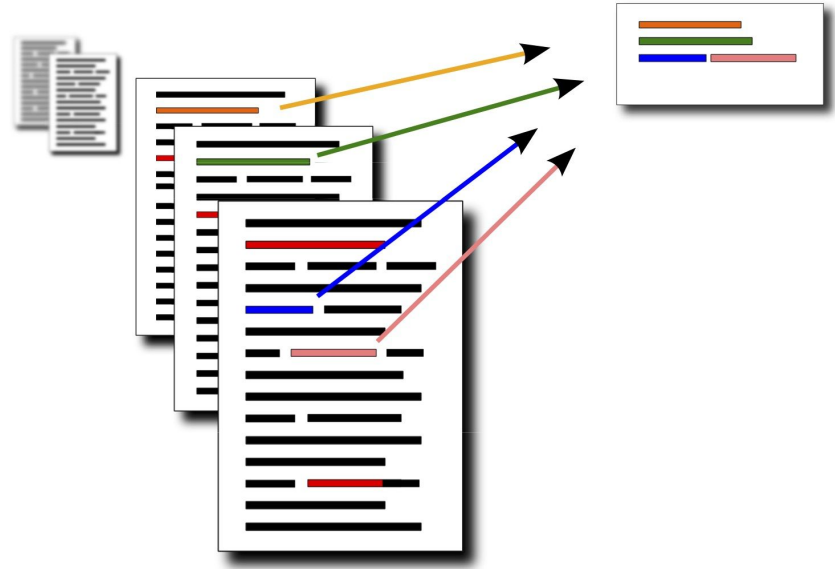


Text Summarization is the process of condensing a source text into a shorter version preserving its information content

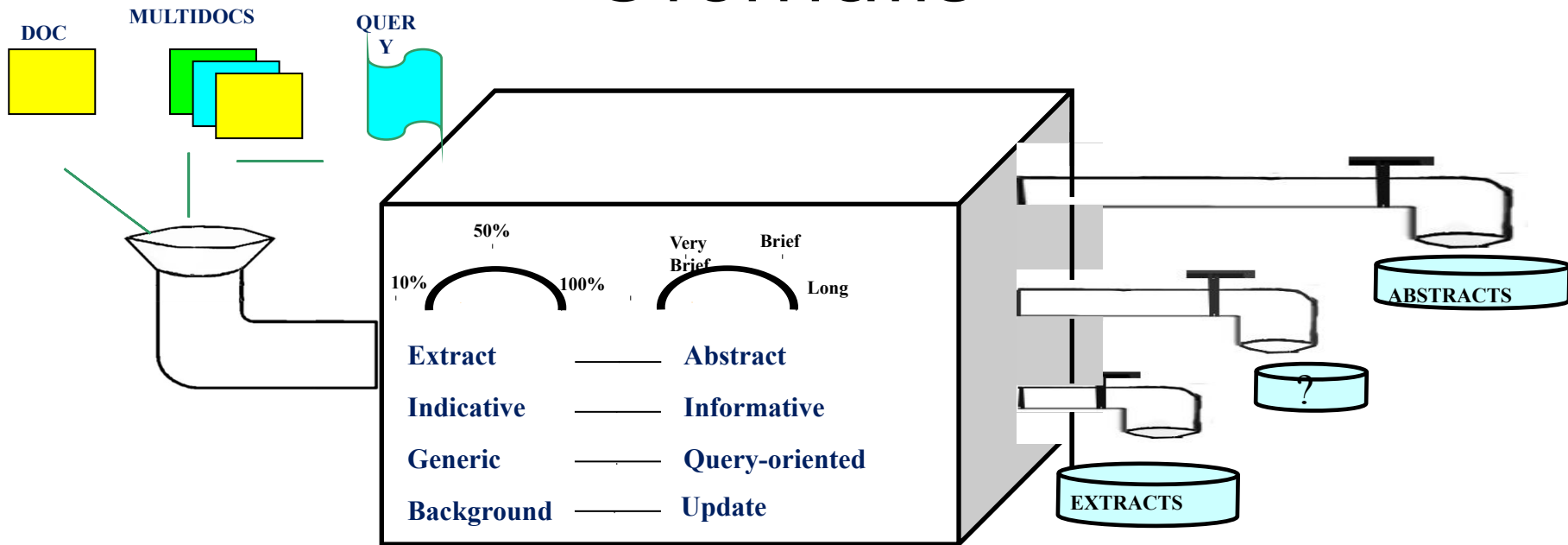
Peringkasan dokumen

Task melakukan identifikasi informasi-informasi penting dari 1 atau kumpulan dokumen dan menghasilkan satu teks pendek (ringkasan).

Pendekatannya: ekstraktif,
abstraktif



Mesin Peringkasan Teks Otomatis



<http://www.isi.edu/natural-language/people/{hovv.html,marcu.html}>

Ringkasan Satu Dokumen, Abstraktif, Generik

KETUA RAPAT (FADLI ZON, S.S., M.Sc./WAKIL KETUA DPR RI BIDANG POLKAM):

Yang terhormat para Anggota DPR RI;
Hadirin yang berbahagia.

Assalamu'alaikum Warrahmatullahi Wabarakatuh.

Pertama marilah kita panjatkan puji syukur kehadiran Allah SWT, Tuhan Yang Maha Esa yang telah melimpahkan rahmat dan karunia-Nya kepada kita semua, sehingga kita dapat mengikuti Rapat Paripurna DPR RI hari ini dalam keadaan sehat walafiat untuk melaksanakan tugas konstitusional kita. Menurut catatan dari Sekretariat Jenderal DPR RI, hadir dalam permulaan rapat hari ini ditandatangani 296 Anggota DPR RI. Dengan demikian kuorum telah tercapai, dan dengan mengucapkan *bismillah* marilah kita berkenankan kami selaku Pimpinan Dewan membuka Rapat Paripurna DPR RI yang ke-11 Masa Sidang II Tahun Sidang 2015-2016 hari Selasa Tanggal 24 November 2015 dan kami nyatakan terbuka untuk umum.

(RAPAT DIBUKA PUKUL 11.20 WIB)

Hadirin yang kami hormati,

Mari kita berdiri menyanyikan Lagu Kebangsaan Indonesia Raya.

(MENYANYIKAN LAGU INDONESIA RAYA)

Sidang Dewan yang kami hormati.

Sesuai hasil keputusan Rapat Konsultasi pengganti Rapat Bamus DPR RI antara Pimpinan DPR RI dan Pimpinan fraksi-fraksi tanggal 19 November 2015 acara Rapat Paripurna hari ini adalah penetapan badan Keamanan Laut RI menjadi mitra kerja Komisi I DPR RI. Sekarang kami menanyakan pada Sidang Dewan yang terhormat apakah acara rapat tersebut dapat disetujui?

F-PAN (H. YANDRI SUSANTO):

Pimpinan,

Yandri Pimpinan.

KETUA RAPAT:

Ya silakan.

A. Pendahuluan

1. Rapat Paripurna dibuka Pukul 11.40 WIB dipimpin oleh Yth. Bapak Dr. Ir. H. Taufik Kurniawan, M.M., Wakil Ketua DPR RI/ Korekku dihadiri oleh 351 orang dari 557 orang Anggota DPR RI.
2. Rapat Paripurna menyetujui secara Rapat pada hari ini adalah Laporan Badan Legislasi terhadap:
 - a. RUU tentang Pengampunan Pajak; dan
 - b. RUU tentang Perubahan atas Undang-Undang Nomor 30 Tahun 2002 tentang Komisi Pemberantasan Tindak Pidana Korupsi; Untuk menjadi RUU Prioritas Prolegnas Tahun 2015 dan dilanjutkan dengan Pengambilan Keputusan. (Didahului oleh pelantikan Anggota Pengganti Antar Waktu DPR RI).
3. Rapat Paripurna didahului dengan Pelantikan Anggota Pengganti Antar Waktu DPR RI yaitu:
 - 1) Saudara *Ir. H. Hari Kartana, M.M.Ph.D., D.Sc.* dari Partai Demokrat Daerah Pemilihan Jawa Barat VII, menggantikan Saudara Sean Mustopa, M.Si
 - 2) Saudari *Hj. Adji Farida Padmo Ardans* dari Partai Demokrat Daerah Pemilihan Kalimantan Timur, menggantikan Saudara Norbali Isran Noor, A.Md.
 - 3) Saudari *dr Anarulita Muchtar* dari Partai NasDem Daerah Pemilihan Bengkulu menggantikan Saudara Patrice Rio Capella, S.H.
 - 4) Saudari *dr Ari Yunita* dari Partai NasDem Daerah Pemilihan Kalimantan Timur menggantikan Saudara Dr. H. Achmad Amira, M.M.
 - 5) Saudara *Djenri Alting Keingem, S.H., M.H* dari Partai Demokrasi Indonesia Perjuangan Daerah Pemilihan Sulawesi Utara menggantikan Saudara Oly Dondokambey, S.E.
 - 6) Saudari *Nurhayati* dari Partai Kebangkitan Bangsa Daerah Pemilihan Lampung II menggantikan Saudari Hj. Chusnunia Chaim, M.Si
 - 7) Saudara *Abdul Latief Hanafiah, M.Sc* dari Partai Kebangkitan Bangsa Daerah Pemilihan Kalimantan Selatan II menggantikan Saudara Dr. H.M. Zainullah Azhar.
4. Pimpinan Rapat membacakan 6 (enam) pokok surat masuk yaitu:
 - 1) Surat dari Presiden R.I. dengan Nomor R-70/Pres/11/2015 tertanggal 26 November 2015, perihal Permohonan Pertimbangan bagi Pencalonan Duta Besar Luar Biasa dan Berkuasa Penuh (Dubes LBBP) Negara Sahabat untuk R.I.
 - 2) Surat dari Presiden R.I. dengan Nomor R-72/Pres/12/2015 tertanggal 10 Desember 2015, perihal penunjukan wakil untuk membahas Rancangan

Risalah rapat DPR

Pokok-pokok Pembicaraan Rapat

Ringkasan Satu Dokumen, Ekstraktif, *Tailored*

Makalah Tugas Akhir (S1 IF-ITB)

Template 1

Title
Background
Objective
Method
Experiment
Conclusion
Suggestion (opt)

Template 2

Title	
Background	Objective
Method	
Experiment	
Conclusion	Suggestion (opt)

 <h2>Ensemble Technique Utilization for Indonesian Dependency Parser</h2>	
<h3>Background</h3> <p>The accuracy indifference may be because of the fact that all of the reparsing algorithms used unweighted weighting scheme which would make the weight of many dependency relations to be the same regardless of the algorithm.</p> <p>With this corpus sentences with deletion can be parsed better but gerunds can't yet because the current available features aren't adequate to differentiate verb and gerund.</p>	<h3>Objective</h3> <p>In this research we propose to add diversity to the treebank with sentences that contain gerund and/or deletion and use ensemble techniques.</p>
<h3>Method</h3> <p>Each directed edge represents the dependency from head to dependant and is ed with the dependency that occurs from dependant to head.</p> <p>Suppose that we want to parse this sentence Adik membeli esau kemarin with its CoNLL representation shown below.</p> <p>The system also used reparsing algorithm to ensure the resulting dependency graphs always form a tree.</p> <p>Up until now there have been only a few researches regarding Indonesian dependency parsing.</p>	
<h3>Experiment</h3> <p>There were six ensemble combinations that were compared all parsing algorithms both from MultiParser and MSTParser all algorithms from MultiParser all algorithms from MSTParser all projective parsing algorithms all nonprojective algorithms and three algorithms with the highest accuracy according to the second scenario.</p> <p>There were eleven base parsers that were compared Nivre eager Nivre standard Covington projective Covington nonprojective Stack projective Stack eager Stack lazy Planar 2Planar Eisner and ChuLiu Edmonds.</p> <p>This is most likely because of the rich feature representations that transition-based model has which depends on the data structures used to represent the parsing state.</p> <p>The first is a set of parsing states which is generally represented by set of tokens in process set of unprocessed tokens and set of created dependencies.</p>	
<h3>Conclusion</h3> <p>We can also see that parsers trained with corpus that were added with only one sentence type deletion or gerund exclusively have lower accuracy than the old parser.</p> <p>There are three researches that are related to ensemble dependency parsing which is Burdakov Manning's work for English and Green et al.</p> <p>This research used Afri's and Firdian's corpus that was enhanced which contains 200 Indonesian sentences added with 362 sentences annotated by annotators with linguistic background.</p> <p>The parsers used Eisner reparsing algorithm with unweighted weighting scheme and were tested using 10-fold cross validation and used the same learning algorithm SVM.</p>	<h3>Suggestion</h3> <p>The results from Table VII shows that ChuLiu Edmonds algorithm is the best parsing algorithm to be used for Indonesian sentences.</p>

Ringkasan Multidok, Abstraktif, Generik

Olahraga



Pada Sabtu, Managing Director Yamaha Motor Racing (VMR) Lin Jarvis mengakui dirinya senang dengan peta kekuatan MotoGP yang mulai merata karena jalannya balapan lebih menarik.

News

- Peta Kekuatan MotoGP Mulai Rata, Bos Yamaha Semringah 10 Dec 2016

WHAT: mengakui dirinya senang dengan peta kekuatan MotoGP yang mulai merata

WHERE:

WHO: Managing Director Yamaha Motor Racing (VMR) Lin Jarvis

WHEN: Sabtu

WHY: jalannya balapan lebih menarik

HOW:

Baca

- MotoGP 2016 Dihiasi Sembilan Pemenang, Livio Suppo: Ini Sult buat Repsol Honda 10 Dec 2016

WHAT: mengatakan hal tersebut justru menyulitkan dirinya

WHERE:

WHO: Team Principal Repsol Honda Livio Suppo

WHEN: Sabtu

WHY: dihiasi sembilan pemenang berbeda

HOW:

Baca

Category

News Aggregator

Olahraga



News

- Peta Kekuatan MotoGP Mulai Rata, Bos Yamaha Semringah 10 Dec 2016
- MotoGP 2016 Dihiasi Sembilan Pemenang, Livio Suppo: Ini Sult buat Repsol Honda 10 Dec 2016
- MotoGP 2016 Jadi Mimpi Buruk Lorenzo 10 Dec 2016

more

Summary



News

- Penambahan Hamilton Jadi Keuntungan Mercedes 10 Dec 2016
- Niki Lauda: Mercedes Akan Umumkan Pengganti Rosberg di Akhir 2016 10 Dec 2016
- Rosberg Berharap Vettel Jadi Penggantinya di Mercedes 10 Dec 2016

more

Cluster

Teknologi dan sains



News

- Jualo.com tinggalkan layanan di Harbinas 10 Dec 2016
- Ponsel dan Komputer Belas Iktu Diburu Saat Harbinas 10 Dec 2016
- Harbinas 2016, Barang Belas Dapat Subsidy Dialok 10 Dec 2016

more



News

- Konsumsi Coleket pada Penderita Kanker Bisa Picu Kematian 08 Dec 2016
- Jangan Buang Air Rebusan Ubi jika Mau Langsing 08 Dec 2016
- Air Rebusan Ubi Mampu Turunkan Berat Badan 08 Dec 2016

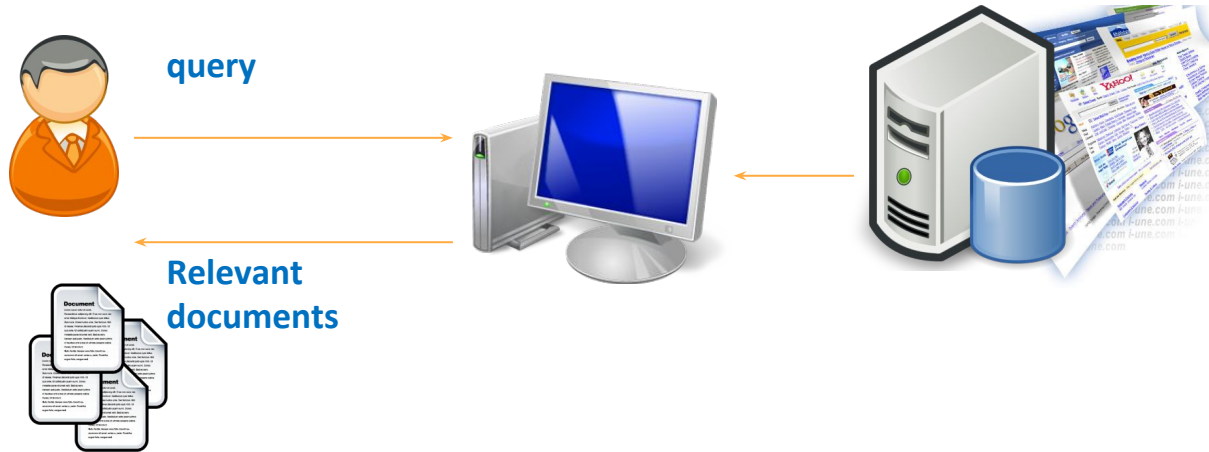
more

Ringkasan Multidokumen, Abstraktif

Generated Summary	Reference Summary
<p>Fujitsu maupun Indosat Ooredoo meyakini kemitraan strategis yang terjalin ini nantinya akan mampu mendukung keduanya dalam mendayagunakan seluruh kekuatan yang mereka miliki dalam mengatasi seluruh tantangan yang akan hadir di masa depan . Ke depan , kerja sama ini akan diperluas ke berbagai sektor industri termasuk sektor publik untuk memenuhi kebutuhan pelanggan korporasi di Indonesia dan perusahaan-perusahaan Jepang yang beroperasi di Indonesia . JAKARTA -- Indosat Ooredoo bersama Fujitsu Indonesia menandatangani nota kesepahaman dalam rangka memantapkan kemitraan dengan para pelanggan bisnis yang telah dibangun dan bekerja sama dalam menghadirkan solusi smart mobility dan internet of things (IoT) .</p>	<p>Indosat Ooredoo bersama Fujitsu Indonesia menandatangani Nota Kesepahaman dalam rangka memantapkan kemitraan dengan para pelanggan bisnis yang telah dibangun dan bekerja sama dalam menghadirkan solusi Smart Mobility dan Internet of Things (IoT) . Sebagai awal , kerjasama ini akan berfokus pada sektor otomotif dan transportasi dan akan diperluas ke berbagai sektor industri termasuk sektor publik untuk memenuhi kebutuhan pelanggan korporasi di Indonesia tetapi juga perusahaan-perusahaan Jepang yang beroperasi di Indonesia . Kemitraan ini mensinergikan piranti , layanan , dan solusi teknologi informasi yang mutakhir dari Fujitsu dengan teknologi telekomunikasi terdepan Indosat Ooredoo untuk menghadirkan produk dan layanan-layanan baru yang bermanfaat bagi pelanggan korporasi di Indonesia .</p>

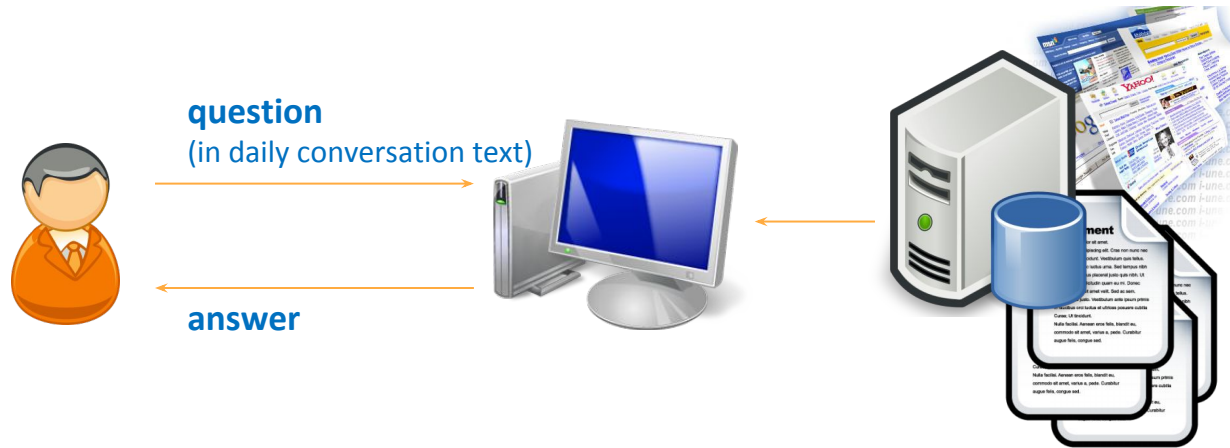
Information Retrieval

System able to return relevant documents related with query input



Question Answering

System able to return answer or snippet of user question



Machine Translation



System able to transform text from
a certain natural language into
other natural language



Natural Language Processing

Text Categorization

Text Clustering

Group 1

Cleanup, Tokenization

Stemming

Lemmatization

Part of Speech Tagging

Query Expansion

Parsing

Topic Segmentation and
Recognition

Morphological Degmentation
(Word/Sentences)

Group 2

Information Retrieval and
Extraction (IR)

Relationship Extraction

Named Entity Recognition
(NER)

Sentiment Analysis/Sentence
Boundary Disambiguation

World sense and
Disambiguation

Text Similarity

Coreference Resolution

Discourse Analysis

Group 3

Machine Translation

Automatic Summarization/
Paraphrasing

Natural Language Generation

Reasoning over
Knowledge Based

Quation Answering System

Dialog System

Image Captioning & other
Multimodel Tasks