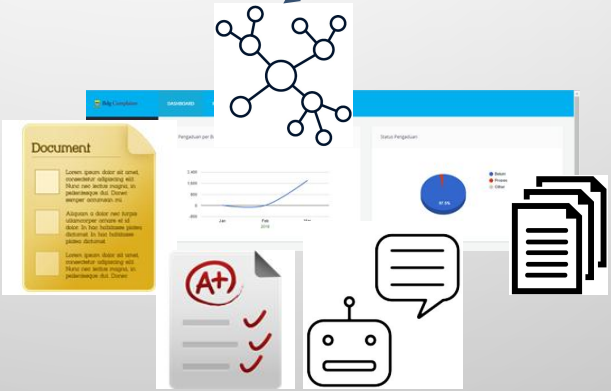
NATURAL LANGUAGE PROCESSING

AYU PURWARIANTI
AYU@INFORMATIKA.ORG



COMPUTATIONAL LINGUISTICS

TEXT MINING

NLU NLG

Text is …

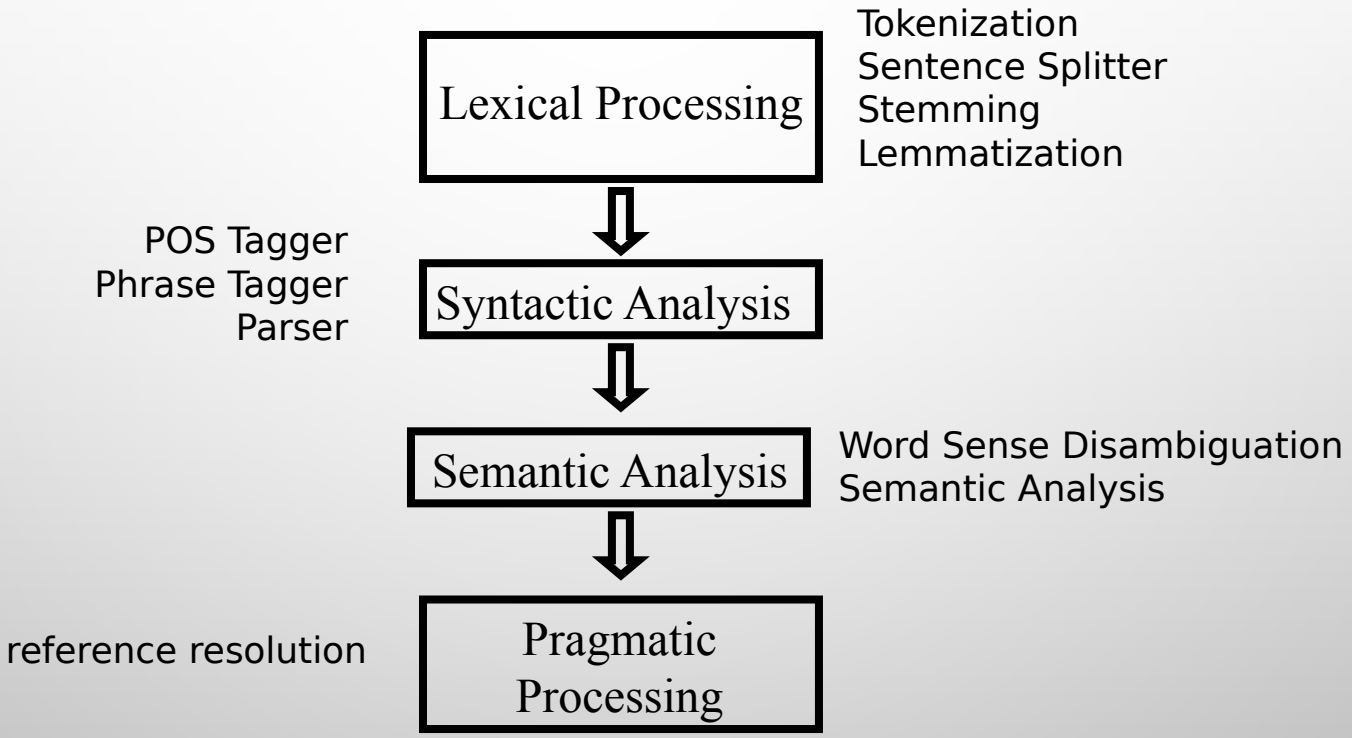
a string

list of tokens

having structure

represent an intention

BASIC NLP TOOLS



NLP TOOLKIT

- NLTK
  - TOKENIZER, POS TAGGER, NAMED ENTITY TAGGER, PARSER, STEMMING, SEMANTIC ANALYZER, SENTIMENT ANALYZER, MACHINE TRANSLATION, TEXT CLASSIFICATION, CLUSTERING, COLLOCATION
- OPENNLP
  - SENTENCE DETECTION, TOKENIZATION, NAMED ENTITY TAGGER, PART OF SPEECH TAGGER, DOCUMENT CATEGORIZER, CHUNKER, PARSER, COREFERENCE RESOLUTION
- STANFORD CORENLP
  - SENTENCE SPLITTER, TOKENIZATION, POS TAGGER, LEMMATIZATION, NAMED ENTITY TAGGER, PARSER, COREFERENCE RESOLUTION, SENTIMENT ANALYSIS, MENTION DETECTION, OPEN IE

https://docs.prosa.ai/syntax/

Prosa NLP API Documentations

Search

Syntactic Analyzer

Extracts syntactic information from a given text. The API uses four separate modules: 1. Sentence splitter: Splits the given text by sentences 2. Tokenizer: Splits each split sentence in the given text by tokens 3. POS tagger: Assign part-of-speech tag to each token in the given text 4. Stemmer: Extracts the lemma and the affixes present in each token in the given text

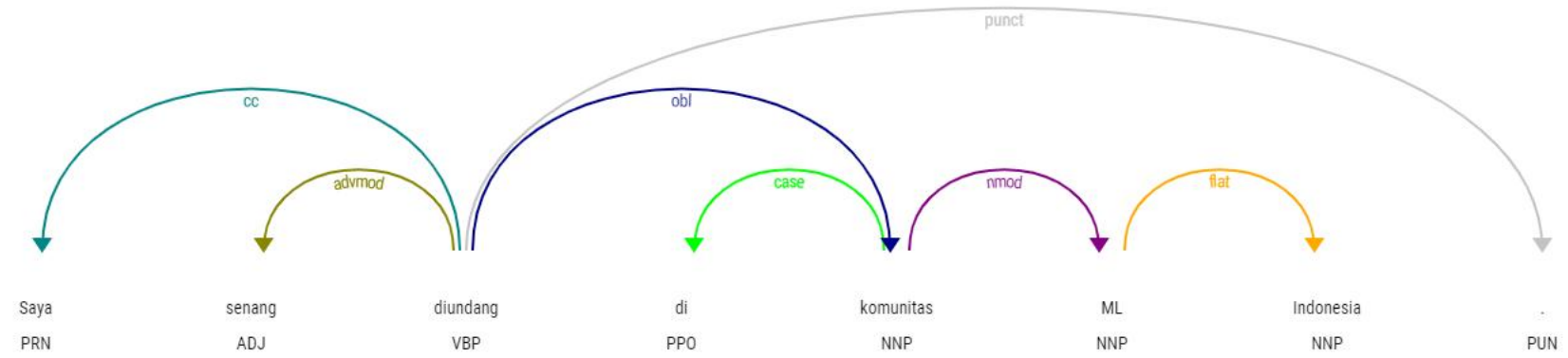
Request Method

`POST`

Request URL

```
https://api.prosa.ai/v1/syntax
```

Dependency Parser



Named Entity Recognizer

ORGANIZATION

Saya senang diundang di komunitas ML Indonesia.


LEXALYTICS


bing


Google


Google translate


LingvoSoft
THE JOY OF UNDERSTANDING

SEVERAL BENEFITS OF USING NLP

improve user experience	• spelling checker, completer, recommendation
automate support	• chatbot
monitor and analyze feedback	• sentiment analysis
make things faster	• search engine, summarization
make things easier	• virtual assistant, machine translation

EMAIL

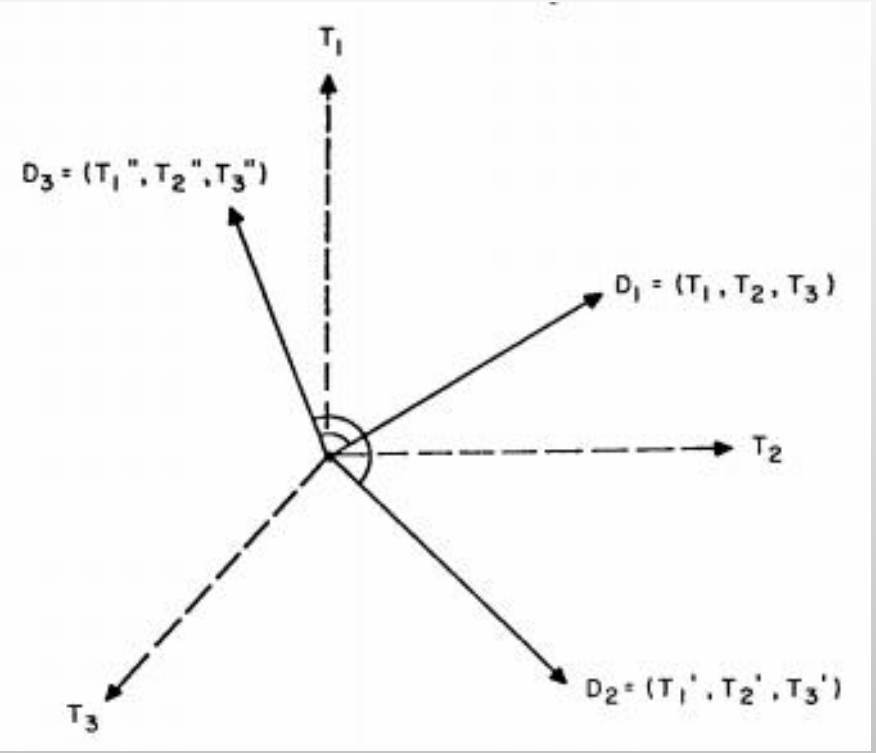
Spam Filtering	**Text classification** Spam vs Not Spam
Email type	**Text classification** Meeting invitation, personal question, broadcast information
Scheduler	**Information Extraction** Date, time, location, agenda
Summarization	Summary on the discussion

DEFINITION OF TEXT CLASSIFICATION

- SYSTEM TO CLASSIFY OR TEXT (OR DOCUMENT) INTO A CERTAIN LABEL
- EXAMPLE:
  - SPAM FILTERING:
    - LABEL THE INPUT (SUCH AS SMS OR EMAIL) INTO SPAM OR NOT SPAM LABELS
  - NEWS CLUSTERING:
    - GROUP THE NEWS ARTICLE INTO CERTAIN CATEGORY SUCH AS POLITICS, SPORTS, ECONOMY, ETC
- = “TEXT CLASSIFICATION”, “DOCUMENT CLASSIFICATION”, “TEXT CATEGORIZATION”, “DOCUMENT CATEGORIZATION”

SPAM FILTERING

- CLASSIFY TEXT INPUT INTO 2 CLASSES: SPAM OR NOT SPAM
- SIMPLE METHOD:
  - USE A WORD LIST THAT CONSISTS OF THE MOST FREQUENT WORDS IN THE SPAM TEXT ; AND THEN USE THRESHOLD RULE, FOR EXAMPLE:
    - IF >= 75% WORDS IN INPUT TEXT ARE AVAILABLE IN THE WORD LIST THEN THE TEXT IS CLASSIFIED AS SPAM
  - WORD LIST 🖂 IS CALLED **FEATURE**
  - RULE WITH >= 75% AS THE THRESHOLD 🖂 IS CALLED THE **CLASSIFICATION TECHNIQUE** WHICH CAN BE RULE OR MODEL



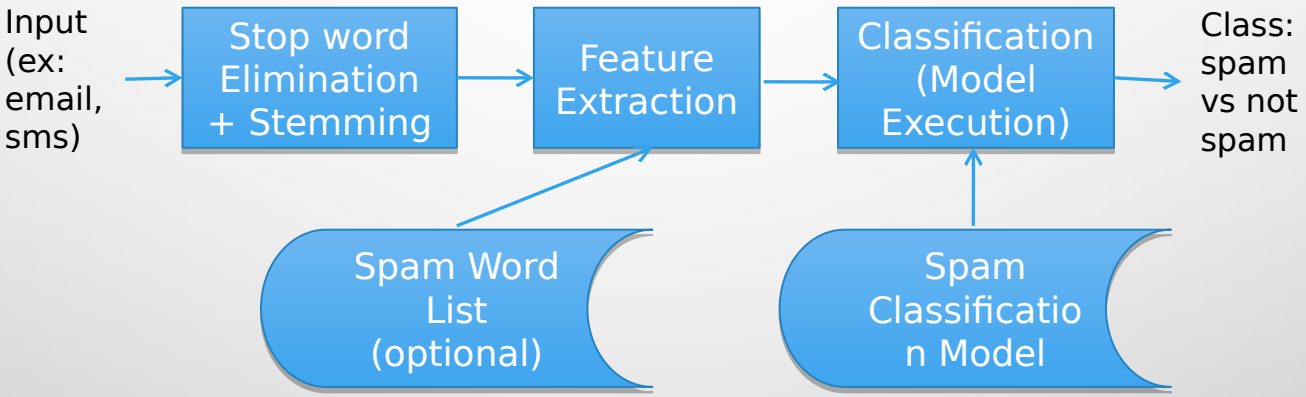
VECTOR SPACE MODEL

- TEXT INPUT AS VECTOR OF TERMS OR WORDS

UNIGRAM APPROACH

- LIST OF WORDS TAKEN FROM ALL TEXT INPUTS AS THE FEATURES
- PROBLEMS: FEATURE SIZE
- DECREASE THE FEATURE SIZE, HOW?
  - STEMMING
  - STOP WORD ELIMINATION
  - TAKE ONLY N-TOP WORDS WITH HIGH WORD WEIGHT
  - SYNONYM

EXAMPLE OF STATISTICAL BASED SPAM FILTERING



EXAMPLE OF LEXICAL 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: 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

LEXICAL FEATURE FOR SPAM FILTERING

- LEXICAL BASED
  - MANUALLY WRITE WORDS OCCUR IN SPAM TEXT
    - BASED ON MANUAL OBSERVATION
    - SMALL DATA SIZE
  - STATISTICAL BASED
    - WORDS IN A WORD LIST HAVE HIGH WORD WEIGHT OR ABOVE CERTAIN THRESHOLD

WORD LIST (LEXICAL BASED)

- WEAKNESSES:
  - SPAM WORDS CAN ALSO BE COMMON WORDS WITH HIGH OCCURRENCE FREQUENCY , SUCH AS *INI* (ENG: THIS), *DI* (ENG: AT), ETC
    - STOP WORD ELIMINATION
  - SPAM WORDS CAN ALSO OCCUR IN NOT SPAM TEXT, FOR EXAMPLE MAMA
    - WORD WEIGHT
      - TF X IDF = TERM FREQUENCY / DOCUMENT FREQUENCY (CONTAINS TERM)
        - IDF = 1/DF
        - IDF = LOG (N/DF)
      - MUTUAL INFORMATION (MI) =
        - N: NUMBER OF DOCUMENT WITH LABEL (L),
        - A: FREQUENCY OF WORD (T) WITH LABEL (L);
        - B: FREQUENCY OF WORD (T) WITHOUT LABEL (L);
        - C: NUMBER OF DOCUMENT WITH LABEL (L) WITHOUT WORD (T)

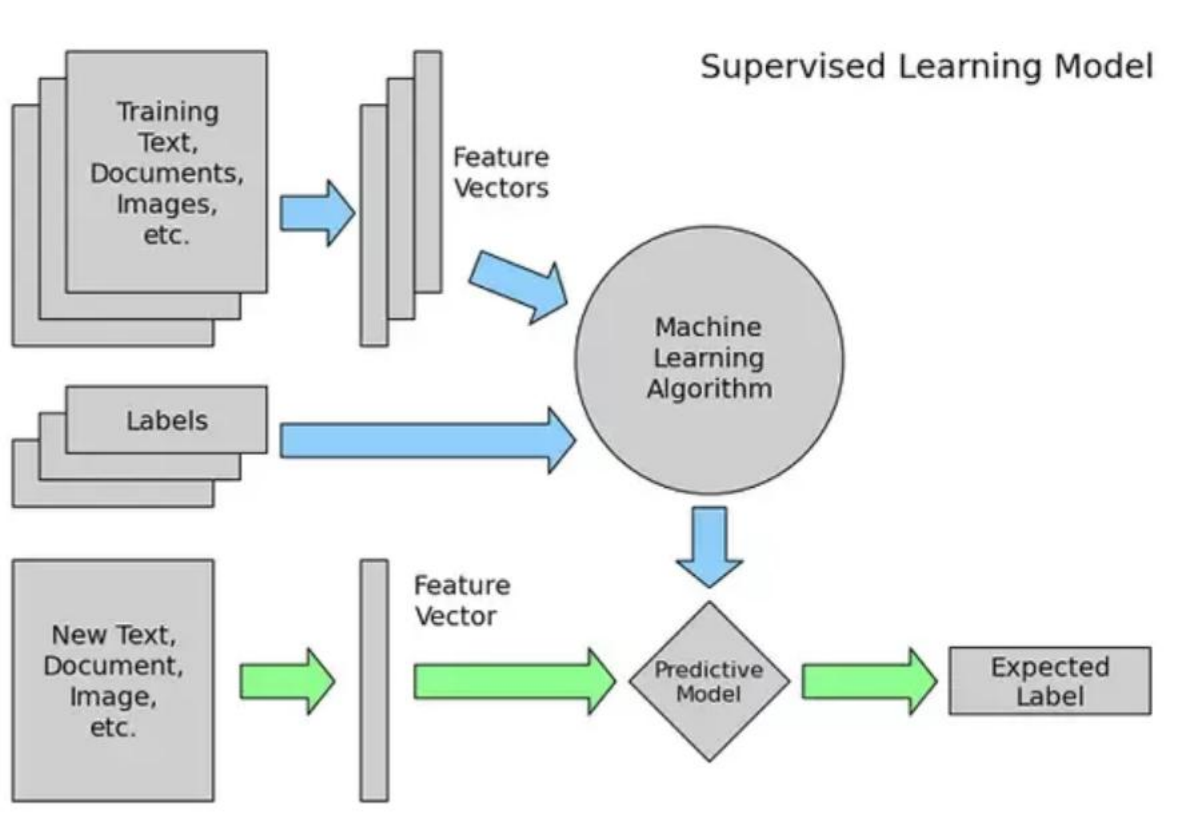
$$MI(t,l) = log_2 \frac{p(t,l)}{p(t)\ x\ p(l)} \approx log_2 \frac{A\ x\ N}{(A+\ B)\ x\ (A+C)}$$

FEATURE

- LEXICAL BASED
  - IDEA: TO CLASSIFY AN INPUT BASED ON ITS WORDS. TO DECREASE THE NUMBER OF WORDS, THE WORDS CAN BE SELECTED FIRST, BY USING:
    - SPECIFIC WORD LIST (MANUALLY SELECTED)
    - NAMED ENTITY
    - TFXIDF
    - MUTUAL INFORMATION
    - POS TAG INFORMATION (FOR EXAMPLE: ONLY NOUN & VERB)
- SYNTACTICAL PARSER
  - IDEA: TO TAKE INTO ACCOUNT, THE WORD ORDER OR GRAMMAR; OR TO DO WORD SELECTION BASED ON ITS SYNTACTICAL INFORMATION
  - SHALLOW PARSER
  - DEEP PARSER
  - N-GRAM
- SEMANTIC

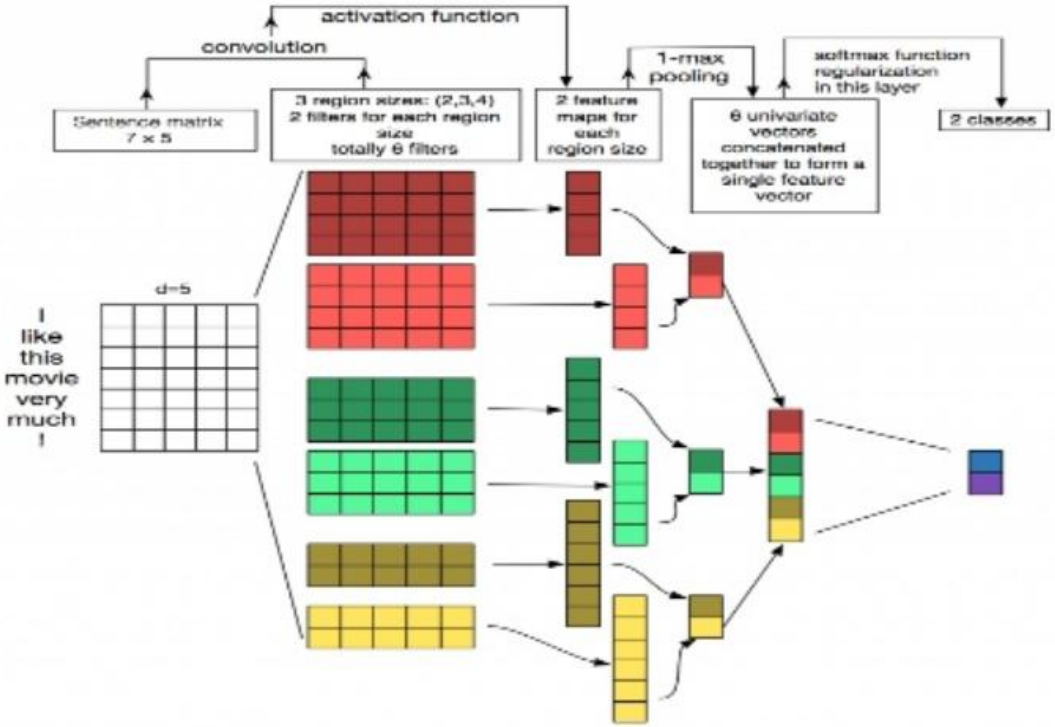
CLASSIFICATION ALGORITHM FOR SPAM FILTERING

- MANUAL
  - IF “NUMBER OF SPAM WORDS > THRESHOLD” THEN SPAM
- STATISTICAL BASED (MACHINE LEARNING)
  - SPAM RULES ARE DISCOVERED BY MACHINE LEARNING ALGORITHM (EXAMPLE: ID3 OR C4.5)
    - EXAMPLE: IF EMAIL CONTAINS W1 AND W2 THEN THE EMAIL IS SPAM
  - PROBABILITY SCORE OF AN INPUT, DEFINED AS SPAM
    - CONTOH: P(W1|SPAM) = 0.3
  - DEEP LEARNING
    - CNN, RNN (LSTM, GRU)

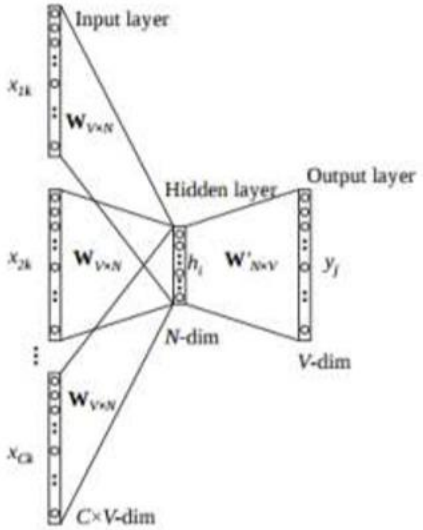


SOURCE: HTTPS://WWW.QUORA.COM/WHAT-DEEP-LEARNING-METHOD-TO-USE-TO-CLASSIFY-TEXT-FILES

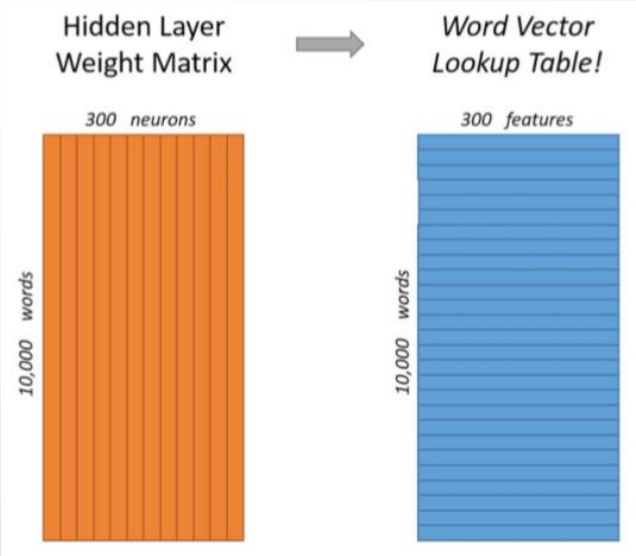
CNN for NLP



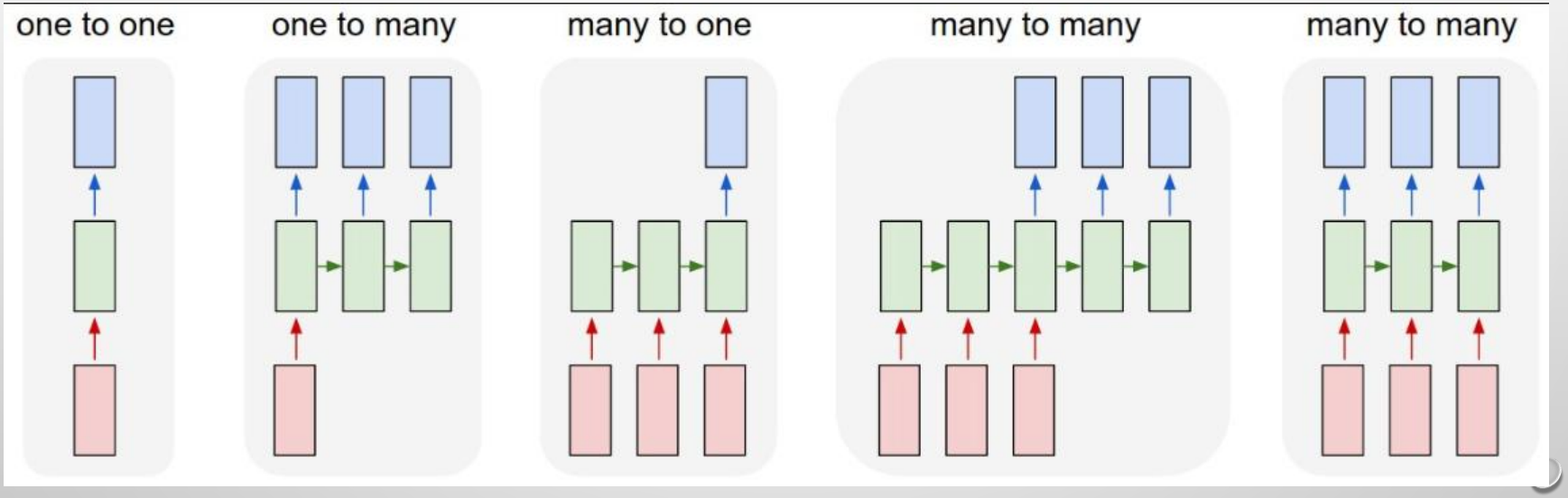
images found in the WildML blog:
http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/
also very good tutorial on CNN for NLP with Tensorflow
http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/



- The weight between the hidden layer and the output layer is taken as the word vector representation of the word.



RNN FOR NLP



- SOURCE: HTTP://KARPATHY.GITHUB.IO/2015/05/21/RNN-EFFECTIVENESS/

ARTICLES

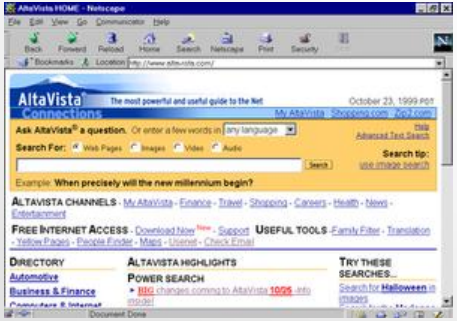
MILLION OF ARTICLES

- SEARCH ENGINE
  - SEARCHING RELEVANT ARTICLES USING GIVEN QUERY
- RECOMMENDATION
  - RECOMMEND A RELATED ARTICLES BASED ON USER HISTORY
- SUMMARIZATION
  - SUMMARIZE ONE OR SEVERAL ARTICLES
- NEWS AGGREGATOR
  - NEWS SUMMARIZATION AND TOPIC DETECTION
- PLAGIARISM DETECTION
  - CHECK THE SIMILARITY OF NEW ARTICLE TO EXISTING ARTICLES

SEARCH ENGINE

MAKE SEARCHING EASIER

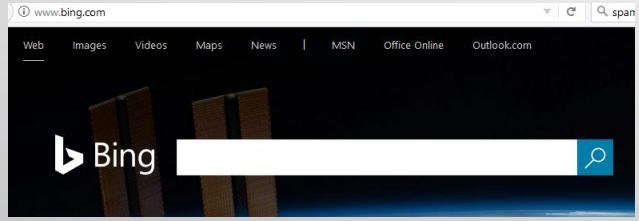
Altavista – 1995, 2003, 2013



Yahoo – 1994

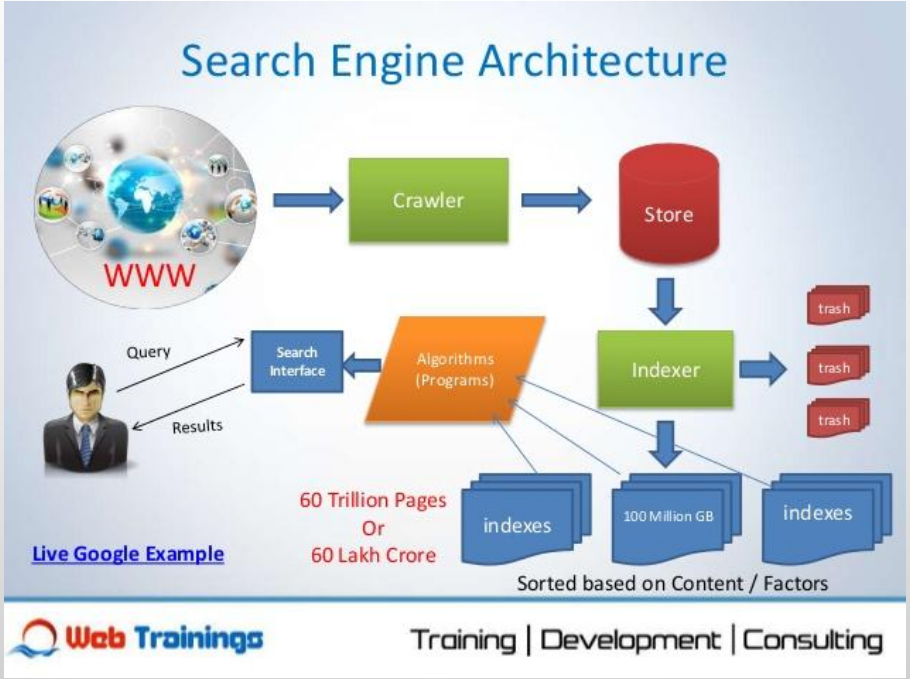
Google - 1998

Bing – 2009



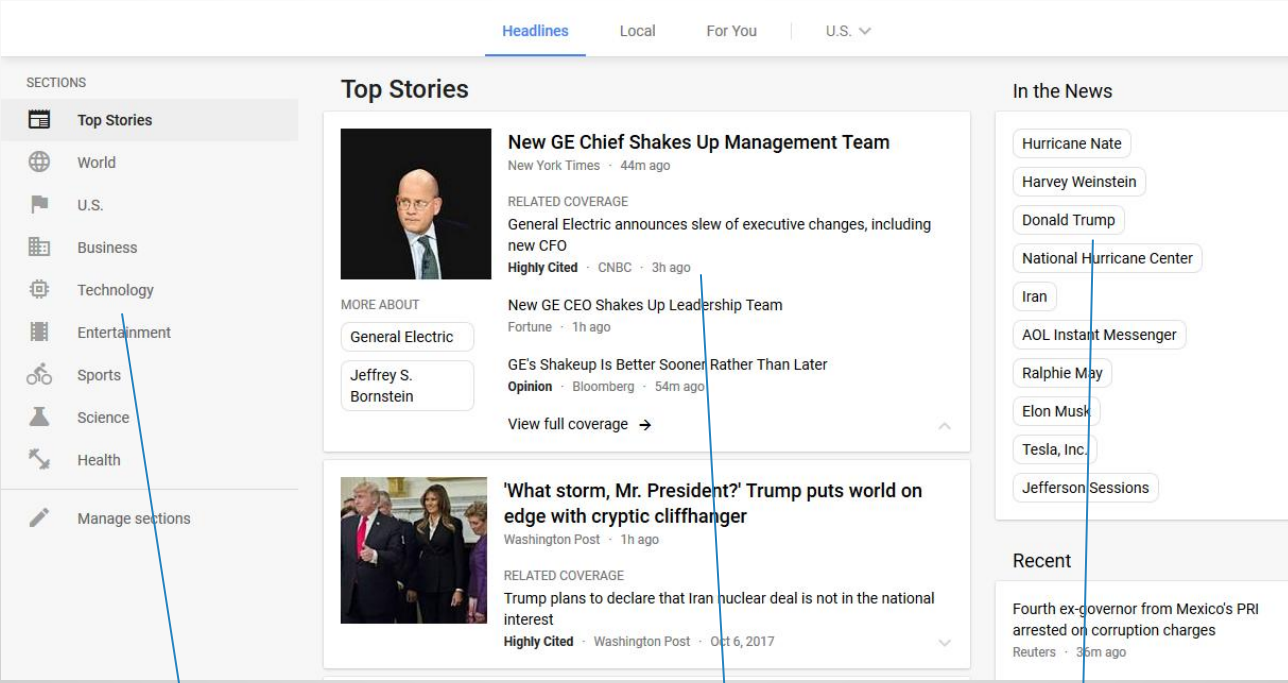
Baidu 百度 2000

SEARCH ENGINE



Source: slide share (Web Trainings)

NEWS AGGREGATOR



Text Classification

Summarization

Topic Detection

PLAGIARISM DETECTION

105.000.000

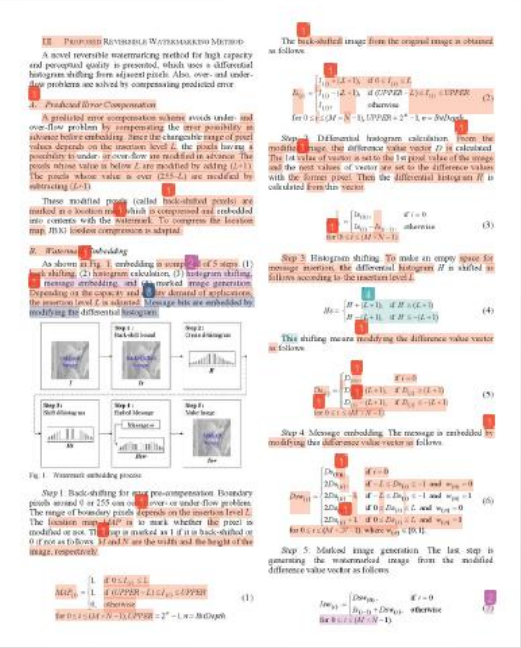
Scientific articles

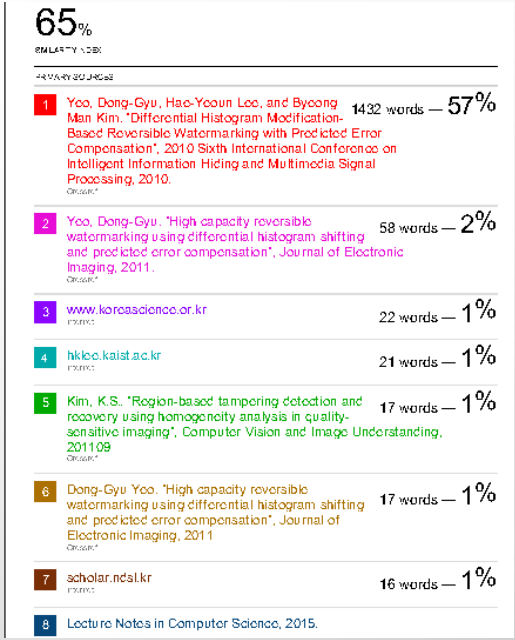
60.000.000.000

Web articles

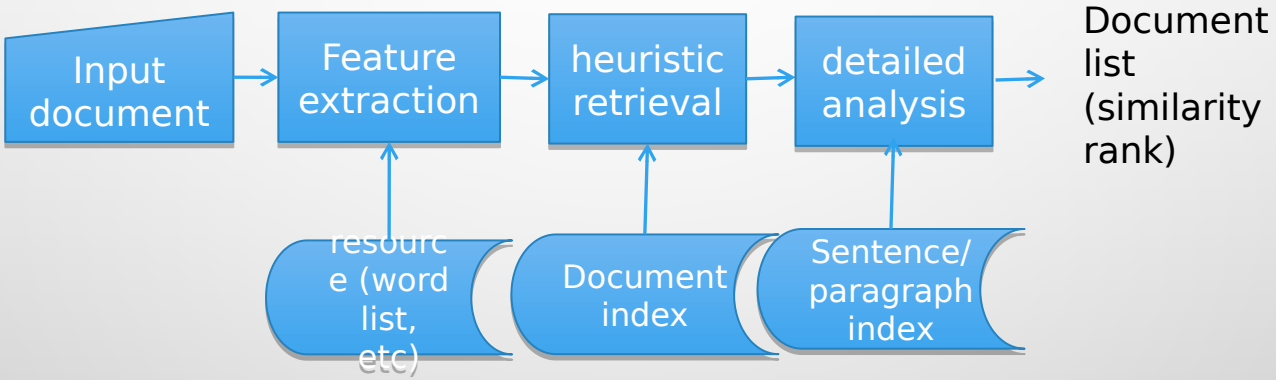
Source: http://www.ithenticate.com/

- ITHENTICATE
- CROSSCHECK
- TURNITIN





SYSTEM ARCHITECTURE



- COMPARE INPUT DOCUMENT WITH ALL DOCUMENTS IN THE COLLECTION (CORPUS)

SOCIAL MEDIA

MILLION OF USER GENERATED CONTENT

- CUSTOMER FEEDBACK
  - OPINION MINING
  - TRENDING TOPIC
  - USER MONITORING
- CUSTOMER SERVICE
  - CHATBOT
- FAKE INFORMATION
  - HOAX CLASSIFIER
- REAL INFORMATION
  - SUMMARIZE EXTRACTED INFORMATION

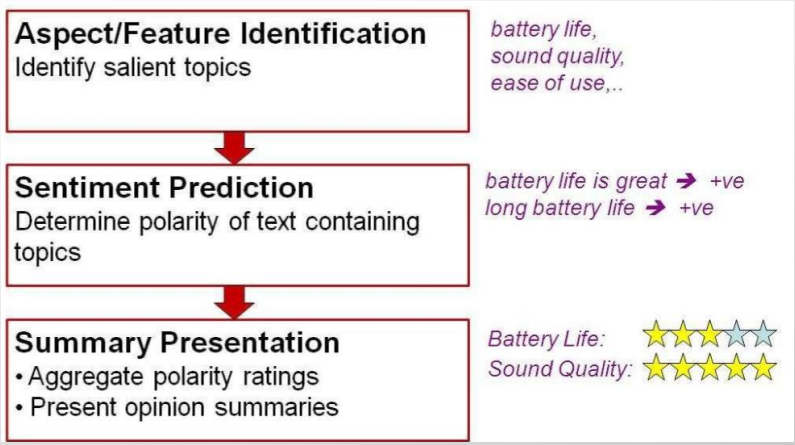
OPINION MINING

- SENTIMENT CLASSIFICATION

Sentiment analysis

Label: positive, negative, neutral

- ASPECT BASED



(1) *I bought a Samsung camera and my friends brought a Canon camera yesterday.* (2) *In the past week, we both used the cameras a lot.* (3) *The photos from my Samy are not that great, and the battery life is short too.* (4) *My friend was very happy with his camera and loves its picture quality.* (5) *I want a camera that can take good photos.* (6) *I am going to return it tomorrow.*

- Entity: samsung, samy, canon. Samsung and Samy are grouped together
- Aspect: picture, photo, battery life. Picture & photo are grouped for the camera
- Holder: the blog author (sentence 3), bigJohn's friend (sentence 4)
- Time; Sept-15-2011
- Sentiment: negative on picture quality, negative on battery life, position on camera as a whole

PROBLEMS IN OPINION MINING

- NE RECOGNITION
  - BARACK OBAMA BERHASIL MENGURANGI …
- ANAPHORA RESOLUTION
  - KAMI BERDUA MENONTON FILM "…". ITU BENAR-BENAR PENGALAMAN BURUK
- SARCASM (IRONI)
  - PEMERINTAH MAKIN HEBAT AJA, NGAMBIL MENTRI DARI PARTAI YG TERKENAL KORUPSI
- NON-FORMAL LANGUAGE
- FINDING MAIN OBJECT AND SENTIMENT ASPECT
  - MENTARI SINYALNYA MAKIN JELEK! MENDING PAKE XL, LEBIH BAGUS
- RELATIVE SENTIMENT
  - HARGA MAHASISWA
- ONE INPUT WITH BOTH POSITIVE AND NEGATIVE SENTIMENT
- COMPARISON
  - XL LEBIH MURAH DARI TELKOMSEL TAPI LEBIH SERING PUTUS-PUTUS

Germany

Brazil

twitter

Indonesia

Nigeria

Events on Twitter not in the News

Aditya Danang R.N #8
@Daraditya_
Follow

Besok aliansi BEM se-Indonesia demo di Istana Negara.. @BemFPsiUHT min gak ada aksi??

View translation

10:50 AM - 20 May 2015

Organizing strikes

Aksi demo mhswa/massa yg besar, masif dan maraton selalu diback up oleh isu, dana, backing yg kuat. Jika tdk, ya pasti tdk efektif

View translation

RETWEETS
10

12:51 AM - 20 May 2015

General Strike discussion

Threatening with Strike

PIDI 林忠龙
@da_vidicode

gelodok dan pecinan seluruh indonesia mogok kerja. kalo tuntutan kami g d jg harga kabel dan elektronik kami n gilaan

Translated from Indonesian by bing

ıt indonesia mogok. Kalo our demands g jg sup
ıan

Benhard Sitorus
@BenhardSitorus
Follow

Ayo Kita dukung aksi sosial #Viking dan #Bomber utk Pengalengan.. fb.me/7b3AdGJBq

View translation

4:51 PM -

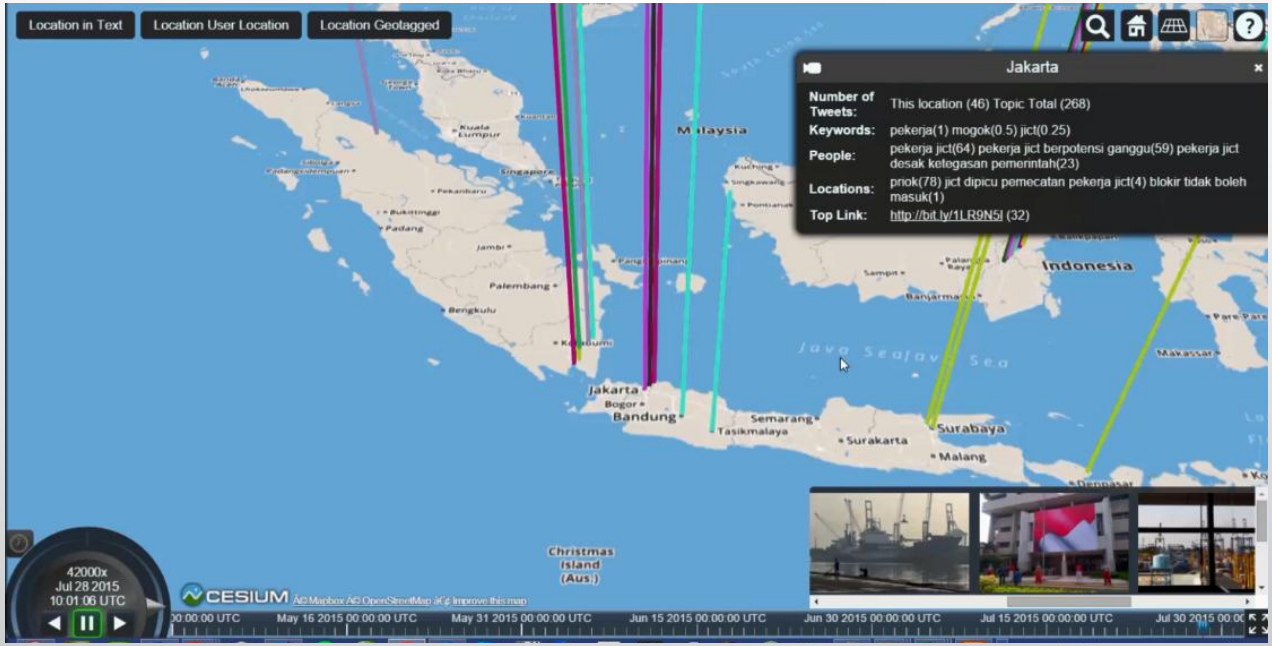
Asking for solidarity

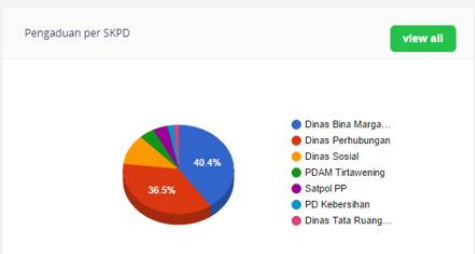
Info Lalulintas
@maslalin
Follow

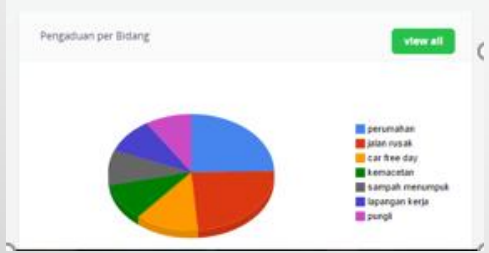
@PuspitaFM demo aksi donasi untuk muslim rohingya di veteran

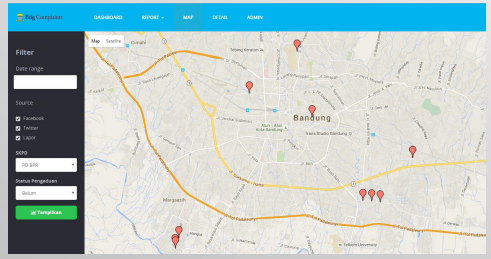
View translation

Asking for Donation


Location in Text
Location User Location
Location Geotagged
Jakarta
Number of Tweets: This location (46) Topic Total (268)
Keywords: pekerja(1) mogok(0.5) jict(0.25)
People: pekerja jict(64) pekerja jict berpotensi ganggu(59) pekerja jict desak ketegasan pemerintah(23)
Locations: priok(78) jict dipicu pemecatan pekerja jict(4) blokir tidak boleh masuk(1)
Top Link: http://bit.ly/1LR9N5l (32)
Malaysia
Indonesia
Kuala Lumpur
Singapore
Pekanbaru
Bukittinggi
Padang
Jambi
Palembang
Bengkulu
Jakarta
Bogor
Bandung
Semarang
Tasikmalaya
Surakarta
Surabaya
Malang
Banjarmasin
Makassar
Denpasar
Java Sea
Christmas Island (Aus.)
42000x
Jul 28 2015
10:01:06 UTC
CESIUM
May 16 2015 00:00:00 UTC
May 31 2015 00:00:00 UTC
Jun 15 2015 00:00:00 UTC
Jun 30 2015 00:00:00 UTC
Jul 15 2015 00:00:00 UTC







COMPLAINT MANAGEMENT SYSTEM

Complaint Classification	Doc classification: complaint vs response; nama dinas
Topic Classification	Doc classification
Information Extraction	Lokasi, waktu, kondisi, penyebab

CLASSIFICATION: TEXT/DOCUMENT VS SENTENCE VS TOKEN

- TEXT/DOCUMENT CLASSIFICATION: ENTIRE DOCUMENT AS INPUT FEATURE
- SENTENCE CLASSIFICATION:
  - SENTENCE SPLITTER
  - RELATION AMONG SENTENCE ✉ REFERENCE RESOLUTION, SENTENCE POSITION
  - SENTENCE AS INPUT FEATURE
- TOKEN CLASSIFICATION:
  - SENTENCE SPLITTER
  - TOKENIZATION
  - TOKEN/KATA/TERM AS INPUT FEATURE

EXAMPLE ON TOKEN CLASSIFICATION

- POS (PART OF SPEECH) TAGGER
  - CLASS: POS TAG, EX: NN (NOUN), VB (VERB), ADJ (ADJECTIVE), ADV (ADVERB)
- NAMED ENTITY TAGGER
  - CLASS: NE TAG, EX: PERSON, ORGANIZATION, LOCATION
- KEYWORD EXTRACTION
  - CLASS: YES/NO

INFORMATION EXTRACTION (NER + RELATION EXT)

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID:
<56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
rea: Voice Mail
req_years_experience: 2
desired_years_experience: 5

For years, Microsoft Corporation CEO Bill Gates was against open source.
But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation

Select Name
From PEOPLE
Where Organization = 'Microsoft'

PEOPLE

Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Bill Gates
Bill Veghte

IE=SEGMENTATION+CLASSIFICATION+ASSOCIATION+CLUSTERING

For years, Microsoft Corporation CEO Bill Gates was against open source.
But today Microsoft appears to have changed his mind. Gates himself says Microsoft will gladly disclose its crown jewels – the coveted code behind the Windows operating system – to select customers.
"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

MICROSOFT CORPORATION
CEO
BILL GATES
MICROSOFT

GATES
MICROSOFT

BILL VEGHTE
MICROSOFT
VP

RICHARD STALLMAN
FOUNDER
FREE SOFTWARE FOUNDATION

INFORMATION EXTRACTION

- IE
  - SEGMENTATION
  - CLASSIFICATION
  - ASSOCIATION
  - CLUSTERING

Named Entity Recognition

Relation Extraction

NAMED ENTITY RECOGNITION

- TYPES:
  - SUPERVISED
  - UNSUPERVISED
- FEATURES:
  - WORD LEVEL FEATURES
  - LIST LOOKUP FEATURES
  - DOCUMENT & CORPUS FEATURES
- SENTENCE: $S= W_1 W_2 \ldots W_{N-1} W_N$
- NAMED ENTITY CLASSIFICATION:
  - SEGMENTATION + CLASSIFICATION
  - RAW SEQUENCE LABELING
    - CLASS: NAMED ENTITY TYPE + “BEGIN/IN/END” INFORMATION
    - W_1 , WORD BEFORE, WORD AFTER, CLASS

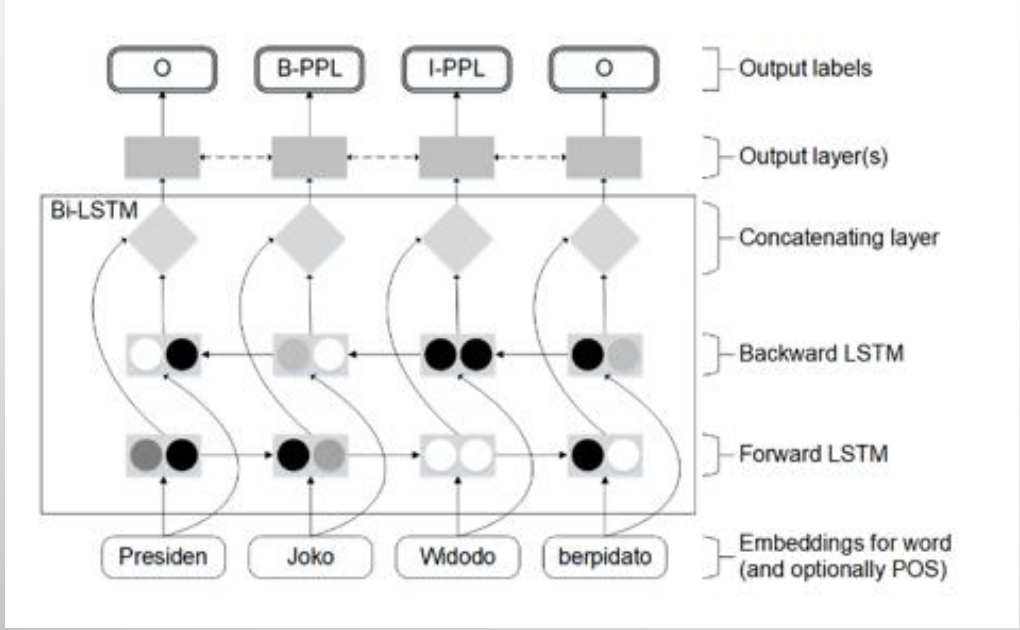
RELATION EXTRACTION

- SENTENCE: S= W_1 W_2 … E_1 … W_I .. E_2 .. W_{N-1} W_N
- RELATION CLASSIFICATION:
  - +1 IF E_1 AND E_2 ARE RELATED BY A RELATION R
    - R BISA BERUPA NAMA RELASI ATAU HANYA MENYATAKAN ADA RELASI ATAU TIDAK
  - -1 OTHERWISE
- FEATURE EXAMPLE:
  - E_1, E_2, WORD BETWEEN, WORD BEFORE, WORD AFTER, RELATION CLASS

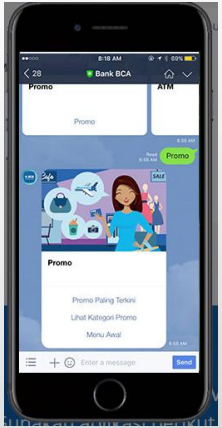
NER WITH MACHINE LEARNING

- FEATURE:
  - IMPORTANT:
    - CURRENT WORD
    - PRECEDING NE TAG
    - POS TAG
  - OPTIONAL:
    - WORD WINDOW: PRECEDING WORDS, SUCCEEDING WORDS
    - WORD LIST ✉ CAN BE USED AS SINGLE FEATURE OR TO LIMIT FEATURE (WORD VOCABULARY)
- CLASS:
  - NE TAGS: PERSON-B, PERSON-I, ORG-B, ORG-I, OTHER, ETC

NER WITH DEEP LEARNING

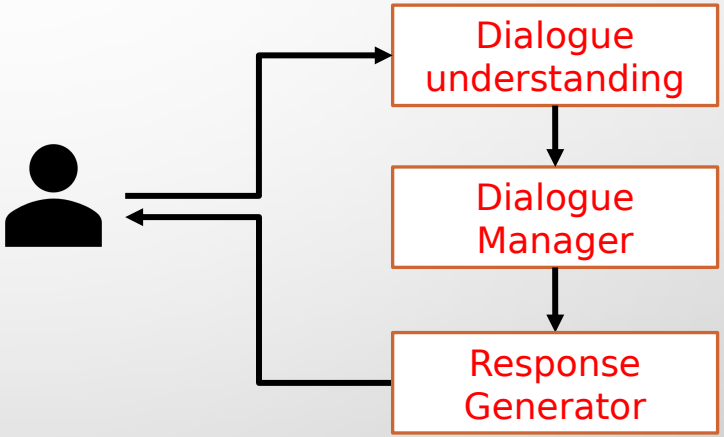


CHATBOT & VIRTUAL ASSISTANT



CHATBOT

- DIALOGUE UNDERSTANDING
  - INTENTION CLASSIFICATION
  - ENTITY EXTRACTION
- DIALOGUE MANAGER
  - EXECUTE DIALOGUE SCENARIO
- RESPONSE GENERATOR
  - TEMPLATE / GENERATOR



CHALLENGES

COMPLETE LANGUAGE PROBLEMS

LOW RESOURCE

DEEP LEARNING

TERIMA KASIH