




Introduction to Natural Language Processing with Python



Radical Rakhman Wahid

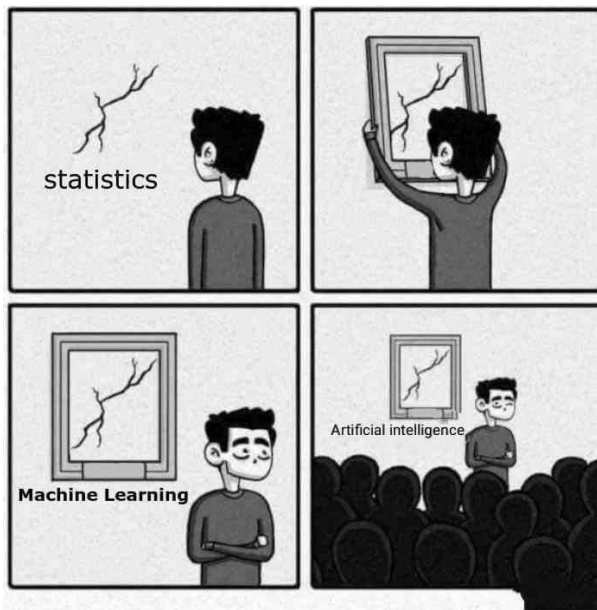
 fb.com/radical.rakhman

 t.me/rakhmanWahid

 linkedin.com/in/rakhid16



Bidang Minat :
Kecerdasan Buatan
Pembelajaran Mesin
Sains Data



KECERDASAN BUATAN

≠

BANYAK *IF-ELSE*



MACHINE LEARNING

=

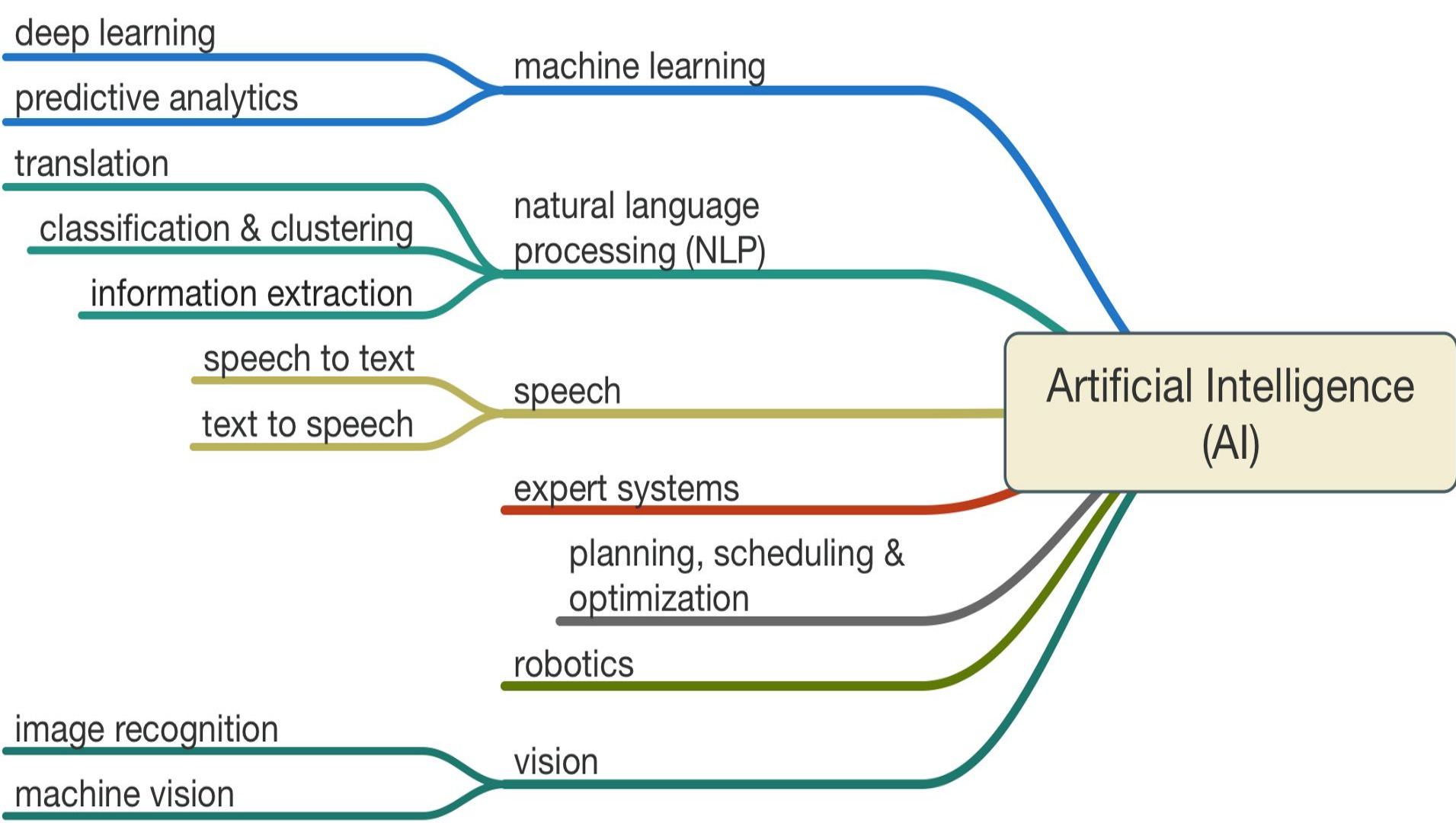
STATISTIKA &
PROBABILITAS,
KALKULUS, ALJABAR
LINEAR & MATRIKS



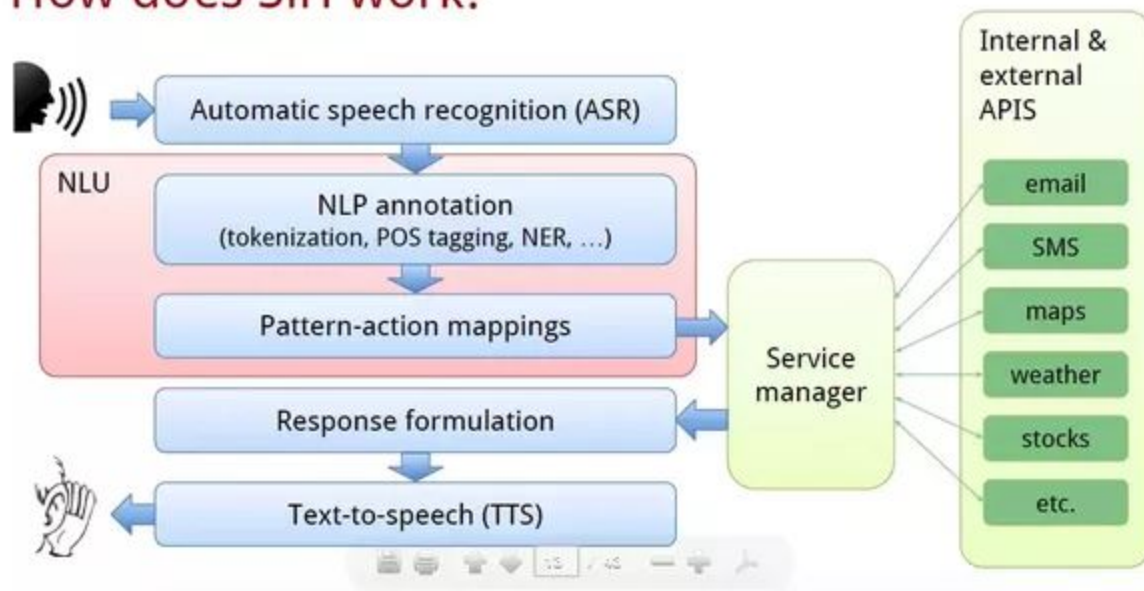
DEEP LEARNING

=

(lebih dari) JARINGAN
SYARAF TIRUAN (biasa)



How does Siri work?



COME AND DO NLP



JOIN THE DARK SIDE
memegenerator.net

Cabang dari **kecerdasan buatan** yang menggabungkan ilmu **linguistik** dan ilmu **komputer** yang bertujuan untuk melakukan **pemodelan komputasi** dari bahasa, sehingga dapat terjadi **interaksi** antara manusia dan mesin dengan perantara **bahasa alami**.

Sebelum melangkah lebih jauh...

- Fonologi ? Berhubungan dengan **bunyi suara** yang menghasilkan kata yang dapat dikenali.
- Morfologi ? Pengetahuan tentang **kata** dan **bentuknya** sehingga bisa dibedakan antara yang satu dengan yang lain.
- Sintaksis ? Pengetahuan tentang **urutan/tata/susunan/ kata** dalam pembentukan kalimat.
- Semantik ? Memelajari gabungan dari kata yang membentuk **arti(makna)** dari suatu kalimat.
- Pragmatik ? Pengetahuan tentang **konteks** kata/kalimat yang berhubungan erat dengan **situasi pemakaian** kalimat tersebut.
- Discourse Knowledge* ? Melakukan pengenalan apakah suatu kalimat yang telah dikenali mempengaruhi kalimat selanjutnya(**hubungan antar kalimat**).
- Word Knowledge* ? Mencakup **arti** kata **secara umum** dan apakah ada **arti khusus** bagi kata tersebut dalam percakapan dengan konteks tertentu

Analisa Semantik

Rule 1 : **IF** determiner adalah bagian pertama dalam kalimat dan diikuti oleh noun **THEN** noun tersebut dianggap sebagai subjek

Rule 2 : **IF** verb diikuti subjek **THEN** verb menjelaskan tentang apa yang dikerjakan oleh subjek

Rule 3 : **IF** noun diikuti subjek dan verb **THEN** noun tersebut dianggap sebagai objek

Rule 4 : **IF** kalimat mempunyai bentuk subjek, verb, objek **THEN** subjek mengerjakan verb yang ada hubungannya dengan objek

A plane flew home

Penerapan NLP?

TEXT MINING

SOCIAL NETWORK ANALYSIS (SNA) DALAM MELIHAT SENTIMEN PENGGUNA TWITTER TERHADAP PENDIDIKAN SEKOLAH MENENGAH KEJURUAN (SMK) DI INDONESIA MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM)



LATAR BELAKANG

Pendidikan kejuruan sebagai salah satu sub sistem dalam sistem pendidikan nasional diharapkan mampu mempersiapkan dan mengembangkan SDM yang bisa bekerja secara profesional di bidangnya, sekaligus bersaing dalam dunia kerja. Namun, data Badan Pusat Statistik (BPS) menunjukkan proporsi pengangguran terbesar adalah lulusan Sekolah Menengah Kejuruan (SMK) sebesar 9,84 persen. Melihat kondisi tersebut, Presiden Jokowi menginstruksikan perubahan sistem pendidikan kejuruan.

Para pelajar pada jenjang SMK tidak lepas dari maraknya kenakalan remaja yang di masa sekarang ini sudah semakin membahayakan. Contoh kasus kenakalan dan kriminalitas tersebut antara lain kasus pembacokan siswa SMK terhadap gurunya di Tangerang (Detik, 2015), siswa SMK membunuh bayinya yang baru dilahirkan (Liputan6, 2018), dan dua siswa SMK menjadi pemeran dan penyebar video mesum (Kompas, 2018).

Media sosial menjadi tempat diskusi masyarakat, salah satunya adalah twitter. Pembahasan mengenai evaluasi pendidikan SMK menjadi topik bahasan yang menarik bagi masyarakat baik itu memberikan tanggapan positif ataupun negatif. Sehingga perlu dilakukan penelitian untuk mengetahui tanggapan masyarakat terhadap evaluasi SMK berdasarkan analisis sentimen di media sosial twitter.

METODE PENELITIAN

DATA DAN ANALISIS

Data diperoleh dari Twitter API (Application Programming Interface) yang diambil sebanyak 2000 tweet beropini mulai 16 Februari 2018 hingga 2 Maret 2018



SVM

PERBANDINGAN KETEPATAN KLASIFIKASI SVM

	Linear	RBF
Training	96%	96%
Testing	94%	90%
Sensitivity	96%	94%
Specificity	94%	90%
AUC	96%	94%



WORD CLOUD

Sentimen Positif: proyek, bina, tinjau, kembang, sekolah

Sentimen Negatif: skandal, viob, anak, pelajar

SARAN

- Untuk Pemerintah: Pembinaan harus mereda ke seluruh SMK di Indonesia serta guru dan kompetensi.
- Untuk Sekolah: Lebih awal memberikan kegiatan yang mengasah aktif untuk mempersiapkan dunia kerja serta memberikan perhatian khusus kepada siswa agar terhindar dari kegiatan seks bebas.

TUJUAN PENELITIAN

- Membantu agar Pemerintah dapat dengan cepat mengetahui tanggapan masyarakat terhadap SMK sehingga dapat menentukan kebijakan lebih cepat.
- Topik apa saja yang dibicarakan oleh masyarakat terkait SMK, salah satu cara menggunakan Social Network Analysis.

Analisa Sentimen Klasifikasi Teks

Analisis Sentimen Masyarakat Terhadap Calon Presiden Indonesia 2019 Berdasarkan Opini Dari Twitter Menggunakan Metode Naive Bayes Classifier

LATAR BELAKANG

Pasal 1 ayat (2) Undang-Undang Dasar 1945 yang menentukan bahwa kedudukan berada di tangan rakyat dan dilaksanakan menurut Undang-Undang Dasar

Perkembangan teknologi komunikasi terus berkembang sehingga memberikan dampak kepada masyarakat

Pendapat/opini melalui tweet bisa digunakan untuk melihat bagaimana sentimen yang dimunculkan, salah satunya mengenai opini seseorang terhadap tokoh politik yang akan maju sebagai calon presiden Indonesia tahun 2019.

Pada penelitian Rozi, Imam F; Pramono Sholeh H; Dahlan, Achmad F. (2012) tentang Implementasi Opinion dilakukan analisis sentimen menggunakan metode *naive bayes* dalam penentuan polaritas sentimen

LANDASAN TEORI

Application Programming Interface

Naive Bayes Classifier

Twitter, Text mining, Analisis Sentimen Data mining, Rule Based

METODE PENELITIAN

Penelitian ini menggunakan opini masyarakat Indonesia dari media sosial Twitter dari bulan Maret 2018 Twitter sebelum deklarasi Calon Presiden 2019 pada tanggal 17-25 Mei 2019.

Data untuk penelitian adalah data primer, yang diambil dari Scraping media sosial Metode analisis data yang digunakan adalah metode Naive Bayes Classifier

TUJUAN PENELITIAN

- Untuk mengetahui klasifikasi pengguna Twitter terhadap Calon Presiden 2019 berdasarkan Periode.
- Untuk mengetahui perbandingan hasil sentimen terhadap Calon Presiden 2019.

ALUR PENELITIAN

Proses awal pada tahapan implementasi adalah pengumpulan dataset yang akan digunakan baik untuk *testing* maupun *training*. Pengumpulan dataset dengan cara memanfaatkan API dari pihak Twitter melalui teknik *streaming data*

Case Folding, Filtering, Tokenizing, Stemming

Preprocessing, Dictionary Build, Training, Testing, Evaluation

KESIMPULAN

- Dari output wordcloud dapat dijelaskan bahwa semakin besar huruf yang ditampilkan maka semakin banyak orang yang memberikan pendapat dengan kata tersebut. Output diatas adalah pendapat dari masyarakat terhadap Jokowi dan Prabowo.
- Persebaran jumlah status tentang Jokowi selama 1 minggu bergerak secara stasioner, status paling banyak terdapat pada tanggal 21-23 Mei 2018 dan tanggal 25-26 Mei 2018
- Periode waktu sebelum pemilu legislatif jumlah percakapan yang membicarakan tokoh Politik Joko Widodo dan Prabowo Subianto adalah sebesar 3778 dan 1828 tweet
- Sebelum pemilu legislatif persentase sentimen joko Widodo adalah 50,37 % sentimen positif dan 49,62% sentimen positif sedangkan Prabowo Subianto adalah 39,33% sentimen positif dan 60,66% sentimen negatif

Naive Bayes Classifier

Akurasi untuk Jokowi Widodo adalah 94,05% dan Prabowo adalah 91,53%

HASIL

wordcloud

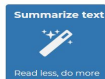
15611160@students.uii.ac.id

FREE SUMMARIZER

Summarize any text online in just a few seconds.



Stop wasting your time and money.



Free Summarizer is a Free service.

esummarizer.com/malay/summarize

TextSummarization Text Summarizer Online Text Summarization API

Ad closed by Google

Report this ad Why this ad?

Text Summarizer

Input the page url you want summarize:

URLs must start with 'http://' or 'https://', support English page summarization only

Or Copy and paste your text into the box:

Type the summarized sentence number you need:

6

Summarize Now

Automatic Text Summarizer

Start generating your online summary

Clear

Summarize

Perangkum Dokumen

Open Text Summarizer

This is a webinterface to the [Open Text Summarizer](#) tool. The tool automatically analyzes texts in various languages and tries to identify the most important parts of the text.

Just paste your text or load it from an [URL](#) to get it summarized.

Input

(or load from URL)

Output

☒ Summary ☐ Keywords

Summarization Ratio

☐ 5% ☐ 10% ☒ 20% ☐ 30% ☐ 40% ☐ 50% ☐ 60% ☐ 70% ☐ 80%

Language

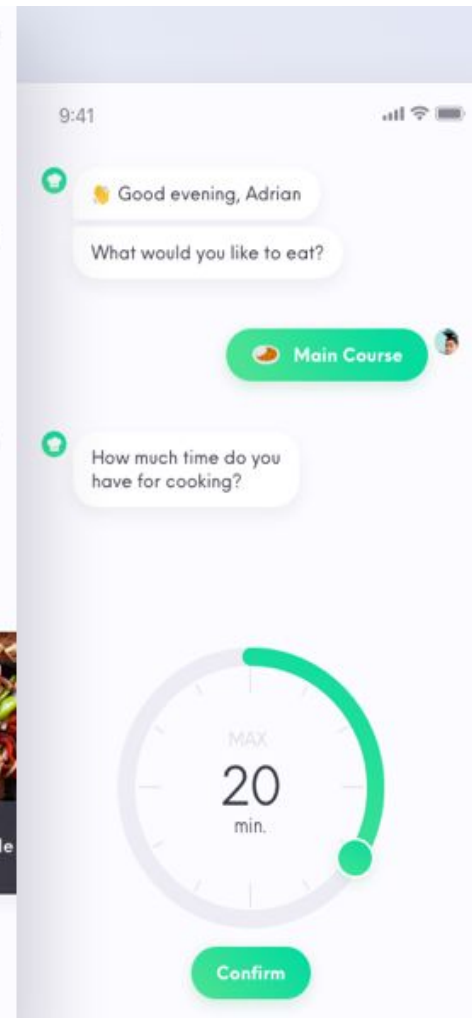
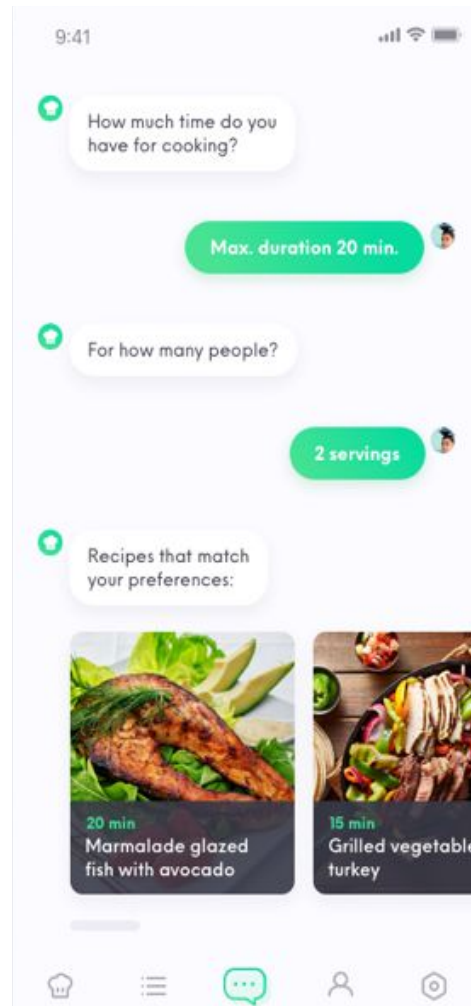
en

Submit

[Webinterface](#) for [Open Text Summarizer](#)



CHATBOT





+ Compose

📁 **Inbox** 19

★ Starred

🕒 Snoozed

🔔 Important

➤ Sent

📄 Drafts

▾ Categories

👤 **Social** 8

📌 **Updates** 8

💬 Forums

🏷️ **Promotions** 3

^ Less

💬 Chats

✉️ All Mail

⚠️ **Spam** 1

🗑️ Trash

⚙️ Manage labels

+ Create new label

☐ ↺ ⋮

1-2 of 2 < > ⚙️

Messages that have been in Spam more than 30 days will be automatically deleted. [Delete all spam messages now](#)

☐ ☆ ⋮ **Gabriel from DataCa.** **Weapons of Math Destruction** - Web Scraping with Python, Weather Data Analysis, Preparing a ... **Aug 7**

☐ ☆ ⋮ **Event Hunter Indone.** **Lomba Cerpen Nasional [HADIAH TRIP EROPA 5 NEGARA]** - webversion | unsubscribe | update p... **Aug 2**

Spam Detektor

0.04 GB (0%) of 15 GB used
[Manage](#)

[Terms](#) · [Privacy](#) · [Program Policies](#)

Last account activity: 13 hours ago
[Details](#)

Python Libraries for Natural Language Processing

Library	Outstanding Function/Feature	GitHub Stars (July 2018)
spaCy	Extremely optimized NLP library that is meant to be operated together with deep learning frameworks such as <i>TensorFlow</i> or <i>PyTorch</i> .	9924
Gensim	Highly efficient and scalable topic/semantic modelling.	7376
Pattern	Web (data) mining / crawling and common NLP tasks.	6387
NLTK	The 'mother' of all NLP libraries. Excellent for educational purposes and the de-facto standard for many NLP tasks.	6633
TextBlob	Modern multi-purpose NLP toolset that is really great for fast and easy development.	5295
Polyglot	Multilingualism and transliteration capabilities.	882
Vocabulary	Retrieve semantic information from individual words.	425
PyNLPI	Extensive functionality regarding FoLiA XML and many other common NLP format (CQL, Giza, Moses, ARPA, Timbl, etc.).	326
Stanford CoreNLP Python	Reliable, robust and accurate NLP platform based on a client-server architecture. Written in Java, and accessible through multiple Python wrapper libraries.	262
MontyLingua	End-to-end NLP processor working with Python and Java. Historical!	-

Algoritma *Term Frequency-Inverse Document Frequency*

Term Frequency :

TF ($word_i$) = banyaknya $word_i$ yang muncul pada satu dokumen

Inverse Document Frequency :

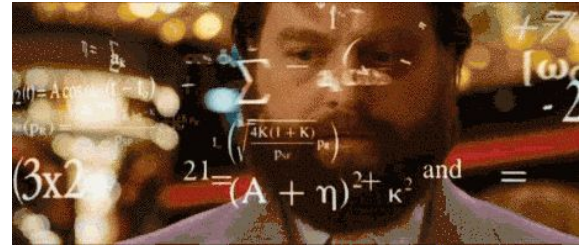
IDF ($word_i$) = \log (total dokumen/jumlah kata keseluruhan)

Term Importance:

$w(word_i) = \text{TF}(word_i) \times \text{IDF}(word_i)$

Word Normalization :

$$w(word_i) = \frac{w(word_i)}{\sqrt{w^2(word_1) + w^2(word_2) + \dots + w^2(word_n)}}$$



1. Ekstrak data

ID	Sentimen	Label
D1	penambangan data itu asik	pos
D2	nillai mata kuliah penambangan data ku jelek	neg
D3	aku tertarik dengan penambangan data	pos



Penambangan Data itu asik
Nillai mata kuliah Penambangan Data ku jelek
Aku tertarik dengan Penambangan Data

2. Hilangkan *stopwords*

Penambangan Data itu asik
Nillai mata kuliah Penambangan Data ku jelek
Aku tertarik dengan Penambangan Data



Penambangan Data itu asik
Nillai mata kuliah Penambangan Data ku jelek
Aku tertarik dengan Penambangan Data

A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia

Fadillah Z Tala

0086975



Master of Logic Project
Institute for Logic, Language and Computation
Universiteit van Amsterdam
The Netherlands

<https://pdfs.semanticscholar.org/8ed9/c7d54fd3f0b1ce3815b2eca82147b771ca8f.pdf>

758 Stopwords in
Bahasa Indonesia

3. Jadikan ke lower-case

Penambangan Data asik
Nilai mata kuliah Penambangan Data jelek
tertarik Penambangan Data



penambangan data asik
nillai mata kuliah penambangan data jelek
tertarik penambangan data

4. Stemming

pen (t)ambangan data asik
nillai mata kuliah pen (t)ambangan data jelek
tertarik pen (t)ambangan data



tambang data asik
nillai mata kuliah tambang data jelek
tarik tambang data

[Help](#)[Donate](#)[Log in](#)[Register](#)

Sastrawi 1.0.1

**Latest version**`pip install Sastrawi`*Last released: Jan 18, 2016*

Library for stemming Indonesian (Bahasa) text

Navigation

[Project description](#)[Release history](#)[Download files](#)

Project links

[Homepage](#)

Project description

Sastrawi is a simple Python library which allows you to reduce inflected words in Indonesian Language (Bahasa Indonesia) to their base form ([stem](#)).

This is Python port of the original [Sastrawi](#) project written in PHP.

build **passing** coverage unknown

Installation

Sastrawi can be installed via [pip](#), by running the following commands in terminal/command prompt : `pip`

```
install Sastrawi
```

Example Usage

5. Hitung frekuensi kata/dokumen

tambang data asik
nillai mata kuliah tambang data jelek
tarik tambang data



asik : 1, data : 1, tambang : 1
data : 1, jelek : 1, kuliah : 1, mata :1, nilai :1, tambang : 1
data : 1, tambang : 1, tarik : 1

6. Bikin *bag of words*

tambang data asik
nillai mata kuliah tambang data jelek
tarik tambang data

asik : 1, data : 1, tambang : 1
data : 1, jelek : 1, kuliah : 1, mata :1, nilai :1, tambang : 1
data : 1, tambang : 1, tarik : 1

ID	Kata	Jumlah
1	asik	1
2	data	3
3	jelek	1
4	kuliah	1
5	mata	1
6	nilai	1
7	tambang	3
8	tarik	1

7. Jadikan *Vector Space Model*

tambang data asik
nillai mata kuliah tambang data jelek
tarik tambang data
asik : 1, data : 1, tambang : 1
data : 1, jelek : 1, kuliah : 1, mata :1, nilai :1, tambang : 1
data : 1, tambang : 1, tarik : 1
1, 1, 0, 0, 0, 0, 1, 0
0, 1, 1, 1, 1, 1, 1, 0
0, 1, 0, 0, 0, 0, 1, 1

ID	Kata	Jumlah
1	asik	1
2	data	3
3	jelek	1
4	kuliah	1
5	mata	1
6	nilai	1
7	tambang	3
8	tarik	1

8. Hitung IDF

1, 1, 0, 0, 0, 0, 1, 0
0, 1, 1, 1, 1, 1, 1, 0
0, 1, 0, 0, 0, 0, 1, 1

Inverse Document Frequency :

$IDF(word_i) = \log(\text{total dokumen} / \text{jumlah kata keseluruhan})$

ID	Kata	Jumlah	IDF
1	asik	1	0.477
2	data	3	0
3	jelek	1	0.477
4	kuliah	1	0.477
5	mata	1	0.477
6	nilai	1	0.477
7	tambang	3	0
8	tarik	1	0.477

9. hitung bobot per kata

1, 1, 0, 0, 0, 0, 1, 0

0.477, 0, 0, 0, 0, 0, 0, 0

0, 1, 1, 1, 1, 1, 1, 0

0, 0, 0.477, 0.477, 0.477, 0.477, 0, 0

0, 1, 0, 0, 0, 0, 1, 1

0, 0, 0, 0, 0, 0, 0, 0.477

Term Frequency :

TF (*wordi*) = banyaknya *wordi* yang muncul pada satu dokumen

Term Importance:

$w(wordi) = TF(wordi) \times IDF(wordi)$

ID	Kata	Jumlah	IDF
1	asik	1	0.477
2	data	3	0
3	jelek	1	0.477
4	kuliah	1	0.477
5	mata	1	0.477
6	nilai	1	0.477
7	tambang	3	0
8	tarik	1	0.477

10. Normalisasi kata di tiap dokumen

1, 1, 0, 0, 0, 0, 1, 0

0.477, 0, 0, 0, 0, 0, 0, 0

0, 1, 1, 1, 1, 1, 1, 0

0, 0, 0.477, 0.477, 0.477, 0.477, 0, 0

0, 1, 0, 0, 0, 0, 1, 1

0, 0, 0, 0, 0, 0, 0, 0.477

1, 0, 0, 0, 0, 0, 0, 0

0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0

0, 0, 0, 0, 0, 0, 0, 1

Word Normalization :

$$w(word_i) = \frac{w(word_i)}{\sqrt{w^2(word_1) + w^2(word_2) + \dots + w^2(word_n)}}$$

$$w(asik) = \sqrt{\frac{0.477^2}{0.477^2 + 0 + 0 + 0 + 0 + 0 + 0 + 0}} = 1$$

Algoritma Naive Bayes

$$P(H | X) = \frac{P(X|H)P(H)}{P(X)}$$

$P(H | X)$ = Peluang hipotesis H jika diketahui data X

$P(X | H)$ = Peluang data X jika diketahui hipotesis H

$P(H)$ = Peluang hipotesis H

$P(X)$ = Peluang data X



1. Siapkan himpunan data

ID	Ulasan netijen	Label
1	Dont buy	neg
2	Phone got hanged	neg
3	Battery drains fast	neg
4	Durable phone	pos
5	Great camera	pos
6	Great phone buy it	?

2. Tentukan *train set* dan *test set*

ID	Ulasan netijen	Label
1	Dont buy	neg
2	Phone got hanged	neg
3	Battery drains fast	neg
4	Durable phone	pos
5	Great camera	pos

Train set

ID	Ulasan netijen	Label
6	Great phone buy it	?

Test set

3. B U A T T A B E L

Kata	pos	neg
DONT	0	1
BUY	0	1
PHONE	1	1
GOT	0	1
HANGED	0	1
BATTERY	0	1
DRAINS	0	1
FAST	0	1
DURABLE	1	0
GREAT	1	0
CAMERA	1	0



STATISTICS

DATA
SCIENCE



4. Hitung probabilitas *prior*

$P(\text{positif}) = \frac{2}{5}$ atau 0.4

$P(\text{negatif}) = \frac{3}{5}$ atau 0.6

ID	Ulasan netijen	Label
1	Dont buy	neg
2	Phone got hanged	neg
3	Battery drains fast	neg
4	Durable phone	pos
5	Great camera	pos

5. Hitung peluang kejadian bersyarat

$$P(\text{GREAT} | \text{positif}) = 1+1 / 4+11 = 2/15 = 0.13$$

$$P(\text{PHONE} | \text{positif}) = 1+1 / 4+11 = 2/15 = 0.13$$

$$P(\text{BUY} | \text{positif}) = 0+1 / 4+11 = 1/15 = 0.07$$

$$P(\text{IT} | \text{positif}) = 0+1 / 4+11 = 1/15 = 0.07$$

$$P(\text{GREAT} | \text{negatif}) = 0+1 / 8+11 = 1/19 = 0.05$$

$$P(\text{PHONE} | \text{negatif}) = 1+1 / 8+11 = 2/19 = 0.11$$

$$P(\text{BUY} | \text{negatif}) = 1+1 / 8+11 = 2/19 = 0.11$$

$$P(\text{IT} | \text{negatif}) = 0+1 / 8+11 = 1/19 = 0.05$$

6. Hitung peluang posterior

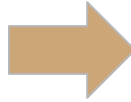
$$\begin{aligned}P(\text{positif}) &= P(\text{GREAT} | \text{positif})P(\text{PHONE} | \text{positif})P(\text{BUY} | \text{positif})P(\text{IT} | \text{positif})P(\text{positif}) \\&= 0.13 \times 0.13 \times 0.7 \times 0.7 \times 0.4 \\&= 0.0033124\end{aligned}$$

$$\begin{aligned}P(\text{negatif}) &= P(\text{GREAT} | \text{negatif})P(\text{PHONE} | \text{negatif})P(\text{BUY} | \text{negatif})P(\text{IT} | \text{negatif})P(\text{negatif}) \\&= 0.5 \times 0.11 \times 0.11 \times 0.05 \times 0.6 \\&= 0.0018149\end{aligned}$$

7. Hasil akhir

Peluang yang mendekati 1 lah yang diambil

ID	Ulasan netijen	Label
6	Great phone buy it	?



ID	Ulasan netijen	Label
6	Great phone buy it	pos

Before we start to build our text classifier :

1. *install python (windows only)* python.org
2. *install pip* pip.pypa.io/en/stable/installing/
3. *install textblob*
 - buka *command prompt* (windows) atau *terminal* (ubuntu & mac)
 - ketik **pip install TextBlob** atau **pip install -U TextBlob** terus *enter*
 - ketik **python -m textblob.download_corpora** terus *enter*
4. *install jupyter notebook (opsional)*
 - ketik **pip install jupyter notebook** terus *enter*

**THANKS FOR
YOUR ATTENTION**

**CLAP AND DON'T ASK
ANY QUESTION, PLEASE**