Natural Language Processing

Ali Akbar Septiandri

December 29, 2020

Universitas Al Azhar Indonesia

Daftar isi

Natural Language Processing

Apa Itu NLP?

Salah satu ilmu multidisiplin yang berfokus pada interaksi manusia dan komputer melalui bahasa alami manusia. Beberapa hal yang dibahas di dalamnya antara lain:

- Part-of-Speech (POS) tagging
- Parsing
- Stemming
- Machine translation
- Named entity recognition (NER)
- Question answering
- Sentiment analysis
- Automatic summarisation
- Speech recognition
- Text-to-speech

Kategori Tugas-tugas NLP

- Syntax
  - Part-of-Speech (POS) tagging
  - Parsing
  - Stemming
- Semantics
  - Machine translation
  - Named entity recognition (NER)
  - Question answering
  - Sentiment analysis
- Discourse
  - Automatic summarisation
- Speech
  - Speech recognition
  - Text-to-speech

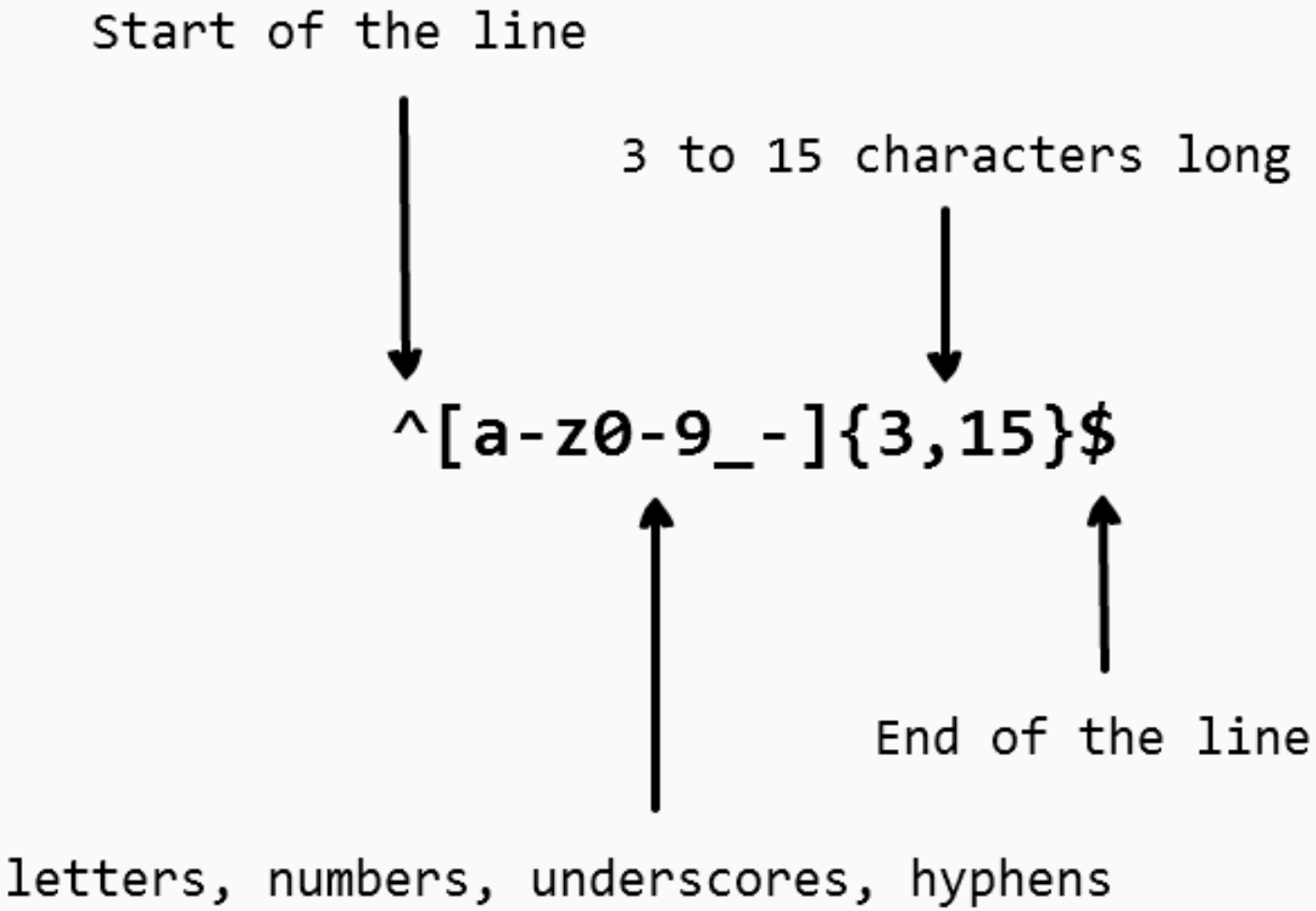


Figure 1: Contoh regular expression. Sumber: tajawal

Part-of-Speech (POS) Tagging

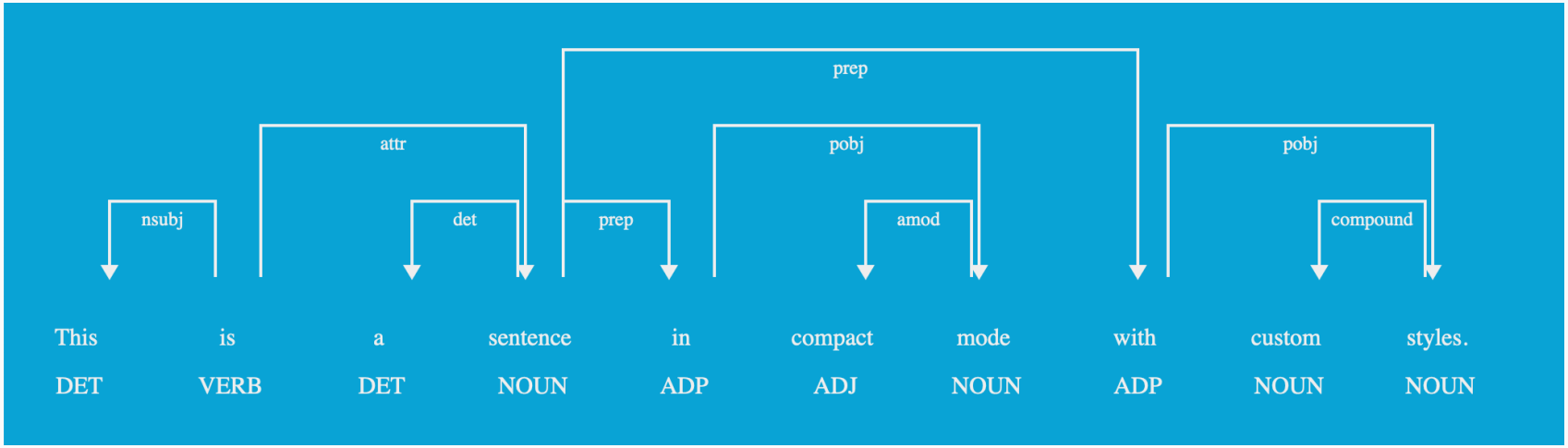


Figure 2: Kategori kata (Sumber: spaCy)

Named Entity Recognition (NER)

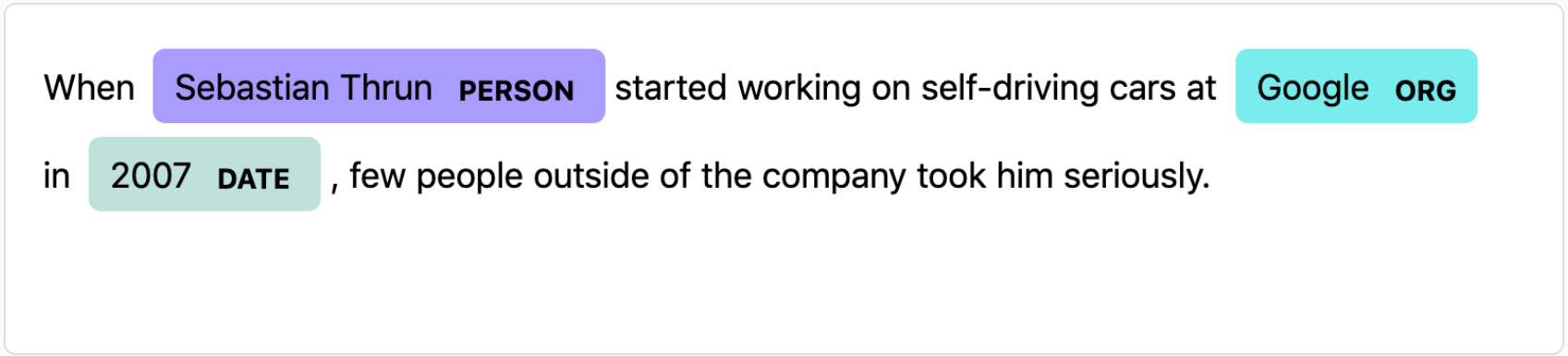


Figure 3: Menemukan orang, organisasi, dan tanggal dalam teks (Sumber: spaCy)

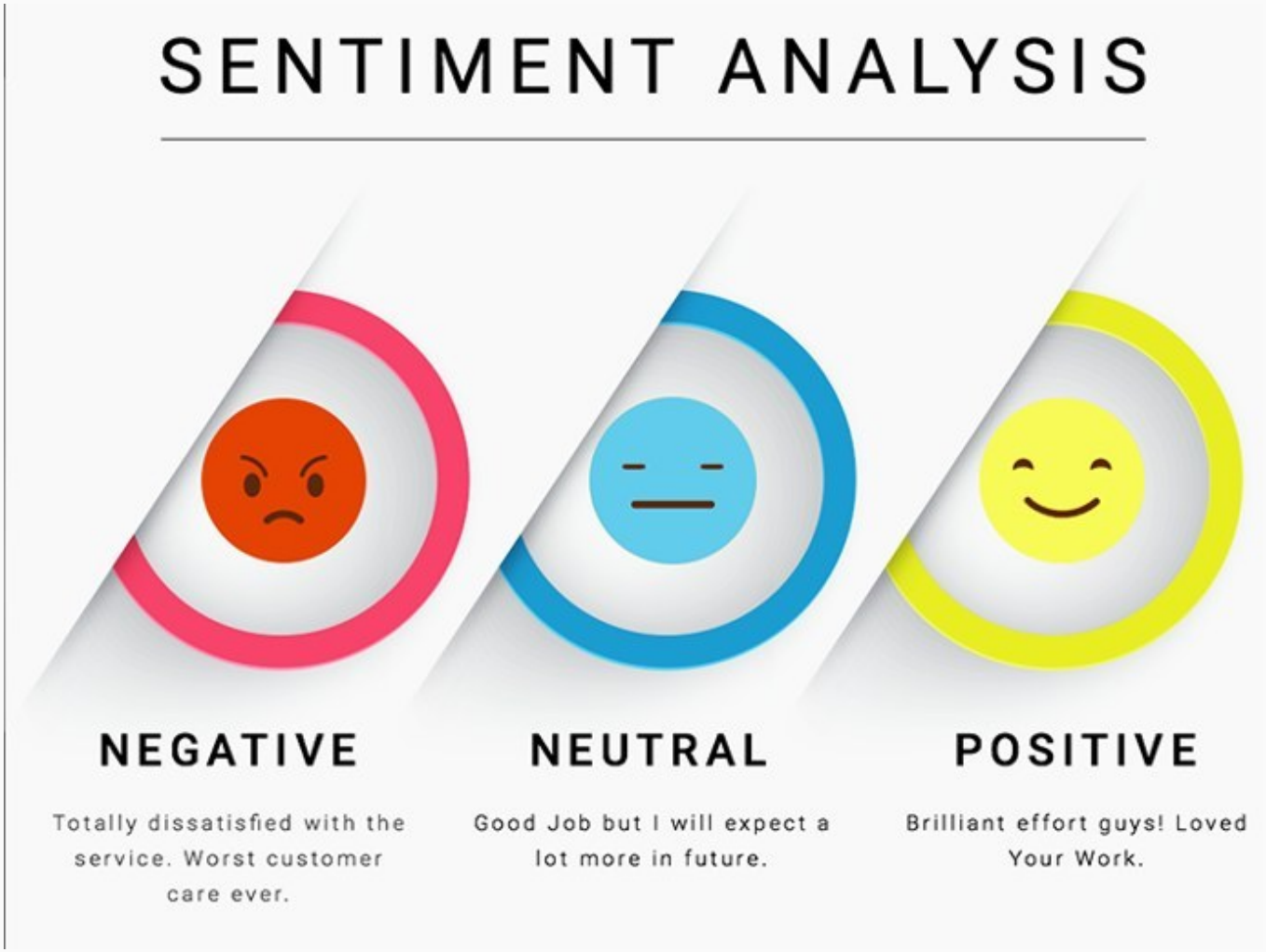
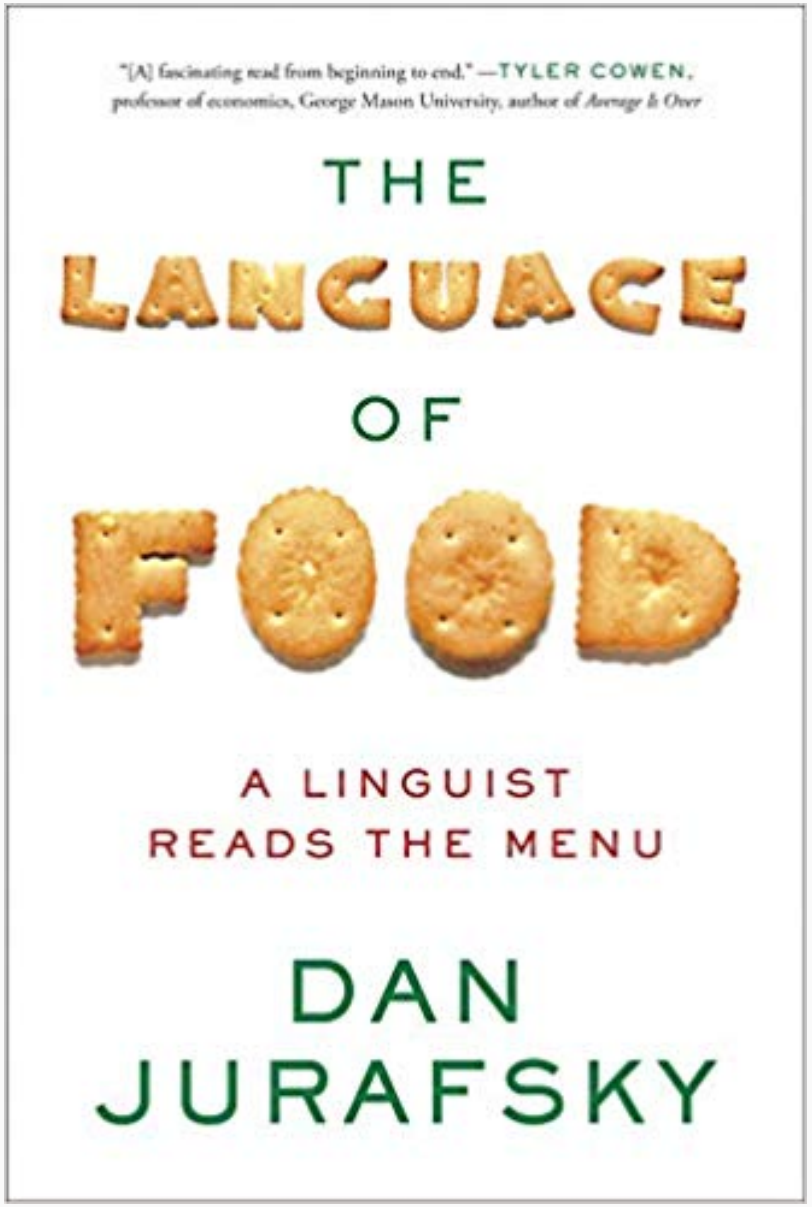


Figure 4: "Apa yang menjadi sentimen dari ulasan ini?" Sumber: KDNuggets


"[A] fascinating read from beginning to end." —TYLER COWEN,
professor of economics, George Mason University, author of Average Is Over
THE
LANGUAGE
OF
FOOD
A LINGUIST
READS THE MENU
DAN
JURAFSKY

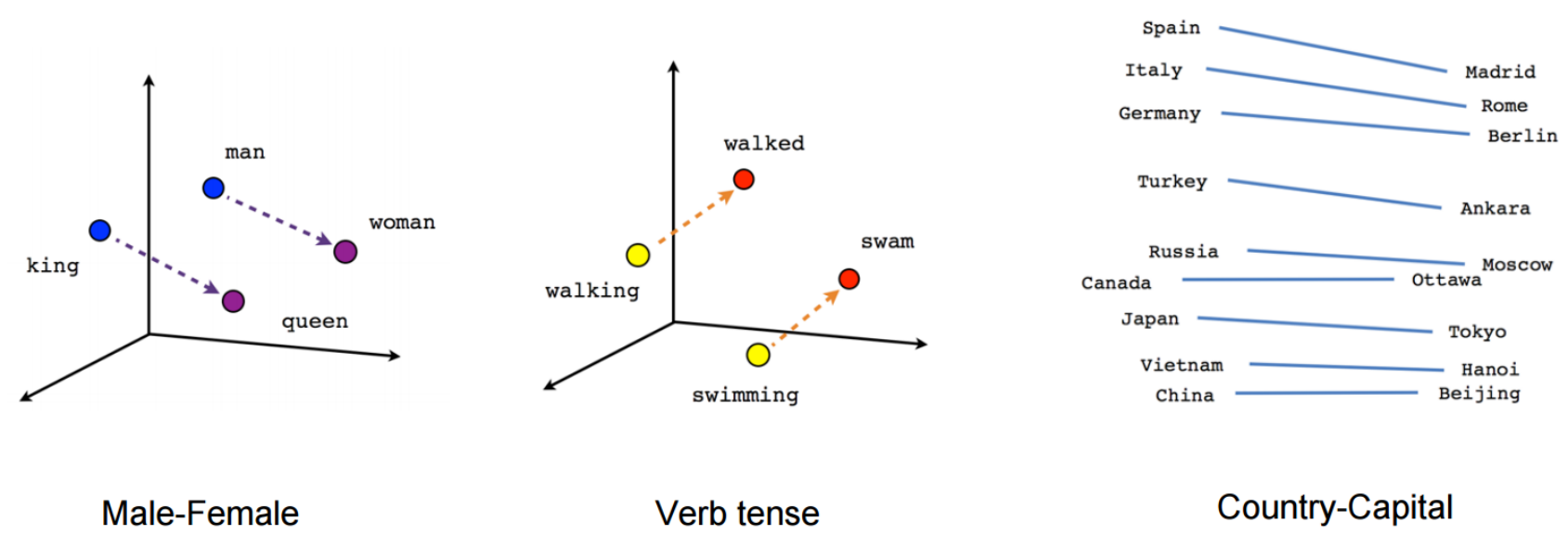


Figure 5: Representasi kata dalam vektor (Mikolov et al., 2013).
Sumber: TensorFlow

Demo word2vec

Sentiment Analysis

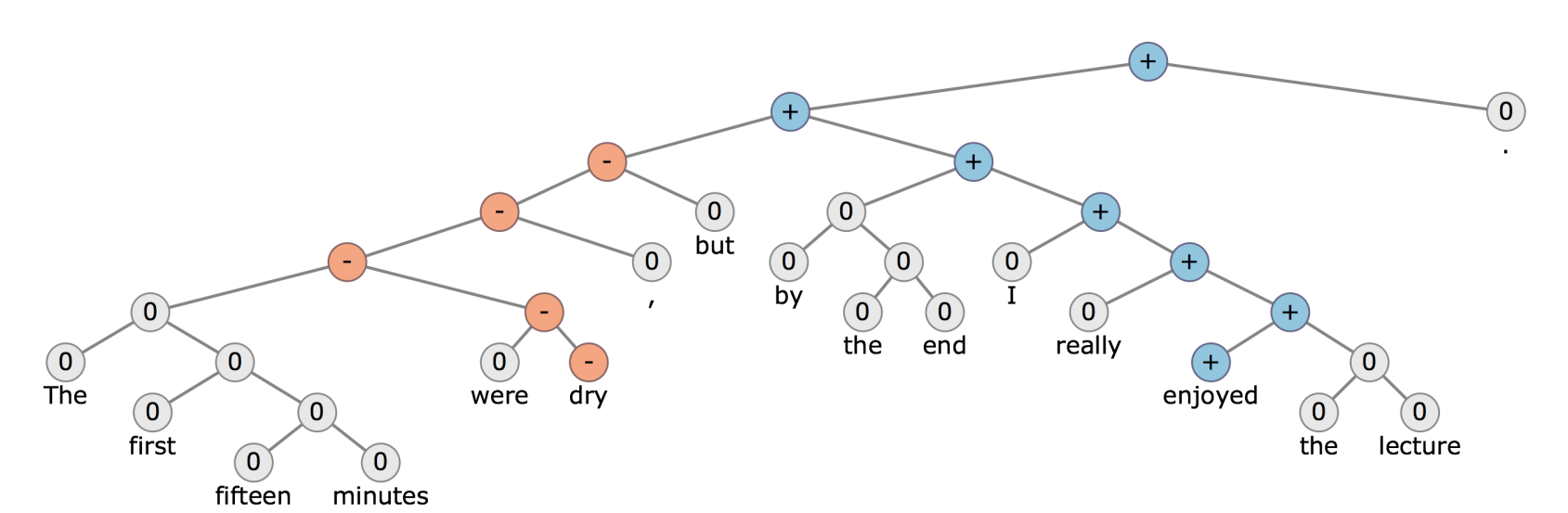


Figure 6: Hasil analisis sentimen dengan *deep learning* [Socher, 2017]

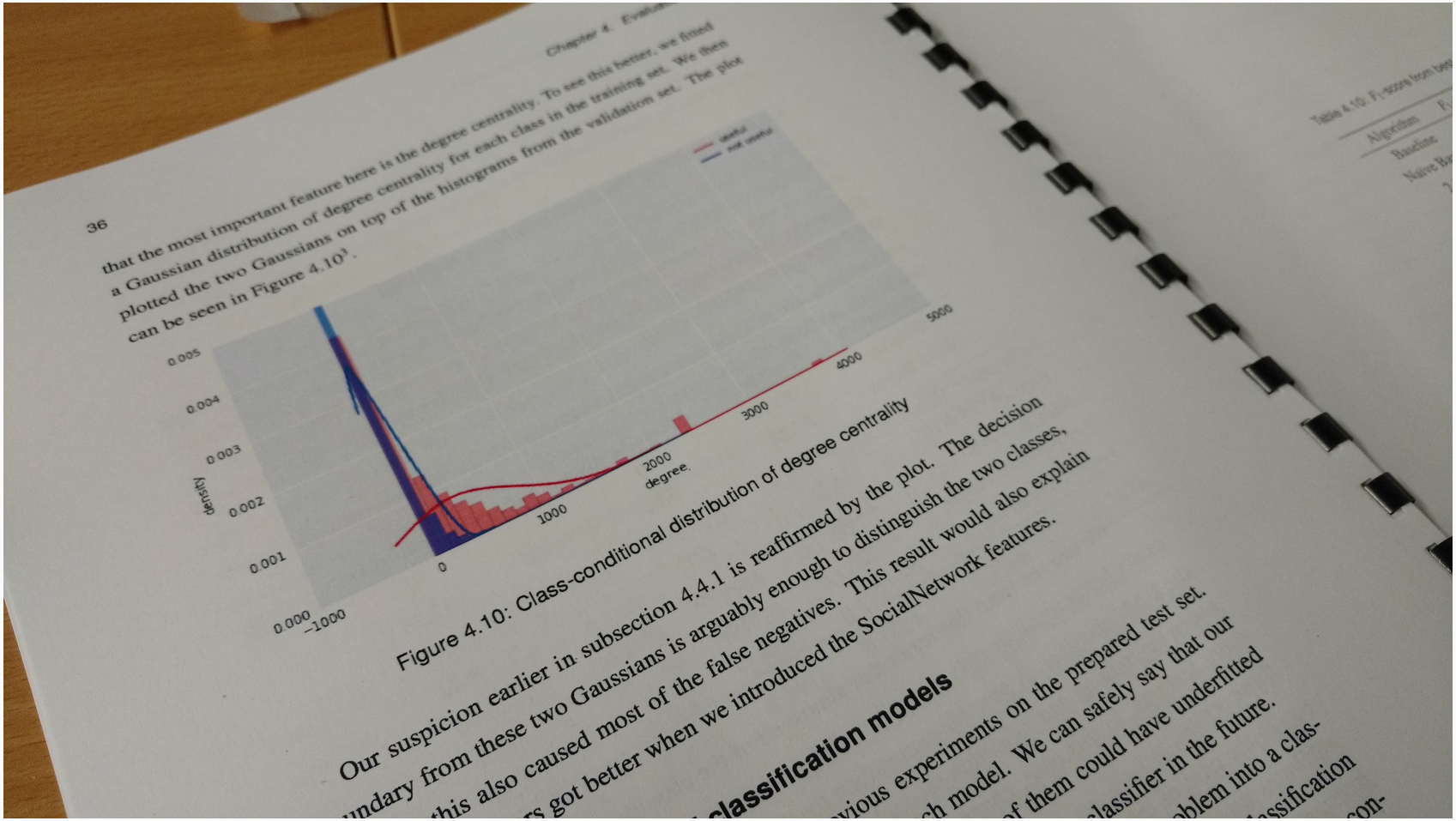


Figure 7: Deteksi plagiarisme dari makalah

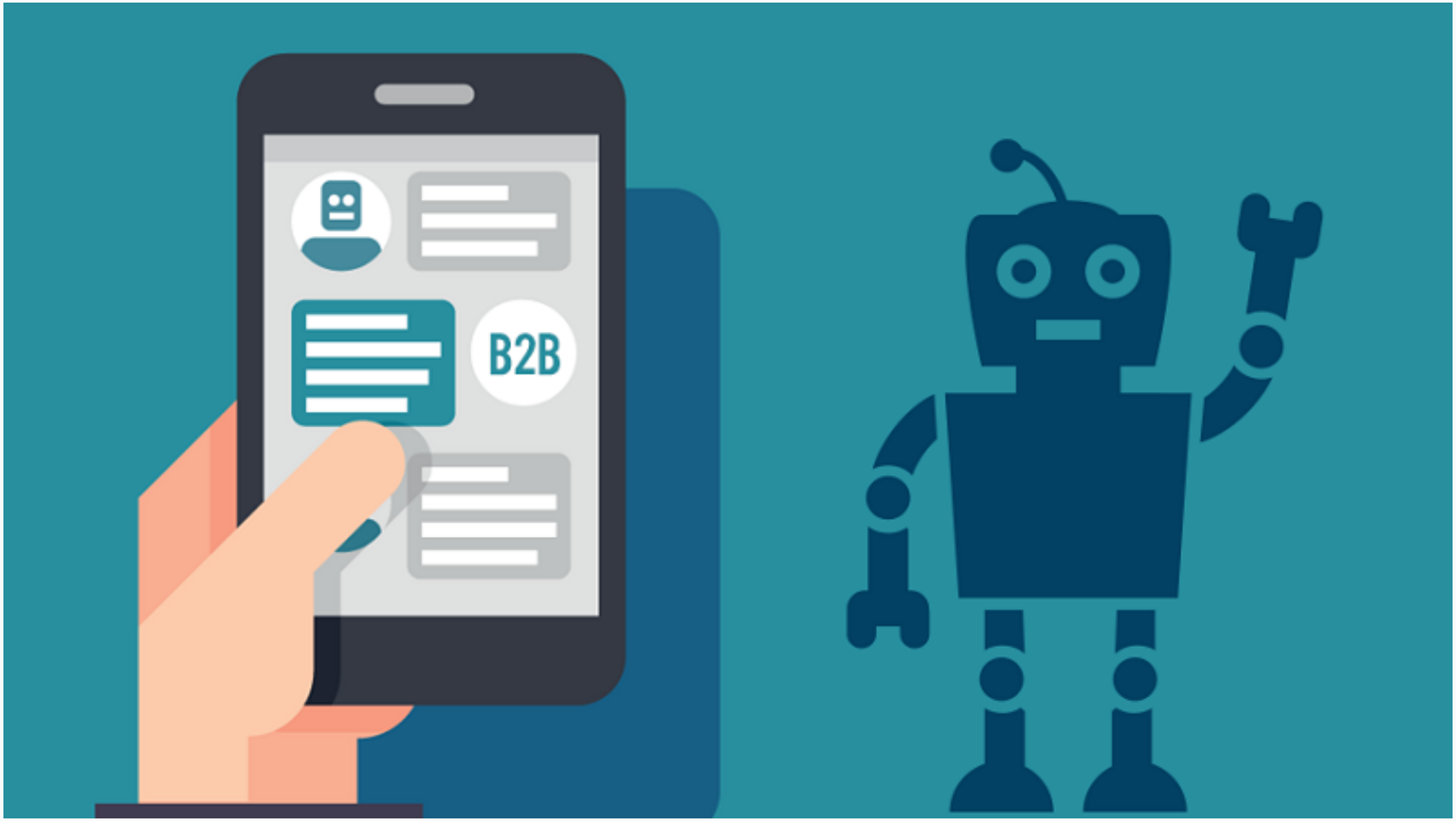


Figure 8: Penggunaan chatbot untuk bisnis. Sumber: Acquire

Computational Linguistics

- NLP juga dikenal dengan nama *computational linguistics*, karena mencoba merepresentasikan makna dari kata, frasa, kalimat, dan dokumen melalui **distribusinya**

Computational Linguistics

- NLP juga dikenal dengan nama *computational linguistics*, karena mencoba merepresentasikan makna dari kata, frasa, kalimat, dan dokumen melalui **distribusinya**
- Distribusi tersebut direpresentasikan dalam **vektor konteks**

Computational Linguistics

- NLP juga dikenal dengan nama *computational linguistics*, karena mencoba merepresentasikan makna dari kata, frasa, kalimat, dan dokumen melalui **distribusinya**
- Distribusi tersebut direpresentasikan dalam **vektor konteks**
- "Dalam suatu dokumen, kata apa saja yang muncul bersamaan?"

Computational Linguistics

- NLP juga dikenal dengan nama *computational linguistics*, karena mencoba merepresentasikan makna dari kata, frasa, kalimat, dan dokumen melalui **distribusinya**
- Distribusi tersebut direpresentasikan dalam **vektor konteks**
- "Dalam suatu dokumen, kata apa saja yang muncul bersamaan?"
- Begitu pula di level semantik $\rightarrow$ **Bag-of-Words (BoW) model**

Computational Linguistics

- NLP juga dikenal dengan nama *computational linguistics*, karena mencoba merepresentasikan makna dari kata, frasa, kalimat, dan dokumen melalui **distribusinya**
- Distribusi tersebut direpresentasikan dalam **vektor konteks**
- "Dalam suatu dokumen, kata apa saja yang muncul bersamaan?"
- Begitu pula di level semantik $\rightarrow$ **Bag-of-Words (BoW) model**
- Bahkan, bisa sampai ke level **karakter**!

Bag-of-Words

Dalam representasi ini, urutan atau letak dari kata tersebut tidak relevan

D1	“send us your password”
D2	“send us your review”
D3	“review your password”
D4	“review us”
D5	“send your password”
D6	“send us your account”

Binary Bag-of-Words

Dalam representasi ini, urutan atau letak dari kata tersebut tidak relevan

dokumen	account	password	review	send	us	your
D1	0	1	0	1	1	1
D2	0	0	1	1	1	1
D3	0	1	1	0	0	1
D4	0	0	1	0	1	0
D5	0	1	0	1	0	1
D6	1	0	0	1	1	1

Apa yang menjadi masalah di sini?

Isu

- Matriksnya jarang
- Tidak ada informasi urutan
- Ada kata-kata yang sangat sering muncul

TF-IDF

$$w_{t,d} = (1 + \log(tf_{t,d})) \log(\frac{N}{df_t})$$

- $tf_{t,d}$ ... frekuensi kata t dalam dokumen d, N ... jumlah dokumen, df_t ... jumlah dokumen yang mempunyai kata t
- Kata yang sering muncul mungkin tidak penting, e.g. kata hubung
- Kata yang langka akan bernilai lebih – lihat posisi df_t!

Menemukan Dokumen yang Mirip

- Euclidean distance adalah metode pengukuran jarak yang umum
- Untuk dokumen, jumlah kemunculan kata *mungkin* tidak begitu penting
- Yang penting adalah keberadaan katanya $\rightarrow$ cosine similarity

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

Memaknai Kata

- Dengan ide yang serupa dan beberapa tambahan algoritma lainnya, e.g. *Latent Semantic Analysis* (LSA), kita bisa menggunakan kakas ini untuk **tes seperti TOEFL**

Memaknai Kata

- Dengan ide yang serupa dan beberapa tambahan algoritma lainnya, e.g. *Latent Semantic Analysis* (LSA), kita bisa menggunakan kakas ini untuk **tes seperti TOEFL**
- LSA berhasil menjawab 64.4% soal

Memaknai Kata

- Dengan ide yang serupa dan beberapa tambahan algoritma lainnya, e.g. *Latent Semantic Analysis* (LSA), kita bisa menggunakan kakas ini untuk **tes seperti TOEFL**
- LSA berhasil menjawab 64.4% soal
- Pengguna bahasa Inggris non-natif rata-rata berhasil menjawab 64.5% soal

Memaknai Kata

- Dengan ide yang serupa dan beberapa tambahan algoritma lainnya, e.g. *Latent Semantic Analysis* (LSA), kita bisa menggunakan kakas ini untuk **tes seperti TOEFL**
- LSA berhasil menjawab 64.4% soal
- Pengguna bahasa Inggris non-natif rata-rata berhasil menjawab 64.5% soal
- *Cukup untuk masuk banyak universitas di US!*

Beberapa Tantangan

- Homograf, kata yang tulisannya sama tetapi maknanya berbeda, e.g. “beruang”

Beberapa Tantangan

- Homograf, kata yang tulisannya sama tetapi maknanya berbeda, e.g. "beruang"
- Token yang tidak dikenali, e.g. salah tik (*typo*), neologisme, slang

Beberapa Tantangan

- Homograf, kata yang tulisannya sama tetapi maknanya berbeda, e.g. "beruang"
- Token yang tidak dikenali, e.g. salah tik (*typo*), neologisme, slang
- Kata dapat berubah makna dalam frasa, e.g. "mahasiswa" itu netral, tetapi "harga mahasiswa" itu positif

Model Bahasa Probabilistik

Manning & Schütze (1999)

$$
\begin{aligned}
P(\mathbf{w}) &= p(w_1, w_2, ..., w_{t-1}, w_t) \\
&= p(w_1)p(w_2|w_1)p(w_3|w_2, w_1)...p(w_t|w_1, w_2, ..., w_{t-1})
\end{aligned}
$$

Nilai ini bisa diaproksimasi sebagai

$$
P(w_1, w_2, ..., w_{t-1}, w_t) \approx p(w_1) \prod_{i=2}^{t} p(w_i|w_{i-1})
$$

Model Bahasa Probabilistik - Contoh

Diberikan tiga contoh kalimat:

1. Saya mau makan
2. Ali dan saya mau bermain
3. Kamu mau makan

Pertanyaan: "Ali mau ..."

$$p(makan|mau) = 2/3$$
$$p(bermain|mau) = 1/3$$

Model Bahasa Neural

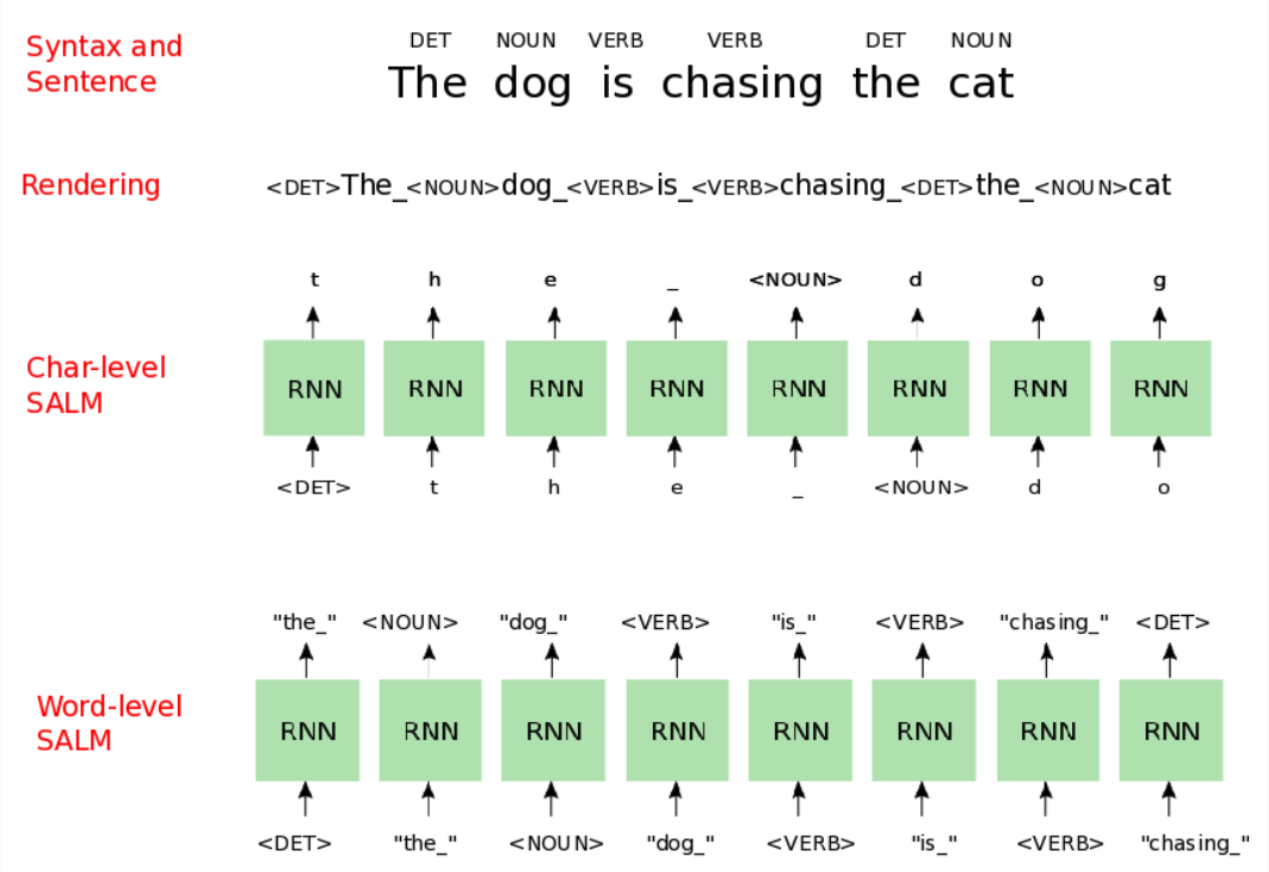


Figure 9: Hubungan antarkata dijelaskan dengan neural network (Sumber: Zalando Research)

Model Bahasa Neural

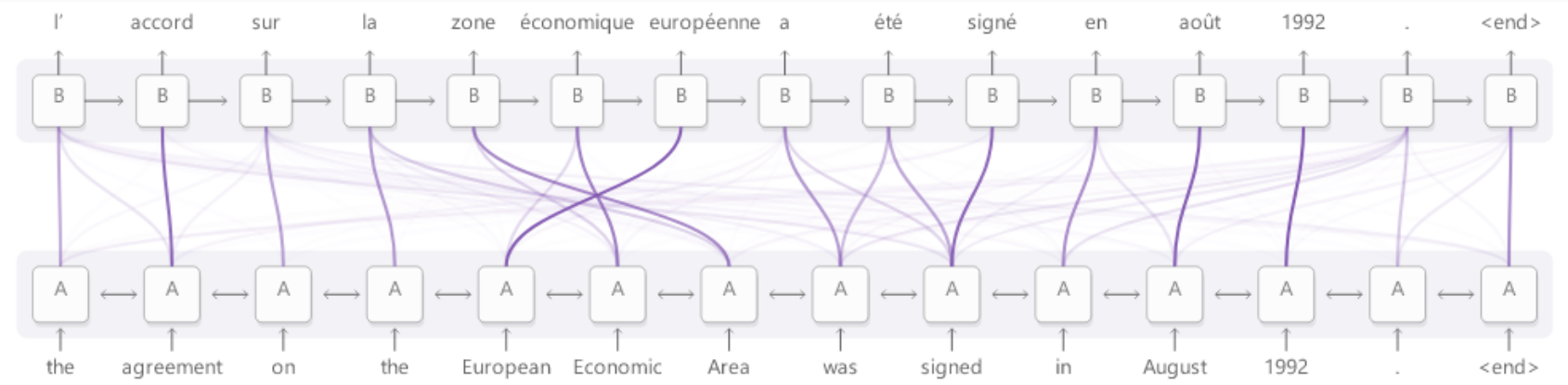


Figure 10: Terjemahan dengan neural network (Sumber: Distill)

Model Bahasa Neural

Figure 11: BERT Transformer (Sumber: Data Folks Indonesia)

NLTK

Natural Language Toolkit

"NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, **tokenization**, **stemming**, **tagging**, **parsing**, and **semantic reasoning**..."

NER Tagging

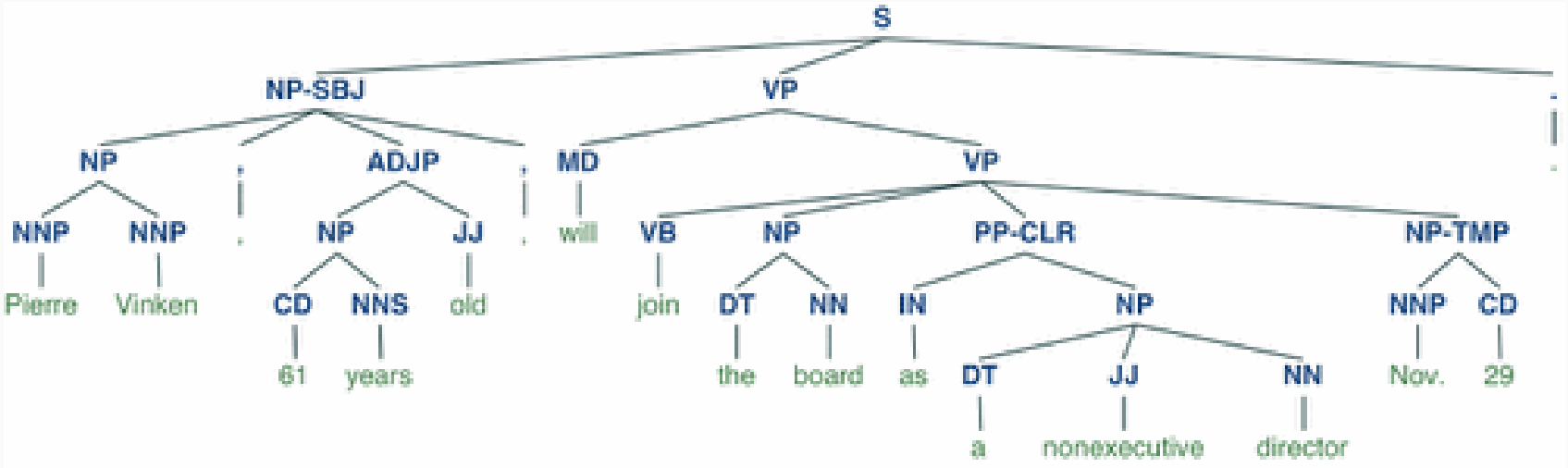


Figure 12: Hasil NER tagging dengan NLTK [NLTK Project, 2017]

Corpora & Models

Beberapa korpus dan model yang terkenal dari NLTK:

- Project Gutenberg Selections
- Penn Treebank
- SentiWordNet
- Stopwords Corpus
- Porter Stemmer

Everything Data

Document Similarity using NLTK and Scikit-Learn

Alternatif untuk NLTK

Beberapa alternatif untuk tugas-tugas spesifik:

- **spaCy**: Industrial-Strength Natural Language Processing in Python
- **gensim**: topic modelling for humans

	SPACY	NLTK	CORENLP
Programming language	Python	Python	Java / Python
Neural network models	✓	✗	✓
Integrated word vectors	✓	✗	✗
Multi-language support	✓	✓	✓
Tokenization	✓	✓	✓
Part-of-speech tagging	✓	✓	✓
Sentence segmentation	✓	✓	✓
Dependency parsing	✓	✗	✓
Entity recognition	✓	✓	✓
Entity linking	✗	✗	✗
Coreference resolution	✗	✗	✓

Figure 13: Perbandingan fitur (Sumber: spaCy)

	SPACY	NLTK	ALLEN-NLP	STANFORD-NLP	TENSOR-FLOW
I'm a beginner and just getting started with NLP.	✓	✓	✗	✓	✗
I want to build an end-to-end production application.	✓	✗	✗	✗	✓
I want to try out different neural network architectures for NLP.	✗	✗	✓	✗	✓
I want to try the latest models with state-of-the-art accuracy.	✗	✗	✓	✓	✓
I want to train models from my own data.	✓	✓	✓	✓	✓
I want my application to be efficient on CPU.	✓	✓	✗	✗	✗

Figure 14: Mana yang harus dipakai? (Sumber: spaCy)

Referensi

Bahan Bacaan

1. Bird, S., Edward L. & Klein, E. (2009). *Natural Language Processing with Python*. OReilly Media Inc.
2. Jurafsky, D. & Martin, J. H. (2018). *Speech and Language Processing (Vol. 3)*. Pearson.
3. Manning, C., Raghavan, P., & Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Kuliah

1. Stanford CS124: From Languages to Information
2. University of Edinburgh: Text Technologies for Data Science
3. Stanford CS276: Information Retrieval and Web Search (advanced)
4. Stanford CS224n: Natural Language Processing with Deep Learning (advanced)

Grup NLP

- Stanford NLP: Chris Manning, Dan Jurafsky, Percy Liang
- EdinburghNLP: Sharon Goldwater, Mirella Lapata, Ivan Titov, Mark Steedman, Shay Cohen, Walid Magdy, etc.
- UniMelb CIS School: Tim Baldwin, Trevor Cohn, Karin Verspoor
- UWNLP: Noah Smith, Luke Zettlemoyer
- QCRI
- Chris Dyer (DeepMind), Sebastian Ruder (DeepMind), Hal Daumé III (UMaryland), Graham Neubig (CMU), Kyunghyun Cho (NYU), Phil Blunsom (Oxford), Richard Socher (Salesforce), Isabelle Augenstein (Copenhagen)

Peneliti Indonesia di NLP

- ITB: Ayu Purwarianti, Masayu Leylia Khodra, Dessi Puji Lestari (speech)
- UI: Mirna Adriani, Rahmad Mahendra
- UGM: Yunita Sari
- Adhiguna Kuncoro (DeepMind), Dani Yogatama (DeepMind), Ruli Manurung (Google Japan), Clara Vania (NYU), Genta Indra Winata (HKUST), Samuel Cahyawijaya (HKUST), Kemal Kurniawan (UniMelb), Yudi Wibisono (UPI)

Referensi

NLTK Project (2 Januari 2017)

Natural Language Toolkit

`http://www.nltk.org/`

Richard Socher (diakses 15 Mei 2017)

CS224d: Deep Learning for Natural Language Processing

`http://cs224d.stanford.edu/`

Terima kasih