Rangkuman Guest Lecture dan Klaim Absensi PKB
Harvest Walukow
164231104

Bentuk-bentuk dari Natural Language
Secara umum bentuk dari natural language dapat dibagi dalam beberapa level, yakni phonetics & phonolg (sounds), morphology (words), syntax (frasa & kalimat), semantics (meanings), pragmatics (notion and context).

Definisi Natural Language Processing (NLP)
Cabang A) yang memungkinkan komputer untuk berinteraksi dengan manusia menggunakan bahasa alami.

NLP menjadi penting saat ini karena meningkatnya volume data berbentuk teks terutama dari sosial media. Diestimasi sekitar 80%-90% dari data ter-generate berbentuk unstructured data.

Contoh Unstructured Data
- Text files
- Email

- Social Media
- Website
- Mobile data
- Communications
- Media
- Business applicatons

Taxonomy of NLP
Ada banyak area NLP yang bisa dieksplor diantaranya Multimodality, Natural Language Interfaces, Semantic Text Processing, Sentiment Analysis, Syntactic Text Procesing, Linguistics & Cognitive NLP, Responsible & Trustworthy NLP, Reasoning, Multilinguality, Information Retrieval, Information Extraction & Text Mining, dan Text Generation.

Area penelitian NLP dapat dibedakan dalam diagram yang terbagi ke dalam empat kelompok berdasarkan total number of papers dan growth rate of number of papers, adapun keempat kelompok tersebut adalah Niche Fields of Study, Foundational Fields of Study, Rising Question Marks, dan Trending Stars.

Level-Level dari NLP
- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis
- Pragmantic Analysis

Pipeline NLP
Jika kita ingin membuat melakukan analisis NLP, terdapat pipeline yang dapat diikuti yang secara garis besar meliputi Text Wrangling and Pre-Processing, Understanding Syntax or Structure, lalu Processing and Functionality.

Untuk proses dasar NLP (basic), tahapannya dapat berupa data (raw text) -> text preprocessing - > vectorization -> models/algorithm -> evaluasi.

Text Preprocessing
Dalam prosesnya metode yang dapat dilakukan dalam text preprocessing adalah:

- **Tokenization**
  Mengubah kalimat menjadi beberapa kata atau beberapa token
- **Lowercasing**
  Mengubah semua huruf ke lower case, terkadang proses ini tidak perlu dilakukan
- **Noise Removal**
  Kebanyakan data dari media sosial mengandung banyak noise sehingga perlu untuk remove
- **Stopwords Removal**
  Daftar stopwords dapat diakses melalui ranks.nl/stopwords

- **Stemming**
  Mengambil kata dasar dari sebuah kata yang memiliki imbuhan

Teknik Vectorization
Terdapat beberapa teknik untuk vektorisasi, diantaranya
- BoW & TF-IDF
- Word Embeddings
- Trained Embeddings

Models/Algorithm
- Rule-based
- Machine Learning
- Deep Learning
- Generative AI

Introduction to LLM & ChatGPT
Large Language Model (LLM) adalah tipe model bahasa yang dilatih dengan teknik deep learning menggunakan banyak data teks. LLM menggunakan arsitektur transformer yang didesain untuk memahami bahasa manusia, codes, dll.

Komponen LLM: Data, Traning, Architecture

ChatGPT adalah model AI conversational yang didevelop oleh OpenAI dengan menerapkan arsitektur GPT (Generative Pre-trained Transformer). Chatgpt juga mengaplikasikan analisis NLP dengan QA, translating, summarizing, generate text, dan engaging in the conversation.

Salah satu aplikasi NLP adalah sentiment analysis. Metode ini semakin populer di dunia e-commerce, perusahaan melakukan analisis pada opini orang pada produk mereka.

Isu Etis NLP
- Bias
- Privasi
- Transparansi