# TUGAS I
## Pemrosesan Bahasa Alami
## Program Studi Informatika
## Fakultas Matematika dan Ilmu Pengetahuan Alam
## Universitas Syiah Kuala

**Dosen Pengasuh**

Prof. Dr. Taufik Fuadi Abidin, S.Si., M.Tech
Razief Perucha Fauzie Afidh, S.Si., M.Sc
Fathia Sabrina, S.T., M.Inf.Tech

---

## PENDAHULUAN

Tugas ini berkaitan dengan tahapan pra-pemrosesan (*preprocessing*) yaitu memeriksa apakah terjadi kesalahan ejaan dalam teks (*spelling correction*). Anda diminta untuk membangun sebuah sistem sederhana yang dapat mendeteksi dan mengoreksi kesalahan ejaan dalam teks berbahasa Inggris. Secara umum, sistem yang dirancang mampu:

1. Mendeteksi kesalahan ejaan dan mengidentifikasi kata-kata yang salah eja dalam sebuah kalimat atau paragraf.

2. Menyarankan koreksi ejaan yang benar berdasarkan konteks kalimat.

Tugas ini merupakan tugas individu, artinya setiap mahasiswa ***DIHARUSKAN MENGERJAKAN TUGAS SECARA PERORANGAN***. Tidak dibenarkan bagi mahasiswa memberikan hasil pekerjaannya kepada mahasiswa yang lain. Apabila dari tugas yang dikumpulkan ditemukan indikasi bahwa tugas tersebut adalah hasil kopian dari teman yang lain maka mahasiswa yang melakukan hal tersebut akan mendapat penalti pengurangan nilai menjadi ***0***.

## LANGKAH PENGERJAAN

Pelajari metode koreksi ejaan, seperti **algoritma Levenshtein Distance, Metode Hamming, dan Algoritma Rabin-Karp**. Pahami konsep N-grams dan carea menggunakannya dalam mendeteksi kesalahan ejaan.

**Gunakan dataset** yang memiliki pasangan kata salah dan koreksi yang benar. Salah satu dataset yang dapat digunakan adalah **TOEFL-Spell**, yang berisi lebih dari 6000 kesalahan ejaan dari esai penulis non-native (https://github.com/EducationalTestingService/TOEFL-Spell). Dataset lain yang dapat digunakan adalah **Spelling Correction Dataset** di Kaggle: https://www.kaggle.com/datasets/bittlingmayer/spelling.

Bersihkan data teks dari karakter atau simbol yang tidak diperlukan. Lakukan tokenisasi untuk memisahkan kata-kata dalam kalimat. Implementasikan model pembelajaran mesin atau *deep learning* untuk mendeteksi dan mengoreksi kesalahan ejaan. Sebagai alternatif, toolkit seperti **NeuSpell** dapat juga digunakan. NeuSpell adalah *open-source toolkit* untuk mengoreksi ejaan dalam bahasa Inggris (https://github.com/neuspell/neuspell). Jika waktu masih ada, coba bagun model berbasis **Recurrent Neural Network (RNN)** atau **Transformer** untuk mengoreksi kesalahan ejaan. Kinerja model dievaluasi berdasarkan indikator Accuracy, Precision, Recall, dan F1-Score.

**Tulis laporan dalam format PDF** yang mendokumentasikan seluruh proses, mulai dari teori pendukung, pengumpulan dataset, pengembangan model atau toolkit, hingga tahapan pengevaluasian kinerja model.

Dokumentasi terkait koreksi ejaan menggunakan **DeepPavlov** dapat diakses melalui tautan: https://docs.deeppavlov.ai/en/master/features/models/spelling_correction.html

## PENGUMPULAN TUGAS

Tugas ini harus dikumpulkan paling lambat pada tanggal **24 Februari 2025 pukul 23.59 WIB** secara elektronik via sistem e-learning (http://www.elearning.usk.ac.id/). File tugas yang diunggah via sistem e-learning harus berupa sebuah file terkompres (**zip** atau **rar** atau **tar**) yang didalamnya terdapat file-file tugas (*source code*) dan sebuah file **README.txt**. File **README.txt** berisi nama dan NPM mahasiswa serta penjelasan tambahan yang dianggap perlu untuk mendukung proses penilaian tugas ini. Nama file yang dikumpulkan harus ditulis dalam format sebagai berikut:

**nama_npm_kelas.zip** atau **nama_npm_kelas.rar** atau **nama_npm_kelas.tar**

Contoh:

**taufik_abidin_2208107010028_A.zip**

FMIPA
UNIVERSITAS
SYIAH KUALA