COMP6882001 – Natural Language Processing
Qualification Case

Last Updated: Odd Semester 2023/2024

Catatan

- Silahkan **mencari dan memasukkan tema atau topik** terlebih dahulu ke **list tema** yang sudah disediakan oleh *assistant development*. Pastikan tema atau topik yang ingin digunakan **belum pernah dipakai** oleh *assistant* lainnya.
- Setiap *assistant* diperbolehkan untuk memilih **lebih dari 1** tema atau topik (boleh berbeda untuk setiap soal).
- Silahkan **mengerjakan soal kualifikasi berdasarkan kriteria** yang sudah ditentukan.
- Setiap assistant akan menyelesaikan tugas kualifikasi ini menggunakan **Python Notebook** dengan **ekstensi file .ipynb**. Harap dipisahkan menjadi **dua file notebook** untuk **kedua kasus** yang diberikan.
- Setiap *assistant* wajib **mengumpulkan kualifikasi** ini dengan menyertakan **jawaban** beserta **dataset** yang digunakan.

Kriteria

Case 1: Text Mining I

Untuk Case 1, anda diminta untuk mencari satu topik dan mengimplementasikannya dalam bentuk *text mining*. Pada implementasi ini, anda akan menyertakan logika yang terdiri dari **tiga validasi** yang *advanced* untuk memastikan hasil yang akurat dan relevan.

Case 2: Text Mining II

Untuk Case 2, anda diminta untuk mencari satu topik dan mengimplementasikannya dalam bentuk *text mining*. Pada implementasi ini, anda akan menyertakan logika yang terdiri dari **tiga validasi** yang *advanced* untuk memastikan hasil yang akurat dan relevan.

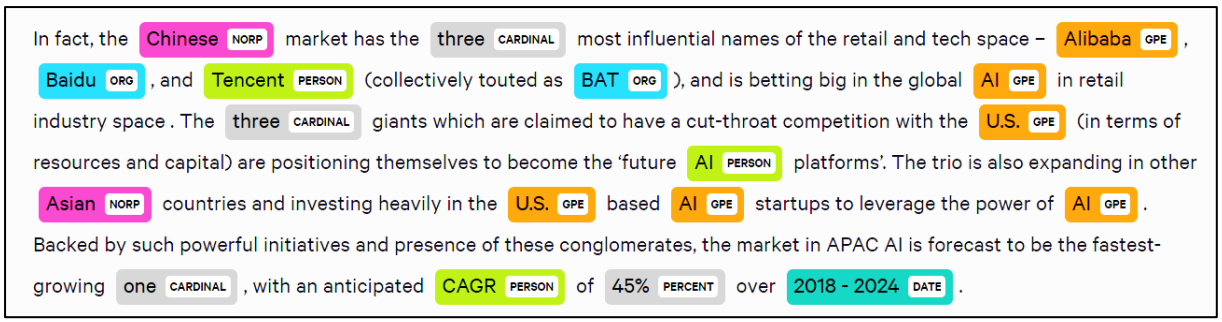
> ***Text mining*** adalah proses ekstraksi informasi berharga dari teks, dan dalam hal ini, kita dapat mengambil contoh topik seperti analisis sentimen terhadap ulasan produk. Pertama, kita dapat menggunakan teknik *natural language processing* untuk mengidentifikasi dan memahami sentimen positif, negatif, atau netral dalam setiap ulasan. Kedua, validasi dapat dilakukan dengan memeriksa keterkaitan antara sentimen yang ditemukan dengan kata-kata kunci tertentu yang berkaitan dengan produk. Ketiga, kita dapat menggunakan *context analysis* untuk memastikan bahwa sentimen yang dihasilkan tidak hanya bergantung pada kata-kata individual, tetapi juga memperhitungkan struktur kalimat dan hubungan antar kalimat. Dengan menerapkan tiga validasi ini, kita dapat meningkatkan ketepatan hasil text mining dalam menganalisis sentimen terhadap ulasan produk.

Case 3: Grammar Parsing using Natural Language ToolKit

Pada Case 3, Anda diminta untuk membuat *code* guna menyelesaikan permasalahan yang tertera dalam soal.

Case

1. **Text Mining I**
   a. Aplikasi yang dibangun harus memanfaatkan sejumlah teknik pemrosesan teks, termasuk **tokenisasi**, **penghapusan kata umum (stop words)**, **stemming atau lemmatisasi**, **POS Tagging**, **NER (Named Entity Recognition)**, **distribusi frekuensi**, **pengambilan korpora dari data NLTK atau situs web**, **pemanfaatan WordNet**, **ekstraksi fitur**, **klasifikasi menggunakan metode Naïve Bayes**, serta kemampuan untuk **menyimpan dan memuat model klasifikasi**.
   b. Model yang dihasilkan dari arsitektur Naïve Bayes harus mencapai tingkat akurasi **minimal sebesar 80%**.
   c. Aplikasi juga diharapkan mampu menampilkan **10 *most informative features*** dari dataset yang digunakan.
2. **Text Mining II**
   a. Aplikasi yang dibuat harus memanfaatkan sejumlah teknik pemrosesan teks, termasuk **model bahasa atau *language modelling* (n-gram)**, ***word embedding*** dengan menggunakan metode **Word2Vec** atau **GloVe**, ***grammar parsing*** dengan menggunakan **Natural Language ToolKit (NLTK)**, ***dependency parsing*** dengan menggunakan **SpaCy**, dan mengimplementasikan **Named Entity Recognition (NER)**.
   b. Dalam bagian ***grammar parsing*** menggunakan **NLTK**, aplikasi harus dapat menampilkan ***grammar parsing tree*** dari **Context Free Grammar (CFG)** yang telah dikonstruksi dengan kalimat yang sudah dibuat sebelumnya. Selain itu, CFG yang dibuat harus memenuhi persyaratan umum, dan Anda harus dapat menjelaskan CFG yang telah Anda buat secara rinci dan jelas.
   c. Pada bagian **Named Entity Recognition (NER)**, Anda diminta untuk menampilkan **Named Entities** dari data yang telah diolah sebelumnya. Ilustrasi dapat dilihat di bawah ini.



3. Grammar Parsing using Natural Language ToolKit

Buatlah representasi ***grammar production*** dalam bentuk **Context Free Grammar (CFG)** beserta ***parsing tree***-nya untuk menangani enam kalimat berikut:

a. The farmer loaded the cart with sand
b. The farmer loaded sand into the cart
c. The farmer filled the cart with sand
d. The farmer filled sand into the cart
e. The farmer dumped the cart with sand
f. The farmer dumped sand into the cart

Komponen Penilaian

- Text Mining I (45%)
- Text Mining II (45%)
- Grammar Parsing using Natural Language ToolKit (10%)