PROPOSAL PROYEK

11S4037 - PEMROSESAN BAHASA ALAMI

Product Matching for Youtube Channels use Term Frequency - Inverse Document Frequency

Disusun oleh:

1. *11S16008 Amzesmoro Sianturi*
2. *11S16017 Herianto Saragi*
3. *11S16050 Evrin Lumbantobing*
4. *11S16051 Anggiat Maruli M*

PROGRAM STUDI SARJANA INFORMATIKA

FAKULTAS TEKNIK INFORMATIKA DAN ELEKTRO

INSTITUT TEKNOLOGI DEL

NOVEMBER 2019

DAFTAR ISI

1. Pendahuluan

Bagian ini berisi latar belakang dan tujuan pengerjaan proyek.

1.1.Latar Belakang

Product Matching adalah sebuah tantangan untuk memeriksa dua representasi berbeda dari product retail dan menentukan apakah keduanya merujuk pada product yang sama [1]. Sehingga berdasarkan fungsi tersebut kami akan melakukan sebuah pembuktian dengan menggunakan algoritma TF-IDF. Dan data yang akan digunakan berasal dari dataset yang telah di download dari internet. Dataset tersebut harus memiliki jumlah data mininum 100,000 data.

1.2.Tujuan

Tujuan dari pelaksanaan proyek ini adalah sebagai berikut:

1. Menerapkan metode TF - IDF dalam melakukan product matching berdasarkan data set yang ada.
2. Menyelesaikan Proyek Mata kuliah Pemrosesan Bahasa Alami.

1.3.Ruang Lingkup

Ruang lingkup dari proyek ini adalah:

1. Data yang digunakan dalam melakukan product matching adalah data set dari YouTube channels ~100000 (Oleh: Ilya Babikov) yang berasal dari kaggle. Data set tersebut berisi sekitar 861973 data.
2. Data set tersebut akan dibagi menjadi dua bagian yaitu data training dan data test. Data training digunakan sebagai alat, agar metode tf-idf dapat belajar berdasarkan data yang ada, dan data test digunakan untuk menguji data yang digunakan.

2. Landasan Teori

Bagian ini berisi teori-teori yang mendukung pengerjaan proyek.

2.1. *Tf-Idf*

Algoritma TF-IDF (Term Frequency – Inverse Document Frequency) adalah salah satu algoritma yang dapat digunakan untuk menganalisa hubungan antara sebuah frase/kalimat dengan sekumpulan dokumen. Contoh yang dibahas kali ini adalah mengenai penentuan urutan peringkat data berdasarkan query yang digunakan.

Inti utama dari algoritma ini adalah melakukan perhitungan nilai TF dan nilai IDF dari sebuah setiap kata kunci terhadap masing-masing dokumen. Nilai TF dihitung dengan rumus TF = jumlah frekuensi kata terpilih / jumlah kata dan nilai IDF dihitung dengan rumus IDF = log(jumlah dokumen / jumlah frekuensi kata terpilih). Selanjutnya adalah melakukan perkalian antara nilai TF dan IDF untuk mendapatkan jawaban akhir.

3. Referensi

[1]https://www.slideshare.net/govind201/5-lessons-ive-learned-tackling-product-matching-for-ecommerce?ref=https://www.semantics3.com/blog/5-lessons-ive-learned-tackling-product-matching-for-e-commerce-bf55f7d4ac07/