PENGEMBANGAN DATA UJI SISTEM KOMPUTASI KEMIRIPAN TEKS SECARA SEMANTIK BERBAHASA INDONESIA

Riza Akbar Nurhadi[1], Faisal Rahutomo[2], Budi Harijanto[3]
[1,2,3]Program Studi Teknik Informatika, Jurusan Teknologi Informasi, Politeknik Negeri Malang
[1]rizakbar2@hotmail.com, [2]faisal.polinema@gmail.com, [3]budi.hijet@gmail.com

Abstrak

Komputer membutuhkan *Natural Lenguage Processing* (NLP) agar komputer dapat memahami bahasa alami manusia. Untuk membuat NLP dibutuhkan standar data uji, seperti yang sudah diteliti pada penelitian *Microsoft Research: Video Description Corpus*. Tetapi *Microsoft Research* hanya menyediakan data uji menggukan bahasa Inggris, maka dari itu diperlukan data uji berbahasa indonesia agar dapat membantu peneliti pemrograman bahasa alami berbahasa Indonesia. Seperti yang dilakukan oleh *Microsoft Research: Video Description Corpus*, penelitian ini menggunakan responden untuk mendeskripsi sebuah video agar mendapatkan ungkapan yang bermacam-macam tetapi memiliki makna yang sama. Maka dari itu dibutuhkan sistem yang dapat menampilkan video dan mendeskripsi video yang dapat digunakan untuk mendapatkan data uji. Data uji tersebut akan dianalisa menggunakan Metode *Jaccard* yang dapat menghitung nilai perbedaan kata yang digunakan dari setiap ungkapan responden. Sehingga dengan metode ini dapat menentukan kualitas data uji yang akan didapatkan. Video yang digunakan sebanyak 150 video dan di dekripsi oleh responden sebanyak kurang lebih 28 orang. Penelitian ini berhasil mendapatkan data uji yang baik sehingga dapat digunakan untuk penelitan NLP seperti Sistem komputasi teks secara semantik.

Kata kunci: ***Natural lenguage processing*****, Ungkapan,** ***Microsoft Research Video Description Corpus*****, Metode** ***Jaccard.***

1. Pendahuluan

Di dalam ungkapan bahasa sehari-hari, sering muncul ungkapan yang berbeda-beda pada momen atau kejadian yang sama, sehingga terdapat banyak kalimat yang dapat menjelaskan satu momen atau kejadian. Agar komputer dapat mengerti makna dari ungkapan yang berbeda-beda dan memiliki makna yang sama, maka terdapat kecerdasan buatan berupa *Natural Lenguage Processing* (NLP) yang berkaitan dengan interaksi antara komputer dengan manusia. Kasus ini membutuhkan komputasi bahasa pada tingkatan makna yang tidak bergantung pada bentuk susunan leksikalnya. Saat ini komputasi bahasa banyak digunakan di aplikasi-aplikasi seperti pengevaluasi kata, mesin terjemahan, mesin tanya jawab, dan lain sebagainya. Tetapi dengan kurangnya standar data uji, maka para peneliti komputasi bahasa membuat data uji sendiri sehingga sulit untuk membandinggan sistem komputasi bahasa lain karena memiliki data uji berbeda.

Pengembangan sistem komputasi yang mampu menangani bahasa pada tingkat makna memerlukan perangkat-perangkat yang mendukungnya. Perangkat tersebut terdiri dari matriks evaluasi dan data uji. Sumber daya tersebut telah terdesia dalam bahasa inggris yang salah satunya dilakukan oleh *Microsoft Research* (MSR) dalam penelitiannya yaitu *Video Description Corpus* yang melakukan matriks evaluasi yang didapatkan dari deskripsi berberapa video pendek. Dengan itu, *MSR Video Description Corpus* dapat mengoleksi kurang lebih 85 ribu kalimat paralel dengan bahasa inggris. Tetapi data uji masih belum tersedia untuk bahasa Indonesia.

Didalam penelitian ini akan dikembangkan data uji untuk sistem komputasi kemiripan teks secara semantik berbahasa Indonesia. Untuk mendapatkan kalimat yang banyak untuk dibandingkan, maka penulis menggunakan video pendek dan meminta sekitar 20 orang untuk mendeskripsikannya dalam 1 kalimat menggunakan bahasa Indonesia. Video yang akan digunakan adalah video yang menampilkan sebuah peristiwa dengan jelas dengan durasi kurang lebih antara 10 sampai 20 detik sebanyak 150 video. Dalam penelitian ini menggunakan metode *Jaccard* untuk mengetahui kualitas dari kalimat tersebut.

Untuk mempermudah dalam menampilkan video, mendapatkan deskripsi dan menganalisanya maka penelitian ini menggunakan web aplikasi bahasa pemrograman *PHP* dengan *framework CodeIgniter* dan *MySql* sebagai *database*-nya.

2. Tinjauan Pustaka

2.1 NLP

Natural Language Processing (NLP) merupakan kecerdasan buatan dalam bahasa (linguistik) yang berkaitan dengan interaksi antara komputer dan bahasa alami manusia. Tujuan utama dari studi NLP adalah membuat mesin yang mampu mengerti dan memahami makna bahasa manusia lalu memberikan respon yang sesuai.

Sejarah NLP dimulai pada tahun 1950-an, Alan Turing mengusulkan tes yang sekarang disebut dengan Tes *Turing*. Tes *Turing* adalah sebuah tes yang mengukur kemampuan mesin untuk menunjukan perilaku cerdas. tes yang dilakukan adalah membuat

seseorang membedakan antara mesin dan manusia, jika orang tersebut tidak bisa membedakannya, maka mesin tersebut dikatakan lulus tes. Contoh aplikasi NLP adalah *Google Translate*.

2.2 *Microsoft Research Video Description Corpus*

Microsoft Research Video Description Corpus (MSRVDC) adalah penelitian yang dilakukan oleh *Microsoft Research* yang melakukan data uji yang didapatkan dari deskripsi berberapa video pendek. Dengan itu, *MSR Video Description Corpus* dapat mengoleksi kurang lebih 85 ribu kalimat paralel dengan bahasa inggris.

MSR Video Description Corpus digunakan untuk membantu para peneliti yang bekerja pada komputasi bahasa sehingga memungkinkan para peneliti untuk membandingkan sistem dan menggunakannya dalam berberapa jenis mesin automatis. Salah satu penelitian yang menggunakan data uji *MSR Video Descrption Corpus* adalah penelitian "*Test Collection Recycling for Semantic Text Similarity*" yang di teliti oleh Faisal Rahutomo, Teruaki Kitasuka, dan Masayoshi Aritsugi. Penelitian tersebut menggunakan data uji *MSR Video Description Corpus* sebagai acuan dasar evaluasi pencocokan teks secara semantik. Dengan pencocokan teks secara semantik maka dapat meyamakan kata-kata atau kalimat yang memiliki makna yang sama meskipun kata-kata atau kalimat tersebut berbeda.

2.4 *Jaccard*

Jaccard adalah salah satu metode yang digunakan untuk mengecek kesamaan antara 2 objek yang bersifat biner. Formula untuk menghitung kesamaaan antara dua objek A dan B adalah sebagai berikut:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jika A dan B kosong maka J(A,B) = 1. Dalam penelitian ini untuk mendapatkan data uji yang baik maka dibutuhkan kalimat yang berbeda-beda tetapi memiliki makna yang sama. Sehingga semakin kecil perhitungan *jaccard* maka semakin baik data uji tersebut.

Sebagai Contoh pertama kalimat A berisi "Saya Mengendarai Sedan" dan Kalimat B "Aku Membawa Mobil", maka:

Diketahui : |A∩B|=0,
|A∪B|=6
Jawab : J(A,B)= (|A∩B|)/(|A∪B|)= 0/6=0

Maka kesamaan antara kalimat A dan B adalah 0, Berikut Contoh kedua, kalimat C berisi "Saya Mengendarai Sedan" dan kalimat D yaitu "Saya mengendarai mobil", maka:

Diketahui : |C∩D|=2 ,
|C∪D|=4
Jawab : J(C,D)= (|C∩D|)/(|C∪D|)= 2/4=0,5

Pada contoh kedua diketahui nilai kesamaan dari C dan D adalah 0,5. Dari kedua contoh tersebut, contoh pertama merupakan data uji yang lebih baik dari pada contoh kedua karena nilai perhitungan contoh pertama (0) lebih kecil dari pada contoh kedua (0,5).

2.5 *CodeIgniter*

CodeIgniter (CI) adalah *framework* pengembangan aplikasi (*Application Development Framework*) yang menggunakan *PHP* yang di rilis pada tahun 2006. Kerangka dalam membuat program *PHP* jadi lebih sistematis. Pemogram tidak perlu membuat program dari awal karena CI menyediakan sekumpulan *library* yang diperlukan untuk menyelesaikan pekerjaan yang umum.

Framework CI menggunakan teknik pemrograman MVC (*Model View Controller*) yang merupakan salah satu teknik pemrograman terbaik dan digunakan oleh banyak bahasa pemrograman saat ini (Betha Sidik, 2012). MVC adalah konsep memisahkan antara logic dengan tampilan dan *database*. Model merupakan kode struktur data. Model berisi fungsi di dalam pengolahan *database*. *View* Merupakan kode untuk menampilkan tampilan suatu program. Tampilan dapat berupa *web page, header, footer* dan apa saja yang berjenis tampilan. *Controller* Merupakan kode untuk *logic*, algoritma dan sebagai penghubung antara *model, view*, dan sumber lain yang di perlukan untuk mengolah *HTTP request* dan *generate web page*.

2.6 *PHP*

PHP atau singkatan dari "*PHP: Hypertext Prepocessor*" merupakan bahasa berbentuk skrip yang di gunakan secara luas untuk penaganan sebuah situs web dan bisa digunakan bersamaan dengan *HTML*. *PHP* dirancang untuk membentuk aplikasi web dinamis. Artinya, *PHP* dapat membentuk suatu tampilan berdasarkan permintaan terkini. Kelahiran *PHP* bermula saat Rasmus Lerdorf membuat sejumlah skrip *Perl* yang yang dapat mengamati siapa saja yang melihat-lihat daftar riwayat hidupnya, yakni pada tahun 1994. skrip-skrip ini selanjutnya dikemas menjadi *tool* yang disebut "*Personal Home Page*". Paket inilah yang menjadi cikal bakal *PHP*. Pada tahun 1995, Rasmus menciptakan *PHP/FI* Versi 2. pada versi inilah pemrogram dapat menempelkan kode terstruktur di dalam tag *HTML*. Yang menarik, kode *PHP* juga bisa berkomunikasi dengan database dan melakukan perhitungan-perhitungan yang kompleks diambil jalan.

PHP bersifat bebas dipakai. tidak perlu membayar apa pun untuk menggunakan perangkat lunak ini alias *free PHP* tersedia dalam bentuk kode biner maupun kode sumber yang lengkap.

2.7 *MySQL*

Adalah sebuah sebuah perangkat lunak sistem manajemen basis data *SQL* atau DBMS yang *multithread*, *multi-user*, dengan sekitar 6 juta instalasi di seluruh dunia.
Beberapa kelebihan MySQL antara lain :

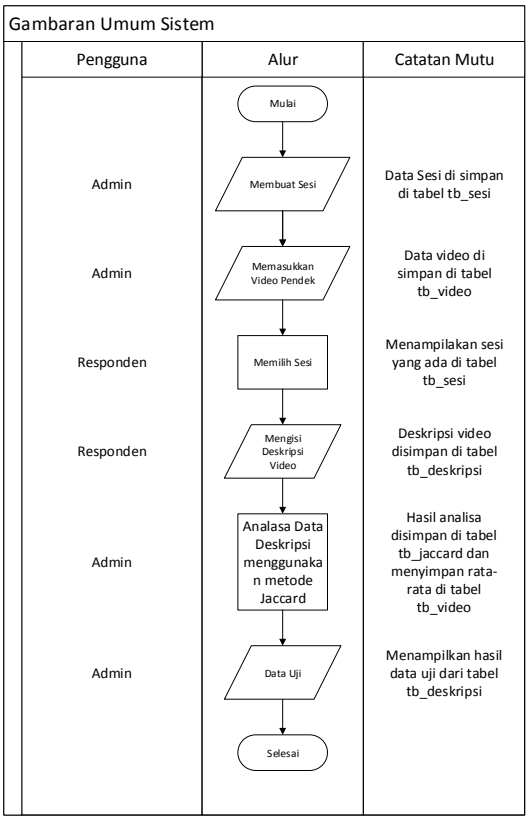
- *Free* (bebas di download)
- Stabil dan tangguh
- Fleksibel dengan berbagai pemrograman
- Keamanan yang baik
- Dukungan dari banyak komunitas
- Kemudahan management database
- Mendukung transaksi
- Perkembangan software yang cukup cepat.

3. Perancangan

3.1 Desain Sistem

3.1.1 Gambaran Umum

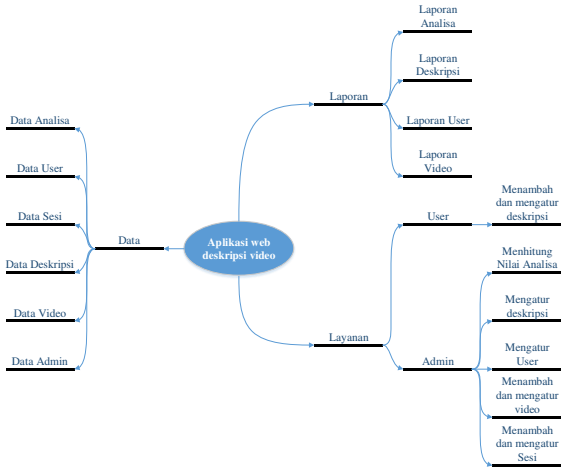
Untuk mendapatkan data uji penelitian ini maka penulis menggunakan cara yang telah dilakukan oleh MSRVDC (*Microsoft Resource Video Description Corpus*) yaitu dengan cara memainkan sebuah video pendek lalu mendeskripsikannya dalam 1 kalimat. Karena membutuhkan kalimat yang banyak, maka dibutuhkan 150 video pendek dan kurang lebih 20 orang relawan yang akan mendeskripsikan video tersebut. Video dapat didapatkan dari berbagai sumber seperti youtube. Video pendek yang dibutuhkan harus menampilkan sebuah kejadian yang jelas. Relawan yang mendeskripsikan dapat dilakukan siapapun yang dapat berbahasa Indonesia dengan baik. Pendeskripsian video dilakukan di satu ruangan dengan mengakses web lokal yang disediakan server. Berikut *flowchart* gambaran umum :



Gambar 3.1 *Flowchart* gambaran umum

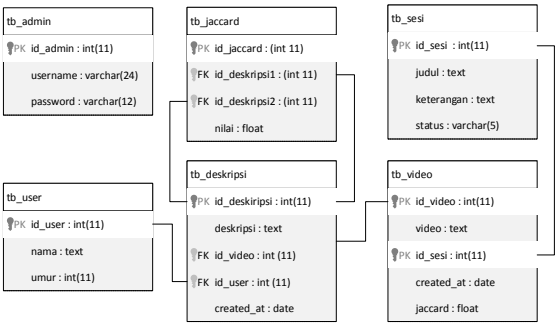
3.1.2 Work Breakdown Structure

Aplikasi web deskripsi video ini memiliki layanan, data, dan laporan. Layanan terbagi menjadi dua yaitu *user* atau sebagai responden, dan admin. layanan *user* dapat menambah dan mengatur data deskripsi. Sedangkan admin memiliki lima layanan yaitu menambah dan mengatur data sesi, data video, data *user*, data deskripsi, dan menghitung nilai analisa yang berelasi di data analisa. admin dapat melihat laporan admin, laporan deskripsi, laporan *user*, dan laporan video. Data laporan diambil dari data analisa, data deskripsi, data sesi, data *user*, dan data video. WBS dari Aplikasi web deskripsi video terdapat pada gambar berikut:



Gambar 3.2 *Work Breakdown Structure*

3.2 Desain Database



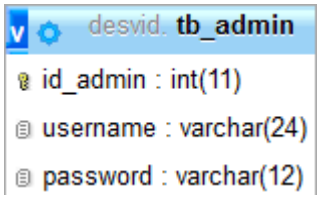
Gambar 3.3 Desain database

Database pada aplikasi ini menggunakan MySql phpmyadmin dengan nama 'desvid'. Aplikasi ini membutuhkan 6 tabel. Tabel tersebut terdiri dari tb_admin, tb_user, tb_sesi, tb_video, tb_deskripsi, dan tb_jaccard. Tb_video berelasi dengan tb_sesi untuk menentukan sesi pada video, tb_deskripsi berelasi dengan tb_video dan tb_user untuk menentukan video yang di deskripsi dan user yang mengisi deskripsi, tb_jaccard berelasi dengan tb_deskripsi untuk menentukan deskripsi yang dibandingkan dan dihitung.

4. Implementasi

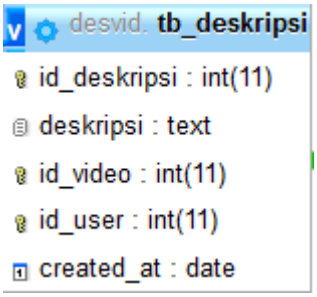
4.1 Implementasi *Database*

Implementasi Basis data sebagai berikut.



Gambar 4.1 Tabel tb_admin

Dalam tabel ini memiliki field id_admin sebagai *primary key* dan *field username* dan *password* dengan tipe data *varchar* untuk menyimpan data *username* dan *password* ketika login sebagai admin.



Gambar 4.2 Tabel tb_deskripsi

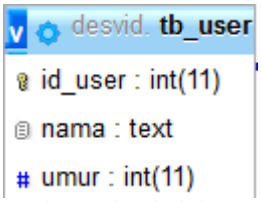
Tabel tb_deskripsi memiliki 5 *field* terdiri dari id_deskripsi sebagai *primary key*, field deskripsi dengan tipe data *text* untuk menyimpan data deskripsi, field id_video sebagai *foreign key* dari tabel tb_video untuk menandakan deskripsi tersebut untuk video yang tercatat, field id_user sebagai *foreign key* dari tabel tb_user untuk menandakan deskripsi tersebut dilakukan oleh user yang terdaftar, dan field created_at untuk menyimpan waktu deskripsi tersebut dibuat.

Gambar 4.3 Tabel tb_jaccard

Tabel tb_jaccard digunakan untuk menyimpan nilai analisa deskripsi yang telah dihitung. *Field* id_jaccard sebagai *primary key*, field id_deskripsi1 sebagai *foreign key* dari tabel tb_deskripsi untuk menentukan perbandingan kalimat pertama, *field* id_deskripsi2 sebagai *foreign key* dari tabel tb_deskripsi untuk menentukan perbandingan kalimat kedua, dan *field* nilai untuk mencatat nilai *jaccard.*

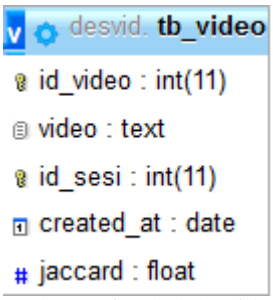
Gambar 4.4 Tabel tb_sesi

Tabel tb_sesi untuk membuat sesi setiap melakukan deskripsi video. Tabel ini berisi id_sesi sebagai *primary key*, *field* judul, keterangan, dan status digunakan untuk memberi judul, keterangan dan status sesi tersebut.



Gambar 4.5 Tabel tb_user

Tabel tb_user digunakan untuk menyimpan nama dan umur responden. *Field* tabel ini berisi id_user sebagai *primary key*, *field* nama, dan *field* umur.



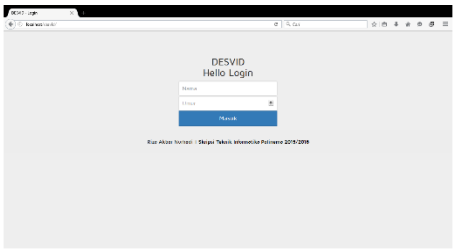
Gambar 4.6 Tabel tb_video

Tabel tb_video digunakan untuk menyimpan informasi video. Tb_video berisi 5 *field* yang terdiri dari id_video sebagai *primary key*, *field* video untuk menyimpan nama video, id_sesi sebagai *foreign key* dari tabel tb_sesi untuk menentukan sesi yang digunakan, *field* created_at digunakan untuk menyimpan kapan video di upload, dan *field* jaccard berisi rata-rata hasil analisa dalam satu video.

4.2 Implementasi Antarmuka

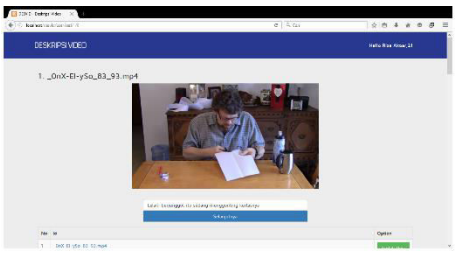
Implementasi antarmuka akan menjelaskan antarmuka sistem yang digunakan untuk penelitian pengembangan data uji sistem komputasi kemiripan teks secara semantik berbahasa indonesia.

Halaman Awal berisi 2 kolom nama dan umur sebagai syarat login masuk ke sistem ini. Implementasi ditunjukan pada Gambar 4.7



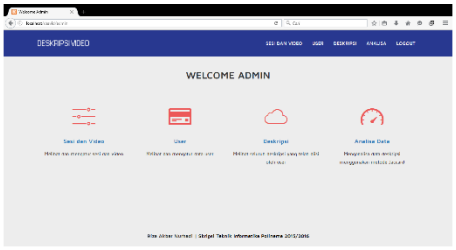
Gambar 4.7 Halaman Login

Paga Gambar 4.8 adalah halaman deskripsi, pada halaman ini responden dapat melihat video dan mengisi deskripsi.



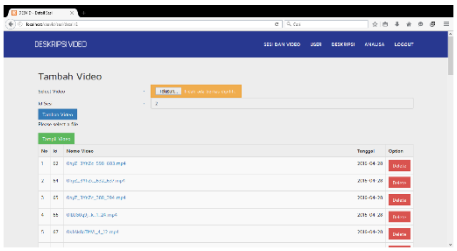
Gambar 4.8 Halaman Deskripsi

Halaman Admin berisi menu pilihan yang dapat digunakan oleh admin. Menu yang dapat dipilih adalah Sesi dan Video, *User*, Deskripsi, Analisa Data. Implementasi halaman admin dapat dilihat di Gambar 4.9



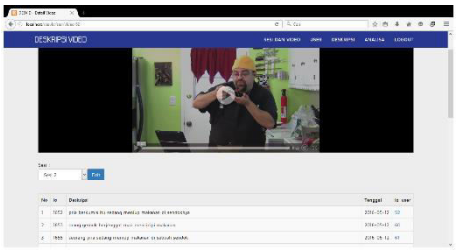
Gambar 4.9 Halaman Admin

Halaman Detail Sesi berisi daftar video dalam sesi tersebut. Pada halaman ini admin dapat menambah atau menghapus video. Implementasi Halaman Detail Sesi terdapat pada Gambar 4.10.



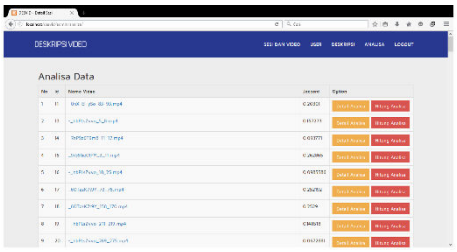
Gambar 4.10 Halaman Detail Sesi

Halaman Video adalah halaman yang menampilkan video dan deskripsi yang sudah diisi pada video tersebut. Di halaman ini, admin dapat merubah sesi video tersebut. Berikut implementasinya.



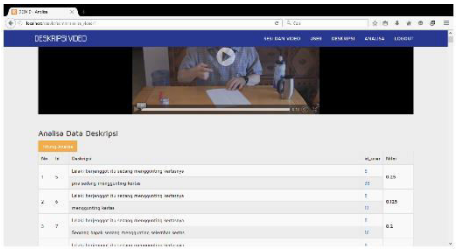
Gambar 4.11 halaman Video

Halaman Analisa adalah halaman yang berisi daftar video yang ada dan rata-rata hasil analisa video tersebut. Di halaman ini terdapat menu Detail Analisa dan Hitung Analisa. Berikut Imlementasinya.



Gambar 4.12 Halaman Analisa

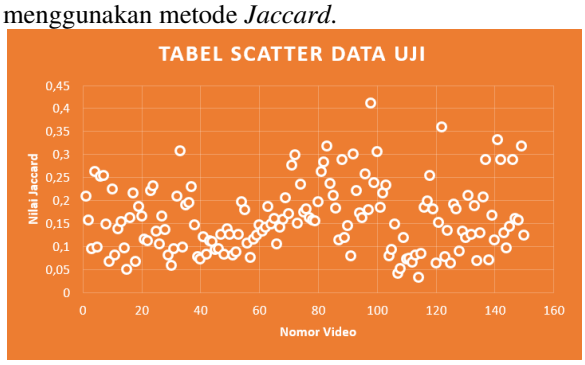
Halaman Detail Analisa terdapat Video dan Tabel Perbandingan dua deskripsi beserta nilai analisanya. Di halaman ini juga terdapat menu Hitung Analisa untuk menhitung analisa deskripsi tersebut.



Gambar 4.13 Halaman Detail Analisa

5. Analisa

Analisa digunakan untuk melihat apakah data uji yang di dapat sudah baik atau belum. Berikut tabel scatter hasil analisa data uji setiap video menggunakan metode *Jaccard.*



Gambar 4.14 Tabel *scatter* data uji

Dari tabel diatas dapat dilihat semua nilai kesamaan kata yang digunakan deskripsi setiap video kurang dari 0,5 dapat disimpulkan bahwa deskripsi berisi kata yang bermacam-macam.

Berikut menghitung rata-rata *jaccard*:

$$\hat{x} = \frac{Jumlah\ Nilai\ Jaccard\ Video}{Jumlah\ Video}$$

$$\hat{x} = \frac{23{,}9325}{150} = 0{,}15955$$

Hasil rata-rata seluruh nilai *jaccard* adalah 0,159. Dari tabel scatter dan nilai rata-rata *jaccard* tersebut dapat disimpulkan bahwa data uji yang telah dibuat adalah data uji yang bagus.

6. Kesimpulan dan Saran

6.1 Kesimpulan

Adapun kesimpulan yang dapat diambil dari hasil penelitian pengembangan data uji sistem komputasi kemiripan teks secara semantik berbahasa indonesia. Berikut kesimpulan:

- Aplikasi web deskripsi video dapat membantu dalam pengumpulan data deskripsi,
- Aplikasi web deskripsi video dapat menghitung nilai *jaccard* dengan akurat,
- Metode *jaccard* dapat mengukur perbedaan penggunaan kata antar kalimat,
- Hasil data uji dapat digunakan untuk sistem komputasi teks secara semantik atau sistem *Natural Language Processing* berbahasa Indonesia.

6.2 Saran

Saran yang diberikan untuk pengembangan penelitian ini dapat dikembangkan dengan meningkatkan jumlah video dan responden sehingga mendapatkan ungkapan yang lebih bermacam dan hasil yang lebih baik.

7. Daftar Pustaka

Alifian Sukma, Bagus Puji Santoso, Dian Ramadhan, Ni Made Ayu Karina Wiraswari, Tiara Ratna Sari. "Klasifikasi Dokumen Bahasa Jawa Menggunkan Metode N-Gram" Jurusan Sistem Informasi, Fakultas Sains Dan Teknologi Universitas Airlangga, Surabaya

Davi L. Chen, William B. Dollan. 2010 "Collecting Highly Parallel Datafor Paraphrase Evaluation" Department of Computer Science The University Texas at Austin and Microsoft Research One Microsoft Way

Faisal Rahutomo, dan Erfan Rohadi. "Pengembangan Piranti Penelitian Sistem Temu Kembali Informasi Bahasa Indonesia"

Faisal Rahutomo, Teruaki Kitasuka, dan Masayoshi Aritsugi. 2012 "Test Collection Recycling for Semantic Text Similarity" State Polytechnics of Malang and Kumamoto University

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu. 2013 "Using of Jaccard Coefficient for Keywords Similarity" Proceedings of the International MultiConference of Engineers and Computer Scientists.

"Microsoft Video Description Corpus" http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/

"Scorring with Jaccard Coefficient" https://class.coursera.org/nlp/lecture/184