

ABSTRAK

Banyaknya informasi wisata di internet saat ini membuat penyedia situs semakin sulit untuk mengelompokkan informasi sesuai dengan kategori yang sebenarnya. Terdapat banyak kategori wisata yang tersebar di permukaan bumi dengan berbagai pilihan mulai dari wisata alam, wisata buatan, wisata religi, wisata budaya dan wisata kuliner. *Klasifikasi* menjadi salah satu cara untuk penyusunan data secara rapi yang mengacu pada beberapa aturan untuk menentukan suatu data tersebut apakah sesuai dengan yang telah dikategorikan sebelumnya. Algoritme klasifikasi banyak digunakan secara luas salah satunya *K-Nearest Neighbor*. Algoritme tersebut melakukan klasifikasi berdasarkan jarak terdekat antara *data testing* dengan *data training* berdasarkan jumlah nilai k (k = banyaknya *data training*). Proses menghitung dekat atau jauhnya jarak dapat dihitung dengan menggunakan metode *Dice Similarity*, metode ini merupakan metode pengukuran jarak kesamaan antar dua objek dengan cara mengalikan keterkaitan setiap pasangan vektor (*inner product*) dalam ruang kata. Pengujian penelitian dengan menggunakan data sebanyak 200 dokumen pariwisata berupa teks dengan melakukan percobaan 5 *k-fold* menunjukkan *k-nearest neighbor* dengan jarak *dice similarity* mendapatkan nilai *recall* 47%, *precision* 45%, *f-measure* sebesar 44%, dan waktu yang dibutuhkan 864,2 detik.

Kata Kunci : *Pariwisata, Klasifikasi, K-Nearest Neighbor, Dice Similarity.*

DAFTAR ISI

HALAMAN PENGESAHAN	ii
HALAMAN PERNYATAAN ORISINALITAS	iii
HALAMAN PERSEMBAHAN	iv
ABSTRAK	v
DAFTAR ISI.....	vi
DAFTAR TABEL	viii
DAFTAR GAMBAR.....	ix
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah	2
1.2.1 Permasalahan	2
1.2.2 Usulan Solusi	2
1.2.3 Pertanyaan Penelitian.....	3
1.3 Tujuan dan Manfaat	3
1.3.1 Tujuan	3
1.3.2 Manfaat	3
1.3.3 Batasan Masalah	3
1.4 Sistematika Proposal	3
BAB II TINJAUAN PUSTAKA.....	5
2.1 Landasan Teori.....	5
2.1.1 Text Mining	5
2.1.2 Klasifikasi	5
2.1.3 Preprocessing Text.....	5
2.1.4 TF-IDF	7
2.1.5 Similarity Measure.....	8
2.1.6 K-Nearest Neighbor	10
2.1.7 K-fold Cross Validation	11
2.2 Penelitian Terkait	12
BAB III METODE USULAN.....	17
3.1 Perancangan Sistem	17
3.1.1 Arsitektur Sistem	17
3.1.2 Flowchart Dice Similarity.....	18
3.1.3 Flowchart Cosine Similarity	19
3.1.4 Flowchart Klasifikasi K-Nearest Neighbor	20
3.1.5 IPO	20
3.1.6 CDM	22
3.1.7 PDM.....	22
3.2 Data Set	23
3.3 Perhitungan Manual.....	23

3.3.1 Perhitungan TF-IDF.....	23
3.3.2 Perhitungan <i>Similarity Measure</i> dengan <i>Dice Similarity</i>	27
3.3.3 Proses Klasifikasi <i>K-Nearest Neighbor</i>	27
3.4 Tahapan Penelitian	28
3.4.1 Metodologi Penelitian.....	28
3.5 Pengujian.....	29
3.5.1 Rencana Pengujian.....	29
3.5.2 Pengujian Ketepatan Perangkat Lunak	30
BAB IV HASIL DAN PEMBAHASAN	39
4.1 Lingkup Uji Coba	39
4.2 Data	39
4.3 Pembangunan Sistem	39
4.3.1 Preprocessing	39
4.3.2 Menghitung TF-IDF	40
4.3.3 Dice Similarity	41
4.3.4 K-Nearest Neighbor	43
4.4 Pengujian.....	45
4.4.1 Proses Pengujian	45
4.4.2 Hasil Pengujian dan analisis	47
4.5 Grafik Analisa Uji Coba.....	48
BAB V KESIMPULAN	49
5.1 Kesimpulan.....	49
5.2 Saran	49
REFERENSI.....	50

DAFTAR TABEL

Tabel 2.1 Hasil <i>Tokenization</i>	6
Tabel 2.2 Contoh <i>Stopword</i>	7
Tabel 2.3 Contoh <i>Stemming</i>	7
Tabel 2.4 Penelitian Terkait	14
Tabel 3.1 Data Training Pariwisata.....	23
Tabel 3.2 Data Testing Pariwisata	23
Tabel 3.3 Hasil <i>Term-Frekuensi</i> Data Training Pariwisata	23
Tabel 3.4 <i>Term-Frequency</i> Data Testing Pariwisata.....	24
Tabel 3.5 Hasil IDF Data Training Pariwisata.....	24
Tabel 3.6 Hasil IDF Data Testing Pariwisata	25
Tabel 3.7 Hasil perhitungan TF-IDF Data Training Pariwisata.....	26
Tabel 3.8 Hasil perhitungan TF-IDF Data Testing Pariwisata	26
Tabel 3.9 Hasil <i>Dice Similarity</i>	28
Tabel 3.10 Tabel Akurasi	30
Tabel 3.11 Hasil Perhitungan TF (<i>term-frequency</i>) data training manual	31
Tabel 3.12 Hasil Perhitungan TF (<i>term-frequency</i>) data training sistem.....	31
Tabel 3.13 Hasil Perhitungan TF (<i>term-frequency</i>) data testing manual	32
Tabel 3.14 Hasil Perhitungan TF (<i>term-frequency</i>) data testing sistem.....	33
Tabel 3.15 Hasil Perhitungan IDF (<i>Inverse Documents frequency</i>) data training manual	33
Tabel 3.16 Hasil Perhitungan IDF (<i>Inverse Documents frequency</i>) data training sistem	34
Tabel 3.17 Hasil Perhitungan IDF (<i>Inverse Documents frequency</i>) data testing manual	35
Tabel 3.18 Hasil Perhitungan IDF (<i>Inverse Documents frequency</i>) data testing sistem	35
Tabel 3.19 Hasil Perhitungan TF-IDF (<i>term frequency-Inverse Documents frequency</i>) data training manual.....	36
Tabel 3.20 Hasil Perhitungan TF-IDF (<i>term frequency-Inverse Documents frequency</i>) data training sistem	36
Tabel 3.21 Hasil Perhitungan TF-IDF (<i>term frequency-Inverse Documents frequency</i>) data testing manual	37
Tabel 3.22 Hasil Perhitungan TF-IDF (<i>term frequency-Inverse Documents frequency</i>) data testing sistem	38
Tabel 3.23 Hasil Perhitungan <i>Dice Similarity</i> manual.....	38
Tabel 3.24 Hasil Perhitungan TF-IDF <i>Dice Similarity</i> sistem.....	38
Tabel 4.1 confusion matrix	46
Tabel 4.2 hasil prediksi dice similarity	39
Tabel 4.3 hasil prediksi cosine similarity.....	44
Tabel 4.4 Hasil Rekapitulasi nilai <i>precision, recall, f-measure</i> dan waktu pada <i>dice similarity</i> dan <i>cosine similarity</i>	47

DAFTAR GAMBAR

Gambar 2.1 Ilustrasi Cosine Similarity	9
Gambar 2.2 Ilustrasi Dice Similarity	10
Gambar 2.3 Ilustrasi K-Nearest Neighbor	11
Gambar 2.4 Ilustrasi <i>K-Fold Cross Validation</i>	12
Gambar 3.1 Arsitektur Sistem.....	17
Gambar 3.2 <i>Flowchart Dice Similarity</i>	18
Gambar 3.3 <i>Flowchart Cosine Similarity</i>	19
Gambar 3.3 <i>Flowchart</i> Klasifikasi <i>K-Nearest Neighbor</i>	20
Gambar 3.4 <i>Input Proses Output</i>	21
Gambar 3.5 CDM.....	22
Gambar 3.6 PDM	22
Gambar 4.1 Perbandingan nilai rata-rata dari <i>precision</i> , <i>recall</i> dan <i>f-measure</i> antara <i>cosine similarity</i> dan <i>dice similarity</i>	48