Analisis Konsep Himpunan dalam Pengolahan Data Seleksi Panitia P3RI Salman ITB

Melati Anggraini - 13522035
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
13522035@itb.ac.id

***Abstract*—Penyeleksian merupakan suatu hal yang krusial dalam sebuah organisasi, lembaga, ataupun perusahaan sebagai upaya regenerasi yang positif. Salah satu tahap yang dilalui dalam proses penyeleksian adalah pengolahan data. Seseorang bisa menggunakan berbagai cara dalam pengolahan data. Di zaman modern yang seperti sekarang, banyak tersedia *tools-tools* yang membantu seseorang dalam mengolah data. Akan tetapi, meskipun sudah banyak tools yang digunakan, namun tak jarang implementasinya masih terpaksa melakukannya secara manual. Apalagi jika pertanyaan bersifat uraian. Dari segi waktu, ini akan menghabiskan waktu begitu pula dari segi energi. Oleh karena itu dibutuhkan sebuah metode tambahan dalam pengolahan data. Pada makalah, akan diuji dua metode pengukuran kesamaan string yang memanfaatkan himpunan untuk mengatahui apakah sebuah jawaban responden cocok dengan indikator yang sudah ditetapkan oleh entitas tersebut melalui perhitungan indeks jaccard dan koefisien tumpang tindih.**

***Keywords*—Himpunan, Indeks Jaccard, Koefisien Tumpang Tindih, Seleksi**

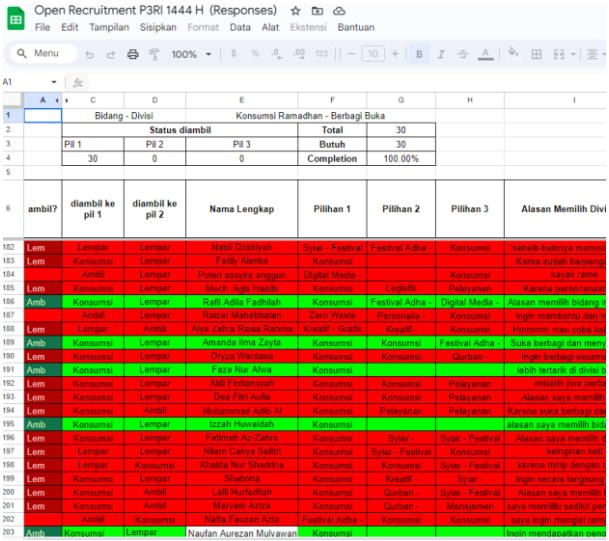
I. PENDAHULUAN

[1] Secara umum, seleksi dapat didefinisikan sebagai sebuah proses memilih atau memilah para pelamar kerja di perusahaan yang sesuai dengan persyaratan atau kualifikasi yang dibutuhkan perusahaan, yang kemudian akan ditempatkan pada posisi tertentu sesuai dengan kebutuhan perusahaan (Garaika dan Margahana,2019).

Dalam proses penyeleksian karyawan oleh sebuah perusahaan ataupun kepanitiaan menjadi langkah penting. Dalam era perkembangan teknologi informasi yang pesat, banyak *tools-tools* pengolahan data yang membantu mempercepat pengolahan data. Akan tetapi meskipun pengolahan data sudah tidak manual lagi, kenyataan yang terjadi di lapangan adalah proses perekrutan masih saja manual yang harus diteliti satu-persatu oleh seorang perekrut. Hal ini tentunya sangat tidak efisien apalagi untuk perusahaan atau organisasi dalam skala besar karena tidak efisien waktu dan energi.

Salah satu cara optimasi untuk memperbaiki efisiensi pengolahan data adalah dengan penggunaan konsep matematika diskrit melalui menggunakan konsep himpunan. Algoritmanya akan mengecek kesamaan semu sebuah *string* dengan menggunakan perhitungan indeks jaccard dan koefisien tumpang tindih.

Penelitian ini bertujuan untuk menerapkan konsep himpunan dalam pengolahan data seleksi panitia P3RI Salman ITB(1444 H) atau data panitia tahun lalu. Selain itu, juga melakukan analisis terhadap kinerja algoritma yang sudah dibuat sehingga selanjutnya bisa dipakai untuk bahan evaluasi dalam pengolahan data di tahun ini.



Gambar1. Penyeleksian "lempar" dan "ambil" secara manual

Melalui penelitian ini diharapkan dapat memberikan kontribusi positif terhadap pengembangan proses seleksi panitia P3RI Salman ITB, serta menggambarkan potensi penerapan konsep himpunan dalam pengolahan data seleksi karyawan secara lebih umum. Dengan demikian, perusahaan, lembaga ,ataupun organisasi dapat memanfaatkan pendekatan ini untuk meningkatkan efisiensi dan efektivitas dalam pengambilan keputusan terkait sumber daya manusia.

II. TEORI DASAR

A. Himpunan

[6]Himpunan(set) adalah kumpulan objek yang berbeda. Dalam konsep matematika, himpunan didefinisikan sebagai suatu kumpulan benda(objek) tertentu dengan batasan yang jelas, sehingga dengan tepat dapat diketahui objek yang termasuk himpunan maupun yang tidak.

Secara sederhana himpunan dinotasikan dengan menggunakan kurung kurawal. Kita juga dapat mendeskripsikan sebuah himpunan dengan lebih spesifik menggunakan notasi

{x|syarat keanggotaan x}. Contohnya adalah A = {x|x bilangan bulat lebih besar dari 5} atau A = { x | x > 5 }.

Simbol	Nama	Contoh	Penjelasan
{}	Kurung kurawal	A = {1,3} B = {2,3,9}	Membungkus elemen-elemen himpunan.
∩	Irisan	A ∩ B = {2}	Elemen-elemen yang sama dari kedua himpunan.
∪	Gabungan	A ∪ B = {1,2}	Gabungan elemen-elemen dua himpunan.
⊆	Himpunan Bagian	{1,3} ⊆ B	{1,3} himpunan bagian B.
ϵ	elemen	3 ϵ A	3 elemen dari himpunan A.
∉	Bukan elemen	3 ∉ A	3 bukan elemen dari himpunan A.

Gambar2. Notasi dalam himpunan

B. Operasi Himpunan

Notation	Name	Meaning
A∪B	Union	Elements that belong to set A or set B or both A and B
A∩B	Intersection	Elements that belong to both set A and set B
A⊆B	Subset	Every element of set A is also in set B
A⊂B	Proper subset	Every element of A is also in B, but B contains more elements
A⊄B	Not a subset	Elements of set A are not elements of set B
A=B	Equal sets	Both set A and B have the same elements
A^c or A′	Complement	Elements not in set A but in the universal set
A-B or A\B	Set difference	Elements in set A but not in set B
P(A)	Power set	The set of all subsets of set A
A×B	Cartesian product	The set that contains all the ordered pairs from set A and B in that order
n(A) or |A|	Cardinality	The number of elements in set A

Gambar3. Operasi dalam Himpunan

Sumber : https://www.storyofmathematics.com/set-notation/

Seperti yang terlihat pada *Gambar3*, himpunan memiliki beberapa operasi. Namun, dalam penelitian ini, penulis hanya menggunakan dua operasi himpunan yaitu operasi gabungan dan operasi irisan.

1. Operasi Gabungan
   Gabungan himpunan A dan B dinotasikan A ∪ $B = \{x|x \in A \; atau \; x \in B\}$ B adalah himpunan semua unsur yang termasuk anggota A atau anggota B.

$$A \cup B = \{x|x \in A \; atau \; x \in B\}$$

2. Irisan
   Irisan himpunan A dan B, dinotasikan A ∩ $B = \{x|x \in A \; dan \; x \in B\}$ B adalah himpunan yang elemen-elemenya merupakan anggota dari A dan B.

$$A \cap B = \{x|x \in A \; dan \; x \in B\}$$

C. Ukuran Kesamaan Teks Berbasis Istilah

1) Kesamaan Jaccard (*Jaccard* similarity)

[4] Kesamaan Jaccard adalah salah satu metode untuk mengukur kesamaan teks yang menghitung jumlah istilah/term bersama terhadap jumlah semua istilah/term yang unik dari kedua string. Atau bisa dirumuskan sebagai berikut :

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

2) Koefisien Dice's

[4] Koefisien Dice's (Dice's coefficient) merupakan ukuran yang didefinisikan sebagai dua kali jumlah suku umum pada string yang dibandingkan dibagi dengan jumlah total elemen pada kedua string. Dirumuskan sebagai berikut:

$$Dice\ Coefficient = \frac{2 \times |A \cap B|}{|A|+|B|}$$

3) Koefisien Pencocokan

[4] Koefisien Pencocokan (*Matching Coefficient*) adalah pendekatan berbasis vektor yang sangat sederhana yang hanya menghitung jumlah elemen yang mirip pada kedua vektor yang bukan nol. Dirumuskan sebagai berikut :

$$Matching\ Coefficient = \frac{|A \cup B|}{|A \cap B|}$$

4) Koefisien Tumpang Tindih

[4] Koefisien Tumpang Tindih(*Overlap coefficient*) mirip dengan koefisien Dice's akan tetapi pengukuran ini menganggap dua string cocok sepenuhnya jika salah satu adalah merupakan bagian dari yang lain. Dirumuskan seperti berikut :

$$Overlap\ Coefficient = \frac{|A \cap B|}{\min(|A|,|B|)}$$

D. Data Cleaning

Sebelum data responden calon panitia dianalisis, data harus terlebih dahulu dibersihkan. Berikut langkah-langkah pembersihan yang penulis lakukan :

1) Penghapusan spasi
   Penghapusan spasi dan kolom yang ingin dianalisis dengan fungsi *str.replace()* dilakukan untuk pembersihan data dan memastikan konsistensi format.
2) Mengubah semua karakter menjadi *lowercase*
   Langkah ini dilakukan agar teks lebih konsisten sehingga keakuratannya bisa optimal.
3) *Stemming*
   Stemming adalah menghilangkan imbuhan atau akhiran kata sehingga hanya menyisakan bentuk dasarnya saja.

III. PEMBAHASAN

A. *Penyeleksian Berdasarkan Kalimat yang paling Relevan untuk Suatu Divisi*

Dalam penelitian ini, penulis melakukan proses penyeleksian panitia dengan menggunakan koefisien tumpang tindih dan indeks Jaccard berdasarkan pertanyaan dasar seperti *"Apa alasan kamu memilih divisi pertama?"*. Penyeleksian tahap pertama ini dilakukan dengan membandingkan dengan kalimat yang paling memenuhi kebutuhan suatu divisi dan telah ditentukan . Artinya jika alasan responden jika memiliki nilai Indeks Jaccard ataupun Overlap yang tinggi maka responden akan lolos ke penyeleksian selanjutnya. Tahap ini dilakukan untuk mempermudah recruter atau panitia agar mereka tidak perlu mengecek semua data respons, mereka hanya mengecek data yang sudah disaring melalui metode kesamaan indeks jaccard dan kesamaan tumpang tindih.

Sebelum membandingkan string-string dalam dataset, pembersihan data perlu dilakukan agar data memiliki format dan kekonsistenan yang baik. Dalam penelitian ini penulis melakukan penghapusan terhadap karakter spasi serta mengubah karakternya menjadi *lowercase()*.

```
# Menghapus spasi pada kolom "Alasan Memilih Bidang/Divisi Pertama? pada divisi"
data.rename(columns=lambda x: x.strip() if isinstance(x, str) else x, inplace=True)

column_name = 'Alasan Memilih Bidang /Divisi Pertama ?'

# Remove spaces and convert to lowercase in the specified column
data[column_name] = data[column_name].str.replace(' ', '').str.lower()

# Now, the values in the specified column have spaces removed and are in lowercase
data[column_name]
```

✓ 0.1s

```
0       saatmasihkuliahqadarullahsayamenjadikoordinato...
1       karenasayasukaberinteraksidenganduniasosmed,di...
2       setiapmenjadipanitiakonsumsientahkenapapunyake...
3       karenasayasukamemotretsuasanadenganaestheticte...
4       inginmencurahkanpotensiyangsudahdimilikisertam...
                              ...
1013    karenasayainginikutsertauntukmenyemarakkansuas...
1014    sayasenangmengikutikegiatansosialsepertiberbag...
1015    sayatertarikdengandivisitersebutdancukupmemumpuni
1016    bismillah.divisiinspirasiramadhanmerupakansebu...
1017    sepertiyangtelahsayajelaskanpadacaptiontwibbon...
Name: Alasan Memilih Bidang /Divisi Pertama ?, Length: 1018, dtype: object
```

Gambar4 Implementasi Data Cleaning

Setelah data dibersihkan penulis menghilangkan imbuhan-imbuhan atau akhiran kata sehingga menyisakan bentuk dasarnya melalui proses stemming.

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

✓ 0.2s

```
stemmer = StemmerFactory().create_stemmer()
```

✓ 0.6s

```
def stemming(string):
    try:
        string = stemmer.stem(string)
    except:
        pass
    return string

data['Alasan Memilih Bidang /Divisi Pertama ?'] = data['Alasan Memilih Bidang /Divisi Pertama ?'].apply(stemming)
```

✓ 4m 54.7s

Gambar5. Implementasi Stemming

Kemudian melakukan pembersihan data terhadap baris yang memiliki nilai yang hanya terdiri dari spasi serta menghapus baris-baris yang memiliki nilai NaN pada kolom "Alasan Memilih Bidang/Divisi Pertama?".

```
df = data.replace(r'^s*$',float('NaN'), regex=True)
df.dropna(subset = ['Alasan Memilih Bidang /Divisi Pertama ?'], inplace = True)
df_reset = df.reset_index(drop=True)
```

✓ 0.2s

Gambar6. Implementasi pembersihan data yang bernilai NaN

```
def calculateJaccardOverlap(string1, string2):
    # Convert each string into a set of characters
    set1 = set(string1.replace(' ', '').lower())
    set2 = set(string2.replace(' ', '').lower())

    # Calculate Jaccard Index
    intersection = len(set1 & set2)
    union_size = len(set1 | set2)
    jaccard_index = intersection / union_size

    # Calculate Overlap Coefficient
    overlap_coefficient = intersection / min(len(set1), len(set2))

    return jaccard_index, overlap_coefficient

# List Indikator perdivisi
indikator_buka = ("kemampuanorganisasiketerampilankomunikasiketerlibatandalamkegiatansosialpeng
indikator_takjil = ("kemampuanorganisasipengetahuantentangvariasitakjilketerampilankomunikasike
indikator_muslimah = ("kemampuanorganisasiuntukacarakhususmuslimahkreativitasdalampenyelenggara
indikator_itikaf = ("kemampuanorganisasiuntukmenyediakanlayananselamaitikafketerampilankomunika
indikator_exhib = ("kemampuanorganisasiuntukpamerankreativitasdalampenyelenggaraanfestivalketer
# PENERAPAN
# Menginisialisasi dataset baru untuk menyimpan hasil seleksi
df["Jaccard_Index"] = 0.0
df["Overlap_Coefficient"] = 0.0
dataset_indexjacard = []
dataset_koeftumpangtindih = []
dataset_indexjacard2 = []
dataset_koeftumpangtindih2 = []
dataset_indexjacard3 = []
dataset_koeftumpangtindih3 = []
dataset_indexjacard4 = []
dataset_koeftumpangtindih4 = []

df.reset_index(drop=True, inplace=True)

# Seleksi dan perbandingan
for index, selected_value in enumerate(df["Alasan Memilih Bidang /Divisi Pertama ?"]):
    # print(df["Pilihan Pertama Bidang -  Divisi"][index])
    # print(selected_value)
    if index < len(df) and df["Pilihan Pertama Bidang -  Divisi"][index] == "Konsumsi Ramadhan
        # print(i)
        # print(df["Pilihan Pertama Bidang -  Divisi"][index] )
        jaccard_index, overlap_coefficient = calculateJaccardOverlap(indikator_buka, selected_v

        # Update the new columns
        df.at[index, "Jaccard_Index"] = jaccard_index
        df.at[index, "Overlap_Coefficient"] = overlap_coefficient
```

Gambar7. Implementasi Indeks Jacard dan Koefisien Tumpang Tindih

Data responden calon panitia P3RI Salman ITB untuk peminat divisi tertinggi mencapai 108 orang yang mendaftar. Terdapat 55 divisi dengan rata-rata jumlah pendaftarnya 23 orang. Akan sangat menguras banyak waktu panitia untuk membaca satu persatu alasan mereka sesuai indikator yang diberikan ke divisi masing-masing. Belum lagi setelah dilakukan pengecekan pertanyaan dasar, mereka diminta untuk mengisi pertanyaan-pertanyaan khusus divisi tersebut yang kurang lebih terdapat 7 pertanyaan lagi yang harus dianalisis dan dilakukan penskoran. Ini akan menguras banyak waktu untuk mengoreksi satu persatu sebanyak(1008 data). Sehingga dengan dilakukannya penyaringan berdasarkan pertanyaan dasar, harapannya rekruter atau panitia hanya membaca satu persatu data yang jumlahnya sudah diolah melalui algoritma yang telah dibuat penulis.

Terakhir, penulis membuat variabel indikator_[namadivisi] yang menyimpan kalimat paling relevan yang nantinya digunakan sebagai indikator kelolosan seseorang dari masing-masing divisi.

Dalam penelitian ini penulis hanya mengolah 4 divisi yang paling banyak diminati sebagai sampel yang bisa dilihat dari *Gambar4*.

```
# Mendapatkan total pendaftar masing-masing divisi
nilai_unik = df['Pilihan Pertama Bidang -  Divisi'].value_counts()
print(nilai_unik[:5])

# Menghitung nilai unik dan mengurutkannya secara descending
nilai_unik_count = df['Pilihan Pertama Bidang -  Divisi'].value_counts().sort_values(ascending=False)
array_of_string_desc = nilai_unik_count.index.tolist()
✓ 0.4s

Pilihan Pertama Bidang -  Divisi
Konsumsi Ramadhan - Berbagi Buka    108
Konsumsi Ramadhan - Takjil           90
Syiar - Festival Ramadan             72
Syiar - Ramadhan Muslimah            55
Pelayanan Jama'ah - I'tikaf          54
Name: count, dtype: int64
```

Gambar8. Lima divisi dengan pendaftar terbanyak

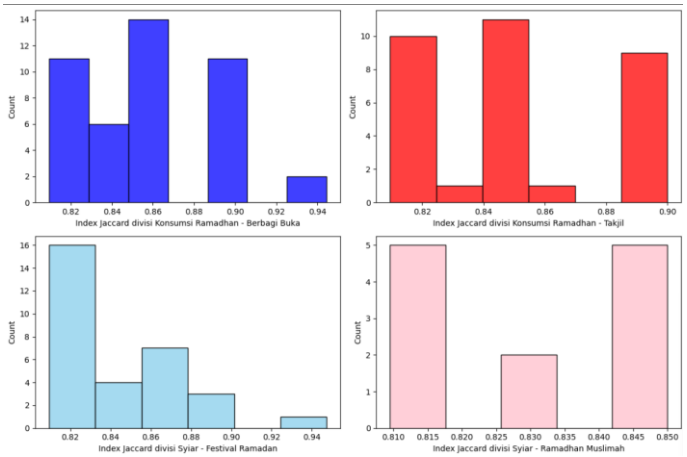
```
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Konsumsi Ramadhan - Berbagi Buka > 0.8:
panitia yang lolos sebanyak : 44
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Konsumsi Ramadhan - Takjil > 0.8:
panitia yang lolos sebanyak : 32
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Syiar - Festival Ramadan > 0.8:
panitia yang lolos sebanyak : 31
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Syiar - Ramadhan Muslimah > 0.8:
panitia yang lolos sebanyak : 12
```

Gambar9. Jacard Indeks empat divisi teratas

```
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Konsumsi Ramadhan - Berbagi Buka >
0.8:
panitia yang lolos sebanyak: 107
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Konsumsi Ramadhan - Takjil > 0.8:
panitia yang lolos sebanyak: 89
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Syiar - Festival Ramadan > 0.8:
panitia yang lolos sebanyak: 72
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Syiar - Ramadhan Muslimah > 0.8:
panitia yang lolos sebanyak: 54
```

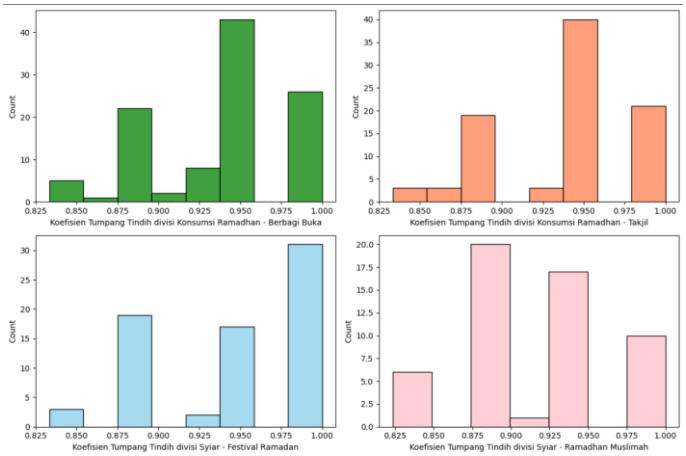
Gambar10. Overlap Coefficient empat divisi teratas

Berdasarkan *Gambar8*, *Gambar9*, dan *Gambar10* terlihat bahwa dengan menggunakan penyeleksian menggunakan indeks jacard dan koefisien tumpang tindih serta dilakukan penetapan kemiripan yaitu lebih dari 80%. Diperolah hasil bahwa data yang memiliki jawaban mirip dengan indikator masing-masing divisi akan terseleksi dan lolos penyeleksian.



Gambar11. Distribusi Koefisien Indeks Jacard

Berdasarkan *Gambar11* histogram untuk indeks jaccard cenderung condong ke kiri. Artinya, kalimat yang berada dalam kolom data tersebut memiliki tingkat kemiripan rata-rata 80%. Selain itu, juga tidak ada yang memiliki kemiripan 100%, hal ini dikarenakan tidak mungkin terjadi duplikasi teks atau penyebaran indikator dari sebuah entitas terhadap calon karyawanya. Angka ini juga berbanding lurus dengan kebutuhan akan sumber daya yang dibutuhkan serta terbagi merata. Jika kita tinjau dari rumus indeks jaccard, kita tahu bahwa Indeks Jacard ini bisa mendeteksi kesamaan irisan kedua kalimat relatif terhadap gabungan kalimat sehingga terjadi pemerataan. Serta bisa disimpulkan juga indeks jaccard bisa mendeteksi suatu string yang memiliki duplikasi semu(sebagian).



Gambar12. Distribusi Koefisien Tumpang Tindih

Berdasarkan histogram pada *Gambar12* di atas koefisien tumpang tindih grafiknya cenderung condong ke kanan. Artinya banyak kalimat yang terdeteksi mirip. Hasil perhitungan memberikan nilai seperti berikut menyeleksi divisi Berbagi Buka sebanyak 107/108, itu artinya 99% kalimatnya dianggap sama, pada divisi Takjil 89/90 (98,9%), divisi Festival Ramadhan 72/72 (100%) dan divisi Ramadhan Muslimah 54/55(98,1%). Nilai-nilai yang diberikan memiliki rata-rata kemiripan yang sangat tinggi(99%). Hal ini terjadi karena dari perhitungan untuk mendapatkan koefisien tumpang tindih sendiri cenderung akan menghasilkan nilai yang besar jika elemen-elemen kunci dari dua himpunan memiliki banyak kesamaan. Sehingga perhitungan ini lebih cocok untuk diimplementasikan pada fitur rekomendasi produk, deteksi spam, dll.

Bidang - Divisi	Konsumsi Ramadhan - Berbagi Buka			
Status diambil			**Total**	30
Pil 1	Pil 2	Pil 3	**Butuh**	30
30	0	0	**Completion**	100.00%

Bidang - Divisi	Konsumsi Ramadhan - Takjil			
Status diambil			**Total**	35
Pil 1	Pil 2	Pil 3	**Butuh**	35
34	1	0	**Completion**	100.00%

Bidang - Divisi	Syiar - Festival Ramadan			
Status diambil			**Total**	17
Pil 1	Pil 2	Pil 3	**Butuh**	17
17	0	0	**Completion**	100.00%

Bidang-Divisi	Syiar - Ramadhan Muslimah			
Status diambil			**Total**	15
Pil 1	Pil 2	Pil 3	**Butuh**	15
12	2	1	**Completion**	100.00%

Gambar13. Jumlah Panitia yang Lolos melalui Penyeleksian Manual

Didukung juga dengan hasil kecocokan pada hasil olah data yang dilakukan secara manual yang telah dilakukan pada *Gambar13* di bawah ini. Penyeleksian dari indeks jaccard nilainya mendekati nilai hasil olah data secara manual. Setelah data diseleksi menggunakan indeks jaccard menghasilnya 44 panitia tersisa pada divisi berbagi buka, 32 panitia tersisa pada

divisi Takjil, 31 panitia tersisa dari divisi Festival Ramdhan, serta 12 panitia tersisa dari divisi Ramadhan Muslimah. Jika kita bandingkan dengan jumlah kebutuhan nilainya mendekati dengan kebutuhan sumber daya panitianya yang bisa kita lihat dalam tabel di bawah ini. Akan tetapi, pada divisi Takjil menghasilkan nilai yang terseleksi lebih sedikit daripada kebutuhan. Hal ini sangat mungkin terjadi jika memang indikator yang diberikan recruter atau panitia memiliki perbedaan yang cukup signifikan. Nilai patokan ini bersifat tentatif dan sangat tergantung kebutuhan. Sehingga, untuk menangani kasus ini, bisa menurunkan atau menaikkan nilai patokan dari indeks jaccard itu sendiri.

Persebaran nilai dari indeks jaccard memberikan persebaran yang lebih baik daripada dengan menggunakan koefisien tumpang tindih. Gambaran kesamaan relatif secara keseluruhan yang dihasilkan oleh indeks jaccard antara dua himpunan ini sangat cocok untuk penerapan proses penyeleksian data.

B. *Penyeleksian Berdasar Kalimat yang Paling Tidak Relevan untuk Suatu Divisi*

Cara kedua yang penulis lakukan untuk melakukan penyeleksian secara otomatis adalah dengan mengidentifikasi kalimat yang paling tidak relevan dengan suatu divisi untuk nantinya penulis mengambil kelolosan panitia berdasarkan nilai indeks jaccard dan koefisien tumpang tindih kurang dari 80%. Penulis menyimpan kalimat yang tidak sesuai tersebut dalam sebuah variabel *lempar_[nama_divisi]* yang nantinya digunakan sebagai bahan perbandingan dengan kalimat responden. Berikut hasil penyeleksian berdasarkan kalimat tidak yang tidak relevan yang bisa dilihat dari *Gambar14* dan *Gambar15*.

```
IDENTIFIKASI BERDASARKAN KALIMAT TIDAK RELEVAN
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Konsumsi Ramadhan - Berbagi Buka < 0.8:
panitia yang lolos sebanyak : 94
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Konsumsi Ramadhan - Takjil < 0.8 :
panitia yang lolos sebanyak : 83
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Syiar - Festival Ramadan < 0.8:
panitia yang lolos sebanyak : 68
DataFrame baru dari hasil seleksi dengan Indeks Jacard divisi Syiar - Ramadhan Muslimah < 0.8:
panitia yang lolos sebanyak : 51
```

Gambar14. Jumlah Panitia yang Tidak Lolos berdasarkan Indeks Jaccard

```
IDENTIFIKASI BERDASARKAN KALIMAT TIDAK RELEVAN
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Konsumsi Ramadhan - Berbagi Buka < 0.8:
panitia yang lolos sebanyak: 9
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Konsumsi Ramadhan - Takjil < 0.8:
panitia yang lolos sebanyak: 8
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Syiar - Festival Ramadan < 0.8:
panitia yang lolos sebanyak: 2
DataFrame baru dari hasil seleksi dengan Overlap Coefficient divisi Syiar - Ramadhan Muslimah < 0.8:
panitia yang lolos sebanyak: 3
```

Gambar15. Jumlah panitia yang Tidak Lolos Berdasarkan Koefisien Tumpang Tindih

C. *Rasio Kesamaan antara Indikator Kalimat Yang Paling Relevan dengan yang Paling Tidak Relevan*

Setelah mendapatkan hasil data seleksi berdasarkan kalimat yang paling relevan dan yang paling tidak relevan, penulis menyamakan kedua hasilnya dan menghitung rasio kesamaannya menggunakan indeks jaccard yang tertera pada *Gambar16*, *Gambar17*, dan *Gambar18*.

```python
def hitung_rasio_kesamaan(list1, list2):
    # Menghitung elemen unik dari masing-masing list
    set_list1 = set(list1)
    set_list2 = set(list2)

    # Menghitung jumlah elemen irisan
    jumlah_elemen_irisan = len(set_list1 & set_list2)

    # Menghitung jumlah elemen gabungan
    jumlah_elemen_gabungan = len(set_list1 | set_list2)

    # Menghitung indeks Jaccard
    rasio_kesamaan = jumlah_elemen_irisan / jumlah_elemen_gabungan

    return rasio_kesamaan
```

Gambar16. Implementasi Rasio Kesamaan

```
Rasio Kesamaan divisi Konsumsi Ramadhan - Berbagi Buka : 87.88%
Rasio Kesamaan divisi Konsumsi Ramadahn - Takjil : 89.66%
Rasio Kesamaan divisi Syiar - Festival Ramadhan : 97.10%
Rasio Kesamaan divisi Syiar - Ramadhan Muslimah : 92.73%
```

Gambar17. Rasio Kesamaan Indeks Jaccard

```
Rasio Kesamaan divisi Konsumsi Ramadhan - Berbagi Buka : 44.44%
Rasio Kesamaan divisi Konsumsi Ramadahn - Takjil : 87.50%
Rasio Kesamaan divisi Syiar - Festival Ramadhan : 100.00%
Rasio Kesamaan divisi Syiar - Ramadhan Muslimah : 100.00%
```

Gambar18. Rasio Kesamaan Koefisien Tumpang Tindih

Berdasarkan *Gambar17* dan *Gambar18* , kesamaan yang diperoleh dari perhitungan indeks jaccard menghasilkan rasio yang lebih baik dan konsisten yaitu 91,84% dibandingkan rasio kesamaan koefisien tumpang tindih yaitu 82,99%. Sehingga bisa disimpulkan indeks jaccard memberi hasil deteksi kesamaan kalimat lebih baik dalam penyeleksian data.

IV. KESIMPULAN

Melalui analisis yang mendalam mengenai pengolahan data seleksi panitia P3RI Salman ITB membawa penulis dalam mencapai kesimpulan bahwa Indeks Jaccard sebagai metrik evaluasi kesamaan teks lebih tepat digunakan dibandingkan dengan koefisien tumpang tindih.

Indeks jaccard, dengan fokus pada elemen-elemen yang dimiliki oleh dua himpunan yang dibandingkan relatif terhadap gabungan dua himpunan, terbukti lebih adaptif dalam mengukur kesamaan antar data calon panitia(responden). Hasil eksperimen dan analisis data juga menunjukkan bahwa indeks jaccard memberikan hasil yang lebih konsisten dan relevan dalam konteks penyeleksian data.

Oleh karena itu, penulis merekomendasikan penggunaaan Indeks Jaccard sebagai metode tambahan untuk mengevaluasi dan menyaring data calon panitia. Impementasi indeks jaccard dapat meningkatkan efisiensi dan akurasi proses penyeleksian data, serta memberikan dasar yang kuat untuk pengambilan keputusan yang informasional.

Kesimpulan ini diharapkan dapat memberikan kontribusi positif dalam pengembangan sistem seleksi panitia P3RI Salman ITB dan dapat menjadi landasan untuk penelitian dan pengembangan lebih lanjut.

V. APENDIKS

Implementasi kode dapat dilihat pada link berikut :
https://github.com/mlatia/Analisis-Konsep-Himpunan-dalam-Pengolahan-Data-Kepanitiaan

Demi menjaga keamanan data, penulis akan memprivasi repository github setelah penilaian makalah ini dilakukan.

VI. Ucapan Terima Kasih

Dengan kerendahan hati penulis, penulis ingin menyampaikan ucapan terima kasih kepada Allah dan semua pihak yang berkontribusi dalam penyelesaian makalah ini. Terima kasih kepada Ibu Dr. Nur Ulfa Maulidevi, selaku dosen pengajar mata kuliah IF2120 Matematika Diskrit K01 yang telah membimbing penulis dan teman-teman selama perkuliahan Matematika Diskrit Tahun Ajaran 2023/2024.

Ucapan terima kasih kepada Panitia P3RI 1444H yang telah mengizinkan penulis melakukan analisis dan evaluasi data guna pemenuhan tugas makalah dengan topik Analisis Himpunan. Keberhasilan makalah ini juga tidak terlepas dari kerja sama dan bantuan dari semua pihak.

Terakhir, terima kasih kepada pembaca yang telah meluangkan waktunya untuk membaca makalah ini. Semoga makalah ini bisa memberikan manfaat dan kontribusi positif dalam bidang penerapan konsep matematika diskrit.

Daftar Pustaka

[1] Dwi Partini, Arni. (2022). Pentingnya seleksi SDM untuk calon Karyawan. Diakses pada (9 Desember 2023) https://digstraksi.com/pentingnya-seleksi-sdm-untuk-calon-karyawan/

[2] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547-579.

[3] Dice, L. (1945). Measures of the amount of ecologic association between species. Ecology, 26(3).

[4]https://download.garuda.kemdikbud.go.id/article.php?article=2449472&val=23384&title=Survei%20Terhadap%20Pengukuran%20Kesamaan%20Teks%20Survey%20of%20Text%20Similarity%20Measurement/ , diakses pada 9 Desember 2023

[5] https://www.storyofmathematics.com/set-notation/ , diakses pada 9 Desember 2023

[6]https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/01-Himpunan(2023)-1.pdf, diakses pada 11 Desember 2023

Pernyataan

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 11 Desember 2023

Melati Anggraini-13522035