**Pembobotan *Vector Space Model* Korpus Twitter Tentang *Data Science Text Mining Text Retrieval* Menggunakan *Cosine Smiliarity***

**TUGAS KELOMPOK**

**Disusun Untuk Memenuhi Tugas Mata Kuliah Temu Kembali Informasi**

**Dosen Pengampu : Retnani Latifah, M.Kom**

**Disusun Oleh :**

|  |  |
| --- | --- |
| **MUHAMMAD REZA** | **2019470055** |
| **SELAMET SAPUTRA** | **2019470069** |
| **SYECHAN AHMAD ZIDAN** | **2019470110** |

PROGRAM STUDI TEKNIK INFORMATIKAFAKULTAS TEKNIKUNIVERSITAS MUHAMMADIYAH JAKARTA2022

# **DAFTAR ISI**

[**DAFTAR ISI** ii](#_Toc120165842)

[**DAFTAR GAMBAR** iii](#_Toc120165843)

[**DAFTAR TABEL** iv](#_Toc120165844)

[**LAMPIRAN** v](#_Toc120165845)

[**BAB I** 1](#_Toc120165846)

[**1.1.** **Latar belakang masalah** 1](#_Toc120165847)

[**1.2.** **Identifikasi Masalah** 2](#_Toc120165848)

[**1.3.** **Rumusan Masalah** 2](#_Toc120165849)

[**1.4.** **Batasan Masalah** 2](#_Toc120165850)

[**BAB II** 3](#_Toc120165851)

[**2.1.** **Data Acquisition** 3](#_Toc120165852)

[**2.2** **Data Exploration** 4](#_Toc120165853)

[**2.3** ***Preprocessing*** 5](#_Toc120165854)

[**BAB III** 6](#_Toc120165855)

[**3.1.** ***Term Frequency Inverse Document Frequency*** 6](#_Toc120165856)

[**3.2.** ***Vector Space Model*** 8](#_Toc120165857)

[**BAB IV** 9](#_Toc120165858)

[**Kesimpulan** 9](#_Toc120165859)

[**DAFTAR PUSTAKA** 10](#_Toc120165860)

# **DAFTAR GAMBAR**

[Gambar korpus 1 3](#_Toc120165814)

[Gambar korpus 2 3](file:///E:\COLLEGE\SEMESTER%207\TEMU%20KEMBALI%20INFORMASI\1%20TUGAS%20KELOMPOK\tugas-tki\LAPORAN\Laporan_Kelompok_TKI.docx#_Toc120165815)

# **DAFTAR TABEL**

[Tabel 2.1 Perubahan teks 4](#_Toc120165834)

[Tabel 2.2 *stopwords* sering muncul 4](#_Toc120165835)

[Tabel 2.3 hasil preprocessing 5](#_Toc120165836)

[Tabel 3.1 *Term frequency* muncul di dokument 6](#_Toc120165837)

[Tabel 3.2 *Term frequency* tidakmuncul di dokument 6](#_Toc120165838)

[Tabel 3.3 tf-idf *query* 7](file:///E:\COLLEGE\SEMESTER%207\TEMU%20KEMBALI%20INFORMASI\1%20TUGAS%20KELOMPOK\tugas-tki\LAPORAN\Laporan_Kelompok_TKI.docx#_Toc120165839)

[Tabel 3.4 Hasil *Cosine Smiliarity* 8](#_Toc120165840)

# **LAMPIRAN**

Kode : [tugas-tki/Tugas.ipynb at main · ackermanjayjay/tugas-tki (github.com)](https://github.com/ackermanjayjay/tugas-tki/blob/main/code/Tugas.ipynb)

Data : [tugas-tki/data/data twitter/data about ai at main · ackermanjayjay/tugas-tki (github.com)](https://github.com/ackermanjayjay/tugas-tki/tree/main/data/data%20twitter/data%20about%20ai)

# **BAB I**

## **Latar belakang masalah**

Saat ini sedang gencarnya *data science, text mining,* dan *text retrieval. Data science* menurut David M. Blei dan Padhraic Smyth adalah turunan atau perhitungan menggunakan statistika untuk melakukan prediksi (Blei & Smyth, 2017). *Text mining* adalah pengolahan dari kumpulan document yang dipecah menjadi teks untuk mengetahui informasi yang bermanfaat dengan menggunakan perhitungan kalkulasi matematika (Sabrani et al., 2020). I*nformation retrieval* atau pengambilan informasi atau text retrieval adalah tugas untuk mengambil informasi yang sesuai atau relevan dari kumpulan korpus yang mewakili permintaan (kueri) (Djenouri et al., 2021).

*Text Mining* adanya *preprocessing* dan ekstraksi fitur, tahap *preprocessing*  terdiri dari *case folding,stop word removal, stemming,word normalization* untuk mengatasi *overfitting* dari hasil *stemming* (Ma’rifah et al., 2020). ekstrasi fitur dalam text mining ada *term* *frequency* melihat setiap kata yang muncul didalam dokumen atau korpus, untuk menghitung *inverse document* diperlukan masing-masing kemunculan *term frequency* di setiap document atau korpus lalu dikalkulasi dengan rumus *Inverse Document Frequency* (IDF) .Maka dari hasil nilai ekstrasi fitur tf-idf ini digunakan untunk perhitung similaritas, dan untuk beberapa metode dalam pendekatan statistika (Setyawan et al., 2021).

*Vector Space Model* (VSM). Sebuah model yang digunakan untuk mengukur sebuah kueri antara suatu dokumen dengan suatu kata kunci atau *keyword* (Susanti et al., 2020). *Vector space* adalah geometri berdimensi besar, ruang yang batas-batasnya ditentukan oleh *vector*. Konsep dasar *vector space model* adalah menghitung jarak vector antara dokumen dengan kata kunci yang dimasukkan kemudian mengurutkan berdasarkan tingkat kedekatannya (Susanti et al., 2020). Salah satu model *Vector Space Model* adalah *Cosine Smiliarity* untuk memodelkan *document* text sebagai vector kata, dengan menggunakan kesamaan antara dua dokumen (Ma’rifah et al., 2020).

## **Identifikasi Masalah**

Berdasarkan permsalahan di latar belakang, permasalahan tersebut dilakukan identifikasi sebagai berikut :

1. Melakukan *preprocessing* *document* teks
2. Melakukan rkstraksi fitur teks menggunakan *Term Frequency Inverse Document Frequency*
3. Menghitung jarak teks dokumen menggunakan *Cosine Smiliarity*

## **Rumusan Masalah**

Berdasarkan permasalahan diatas, akan dilakukan perumusan atau kajian sebagai berikut :

1. Bagaimana melakukan preprocessing data teks *document*  ?
2. Bagaimana cara melakukan ekstraksi fitur text dari *document* menggunakan *Term Frequency Inverse Document Frequency* ?
3. Bagaimana cara menghitung jarak teks *document* twitter menggunakan *Cosine Smiliarity*  ?

## **Batasan Masalah**

1. Mengolah teks *document* dari twiiter tentang *data science text mining text retrieval*
2. Menggunakan bantuan *library* regex*,*NLTK, Pandas, Numpy untuk mengolah text data dan menghitung hasil jarak kedekatan dokument
3. Menggunakan data dari *twitter* berjumlah 84 dokument hanya bahasa inggris
4. Mengetahui hasil kedekatan jarak antara *document*

# **BAB II**

## **2.1. Data Acquisition**

Data yang digunakan dalam laporan kali ini adalah data atau korpus yang  
diambil berasal dari twitter dengan cara *scrapping*, korpus yang diambil  
adalah tentang text mining dan *information retrieval*. Berjumlah 84 dokument

Text

Description automatically generated

Gambar korpus 1

Text

Description automatically generated

Gambar korpus 2

## **2.2 Data Exploration**

Karena dokument terpisah maka dilakukan penggabungan antara korpus 1 dan korpus 2, maka total data yang digabungkan berjulam 84 dokument. Dokument 4 terdapat kalimat bukan bahasa inggris maka dilakukan perbaikan agar kalimat tersebut menjadi bahasa inggris.

### Tabel 2.1 Perubahan teks

|  |  |
| --- | --- |
| Bukan bahasa inggris | Bahasa Inggris |
| Se puede crear gráficos espectaculares que incluyan los resultados de las pruebas estadísticas con rigor? statsplot lo vuelve simple y listo para publicar | Can you create spectacular graphics that include the results of rigorous statistical tests? |

Pada tabel 2.1 kalimat diubah menjadi bahasa inggris, agar pada teks *preprocessing* tidak perlu memusingkan stemming bahasa non inggris karena bahasa sebelum dilakukan perubahan adalah bahasa spanyol.

Dokument yang sudah digabungkan memiliki stopwords yang semuanya bahasa inggris dapat dilihat pada tabel 2.2

### Tabel 2.2 *stopwords* sering muncul

|  |  |
| --- | --- |
| *Stopwords* | *Frequensi* |
| to | 53 |
| in | 46 |
| and | 46 |
| a | 44 |
| the | 38 |
| of | 33 |
| is | 26 |
| for | 24 |
| you | 19 |
| on | 18 |

## **2.3 *Preprocessing***

Tahap *preprocessing* terdiri dari tahap :

1. *Casefolding*

Adalah tahap untuk mengecilkan huruf yang sebelumnya kapital, penghapusan tanda baca, penghapusan nomor dan link

1. *Stopword removal*

Adalah tahap untuk menghapus kata pengubung

1. *Stemming*

Adalah tahap untuk mereduksi kata menjadi kata dasar

1. *Word normalization*
2. Adalah tahap opsional yang digunakan untuk mencegah hasil overfitt pada saat tahap *Stemming*

Setelah tahap sudah ditentukan maha buat *pipeline* agar data dokument dapat di *preprocessing* yang dapat dilihat pada tabel 2.3

### Tabel 2.3 hasil preprocessing

|  |  |
| --- | --- |
| Sebelum *preprocessing* | Sesudah *preprocessing* |
| I think the message in Data Science needs to be: Don't believe everything you read. 🧵  #stats #datascience https://t.co/4jGMgmX8Nw | think message data science needs believe everything read stats datascience jgmgmx nw |

Hasil dari preprocessing pada Tabel 2.3 dapat dilihat bahwa kata seperti “I” dihapus pada saat preprocessing, serta link seperti “<https://t.co> “ juga dihapus karena ingin mengambil teksnya saja, untuk dilakukan pencarian informasi yang bermanfaat.

# **BAB III**

## **3.1. *Term Frequency Inverse Document Frequency***

Pada Tahap ini dokument yang sudah dilakukan *preprocessing* dipecah menjadi per kata, serta pada masing-masing kata yang muncul di dokument diberikan nilai 1, akan tetapi jika tidak muncul di dokument diberikan nilai 0. Pada tabel 3.1 dan tabel 3.2 adalah salah satu sampel term yang muncul di dokumen dan tidak muncul di dokumen.

### Tabel 3.1 *Term frequency* muncul di dokument

|  |  |  |
| --- | --- | --- |
| *Term*(kata) | *Dokument frequency* | *Frequency in dokument* |
| data | 23 | 1 |
| text | 44 | 1 |
| mining | 27 | 1 |
| retrieval | 18 | 1 |
| python | 11 | 1 |

### Tabel 3.2 *Term frequency* tidakmuncul di dokument

|  |  |
| --- | --- |
| *Term*(kata) | Frequensi |
| data | 0 |
| text | 0 |
| mining | 0 |
| retrieval | 0 |
| python | 0 |

Tahap *term* frequensi sudah dilakukan maka dilakukan *Inverse Document Frequency*(IDF) dengan rumus :

N = Jumlah dokumen

*Document frequency*(DF)= jumlah kemunculan term yang muncul di dokument

*Frequency Qorpus*(FQ)

Maka perhitungannya dengan menggunakan sampel dari tabel 3.1

= 0,563

Pada tahap *term frequency inverse document query* adalah tahap untuk mencari bobot berdasar kueri yang ingin ditentukan dilakukan dengan hasil *Inverse Document Frequency*(IDF), semisal kueri “data” dilakukan pencarian apakah data ada di *term* jika ada maka bernilai 1, sebagai berikut cara perhitungannya:

*idf(term)\** FQ

*idf(data)\*FQ*

*0,563\*1*

Pada tabel *term frequency document inverse query,* dengan kueri masukkan “data”, “text”, “mining” menghasilkan hasil pengujian yang sudah dilakukan preprocessing, bahwa term “data” menghasilkan *inverse document frequency* yang paling besar diantara ketiga term.

### Tabel 3.3 tf-idf *query*

|  |  |  |  |
| --- | --- | --- | --- |
| Term | DF | N/df | IDF |
| data | 23 | 3.652173913043478 | 0, 5625514500442887 |
| text | 44 | 1.9090909090909092 | 0.2808266095756942 |
| mining | 27 | 3.111111111111111 | 0.49291552190289434 |
| retrieval | 18 | 4.666666666666667 | 0.6690067809585756 |
| python | 11 | 7.636363636363637 | 0.8828866009036567 |

## **3.2. *Vector Space Model***

Setelah tahap *Term Frequency Inverse Document Frequency* dilakukan hasil bobot dari perhitungan tahap tfidf, dari masing-masing kueri di kalkulasikan menggunakan metode *Cosine Smiliarity* untuk mengetahui jarak *term* dari masing-masing document, metode *Cosine Smiliarity* menggunakan perhitungan sebagai berikut:

= bobot tf-idf *term i* sampai *j*

= bobot tf-idf *term i* sampai *q*

### Tabel 3.4 Hasil *Cosine Smiliarity*

|  |  |  |
| --- | --- | --- |
| *Query* | Dokument terkuat | Hasil smiliaritas |
| data text mining retrieval python | 20 | 0.1838423837595189 |

# **BAB IV**

## **Kesimpulan**

Kesimpulan mengenai data korpus yang digunakan, yaitu data dari media sosial twitter, yang kemudian digabungkan menggunakan *library pandas*, lalu dilakukan *preprocessing* terdiri dari *casefiolding, stemming* menggunakan algoritma porter*, remove stopwords* menggunakan NLTK*, word normalization,* melakukan ekstraksi menggunakan tf-idf dengan *library scikit learn* agar dapat dilakukan kalkulasi menghitung jarak *term* dari masing-masing dokumen menggunakan *vector space model* metode *cosine smiliarity,* serta menghasilkan dokumen paling terkait adalah dokumen 20.

# **DAFTAR PUSTAKA**

Blei, D. M., & Smyth, P. (2017). Science and data science. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 114, Issue 33, pp. 8689–8692). National Academy of Sciences. https://doi.org/10.1073/pnas.1702076114

Djenouri, Y., Belhadi, A., Djenouri, D., & Lin, J. C. W. (2021). Cluster-based information retrieval using pattern mining. *Applied Intelligence*, *51*(4), 1888–1903. https://doi.org/10.1007/s10489-020-01922-x

Ma’rifah, H., Wibawa, A. P., & Akbar, M. I. (2020). Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing. *Sains, Aplikasi, Komputasi Dan Teknologi Informasi*, *2*(2), 70. https://doi.org/10.30872/jsakti.v2i2.2681

Sabrani, A., Wedashwara W., I. G. W., & Bimantoro, F. (2020). Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia. *Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (JTIKA )*, *2*(1), 89–100. https://doi.org/10.29303/jtika.v2i1.87

Setyawan, C., Benarkah, N., & Prasetyo, V. R. (2021). Automatic Text Summarization Berdasarkan Pendekatan Statistika pada Dokumen Berbahasa Indonesia. *KELUWIH: Jurnal Sains Dan Teknologi*, *2*(1). https://doi.org/10.24123/saintek.v2i1.4045

Susanti, S., Azmi, M., Ali, E., Rahmaddeni, R., & Saputra Wijaya, Y. (2020). Perbandingan Boolean Model Dan Vector Space Model Dalam Pencarian Dokumen Teks. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *11*(2), 268–277. https://doi.org/10.31849/digitalzone.v11i2.4168

Blei, D. M., & Smyth, P. (2017). Science and data science. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 114, Issue 33, pp. 8689–8692). National Academy of Sciences. https://doi.org/10.1073/pnas.1702076114

Djenouri, Y., Belhadi, A., Djenouri, D., & Lin, J. C. W. (2021). Cluster-based information retrieval using pattern mining. *Applied Intelligence*, *51*(4), 1888–1903. https://doi.org/10.1007/s10489-020-01922-x

Ma’rifah, H., Wibawa, A. P., & Akbar, M. I. (2020). Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing. *Sains, Aplikasi, Komputasi Dan Teknologi Informasi*, *2*(2), 70. https://doi.org/10.30872/jsakti.v2i2.2681

Sabrani, A., Wedashwara W., I. G. W., & Bimantoro, F. (2020). Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia. *Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (JTIKA )*, *2*(1), 89–100. https://doi.org/10.29303/jtika.v2i1.87

Setyawan, C., Benarkah, N., & Prasetyo, V. R. (2021). Automatic Text Summarization Berdasarkan Pendekatan Statistika pada Dokumen Berbahasa Indonesia. *KELUWIH: Jurnal Sains Dan Teknologi*, *2*(1). https://doi.org/10.24123/saintek.v2i1.4045

Susanti, S., Azmi, M., Ali, E., Rahmaddeni, R., & Saputra Wijaya, Y. (2020). Perbandingan Boolean Model Dan Vector Space Model Dalam Pencarian Dokumen Teks. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, *11*(2), 268–277. https://doi.org/10.31849/digitalzone.v11i2.4168