Jovita Nurvania (1301174215)

LAPORAN TUGAS NLP PEKAN 3 – TFIDF & PPMI

I. Deskripsi Masalah

Term Frequency — Inverse Document Frequency atau TF — IDF adalah suatu metode algoritma yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus.

PPMI atau *Positive Pointwise Mutual Information* adalah metode yang menyatakan apakah kedua kata pada sebuah dokumen sering muncul secara bersamaan atau independent. Nilai PPMI bersifat positif beda dengan PMI yang bisa saja negatif.

II. Penyelesaian Masalah

Pada penelitian kali ini, dilakukan pelatihan pada 20 file yang terdiri dari 20 artikel juga, yaitu 10 artikel pertama dengan topik ‘pembelajaran jarak jauh karena covid-19’ dan 10 artikel kedua dengan topik ‘kejahatan media sosial terutama pada whatsapp’. Semua artikel bersumber dari internet, salah satunya adalah website berita seperti kompasiana. Kemudian ke-20 file tersebut digabung menjadi satu dokumen untuk dihitung frekuensinya dan memperlihatkan kata mana yang memiliki frekuensi tertinggi. Disini dilakukan pembatasan yaitu hanya sebanyak 200 kata tertinggi yang ditampilkan.

Sebelum diproses dengan metode lainnya, pada dokumen dilakukan *preprocessing* yaitu dengan mengubah huruf kapitalnya menjadi huruf kecil semua dengan metode *lowercasing*. Setelah itu, penghilangan tanda baca dan spasi yang berlebihan pada dokumen untuk mempermudah proses setelahnya. Setelah it baru dilakukan tokenisasi dan kemudia dihitung frekuensinya,

Frekuensi dari setiap kata ini nantinya akan digunakan di proses manapun baik itu TF-IDF maupun PPMI. Pada TF-IDF, dihitung terlebih dahulu nilai TF dari perkatanya, kemudian baru dihitung nilai IDF nya dari perkata yang ada. Jika masing-masing nilai sudah ada maka dikalikan menjadi nilai TF-IDF. Nilai tersebutlah yang jika ditranspose akan menjadi matriks yang akan digunakan untuk menghitung *cosine similarity* perdokumennya.

Untuk metode PPMI yang pertama dilakukan adalah menghitung matriks *co-occurrence* nya dengan menentukan window, jika yang digunakan 2 window maka akan terbagi menjadi sebelah kiri dan kanan. Setelah dihitung, maka hitung frekuensi bigram dari dokumen tersebut. Dari nilai frekuensi bigram tersebut, dapat dihitung nilai untuk matriks *co-occurrence* nya. Dari nilai yang ada pada matriks tersebut bisa diolah menjadi nilai probabilitas matriks term context dengan membaginya dengan total dari term context. Hasil akhirnya berubah matriks probability dari term dan context.

Selain itu, juga diharuskan menghitung nilai PPMI dari pasangan kata bigram dengan menghitung probabilitas dari masing-masing kata dan juga probabilitas dari bigram tersebut. Setelah itu baru dihitung sesuai dengan rumus. Karena pada PPMI tidak terdapat nilai negative, maka nilai negativenya diubah menjadi 0 dan jika terdapat nilai 0 yang mana tidak bisa dikalikan dengan log2 hasil matriks akan none.

III. Analisis Hasil

Setelah didapatkan masing-masing matriksnya maka dihitung ukuran matriks dari masing-masing metode. Ukuran matriks dari TF-IDF adalah 303x200. Ukuran matriks *co-occurrence term-context* adalah 1582x1582. Ukuran matriks PPMI adalah 1582x1582.

Dari hasil penelitian untuk menjawab persoalan, dapat diketahi bahwa elemen matriks TF-IDF yang bernilai tidak sama dengan 0 adalah 5,17% sedangkan untuk elemen matriks PPMI yang bernilai tidak sama dengan 0 adalah 0,67%. Selanjutnya beralih pada penghitungan nilai *cosine similarity* antar dokumen, untuk dokumen yang bertopik sama yaitu dokumen ke-10 dengan ke-15 dan dokumen ke-200 dengan ke-300 bernilai lebih besar jika dibandingkan dengan nilai *cosine similarity* pada dokumen ke-100 dengan dokumen ke-302 yang hanya bernilai 0 karena terdapat pada topik yang berbeda.

Setelah itu, berdasarkan dari matriks tersebut dapat dihitung *cosine similarity* antar kata. Disini dilakukan pengambilan kata dari kata dengan frekuensi tertinggi, lalu dibandingkan kedua kata pertama dari topik 1 (nomor&whatsapp) kedua dari topik 2 (jarak&jauh) yang ketiga dari campuran topik 1 dan topik 2 (peretasan&siswa). Dari perhitungan nilai *cosine similarity* tersebut, dapat dilihat bahwa perhitungan dengan kata yang berasal dari topik berbeda bernilai 0 karna tidak ada kemiripan sama sekali.

Untuk membandingkannya, dilakukan perhitungan *cosine similarity* berdasarkan metode berbeda yaitu matriks *co-occurrence term-context.* Setelah dibandingkan hal yang sama adalah untuk topik berbeda nilai *cosine* similaritynya tetap lebih kecil dibandingkan dengan topik yang sama. Nilai berbeda terdapat dinilai ketiganya, yaitu dengan menggunakan matriks *co-occurrence term context* nilai menjadi lebih tinggi karena perkalian dan penjumlahan serta pembagiannya yang berbeda, jadi nilainya lebih terisi.

Setelah itu, dilakukan lagi perhitungan nilai *cosine similarity* pada ketiga bigram yang telah digunakan pada nomor sebelumnya tapi dengan berdasarkan matriks PPMI. Dapat dilihat bahwa untuk nilai *cosine similarity* pada topik yang sama untuk topik pertama bernilai 0 sedangkan topik kedua bernilai lebih tinggi dari metode-metode sebelumnya, dan untuk topik berbeda bernilai none karena nilai probabilitasnya 0 dan tidak bisa dihitung dengan log2.

IV. Kesimpulan

Kesimpulan yang dapat diambil dari hasil laporan adalah dalam penilaian probabilitas munculnya atau frekuensi munculnya sebuah pasangan kata lebih baik dihitung berdasarkan matriks *co-occurrence term-context* karna memiliki nilai *cosine similarity* yang lebih besari dibandingkan dengan nilai yang berdasarkan matriks TF-IDF maupun berdasarkan matriks PPMI. Jika berdasarkan matriks *co-occurrence term-context* tidak ada hasil *cosine similarity* yang bernilai 0 atau bahkan tidak ada yang tidak dapat didefinisikan atau none.