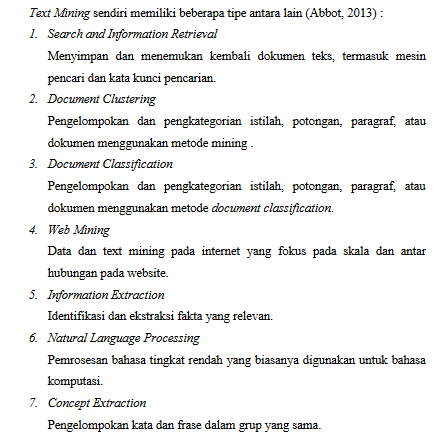
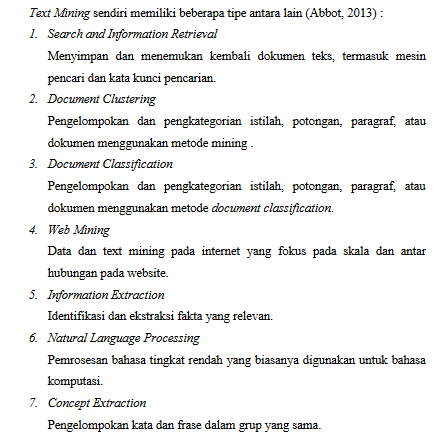
**KUMPULAN JAWABAN**

**Text mining :**

Text mining adalah salah satu bidang khusus dalam data mining yang memiliki definisi menambang data berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Penerapan yang paling umum dilakukan text mining saat ini misalnya penyaringan spam, analisa sentimen, mengukur preferensi pelanggan, meringkas dokumen, pengelompokan topik penelitian, dan banyak lainnya.

~~

Cabang keilmuan yang focus pada pengolahan data teks dikenal dengan nama *text mining*. Salah satu kategori dalam text mining adalah klasifikasi teks (text classification). Klasifikasi teks adalah sebuah proses yang bertujuan untuk menentukan kelas atau kategori dari suatu teks. Teks disini bias berupa frase, kalimat, paragraph, atau bahkan dokumen teks.

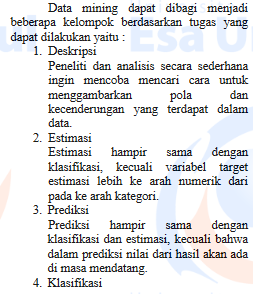
**Jenis2 Text Mining :**

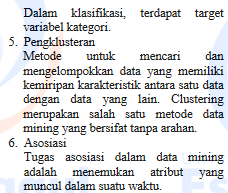
**Data Mining :**

  Data Mining merupakan prinsip dasar dalam mengurutkan data dalam jumlah yang sangat banyak dan mengambil informasi – informasi yang berkaitan dengan apa yang diperlukan seperti apa yang biasa dilakukan oleh seorang analis.

~~~~

Data mining merupakan suatu istilah yang diciptakan untuk menggambarkan proses peralihan melalui database yang besar untuk mencari pola yang menarik dan yang sebelumnya tidak diketahui.

****

****

**Scikit Learn :**

Scikit –Learn adalah modul python yang mengintegrasikan berbagai algoritma pembelajaran mesin state-of-the-art untuk masalah yang diawasi dan tidak diawasi skala menengah. Paket ini berfokus pada membawa pembelajaran mesin ke non-spesialis menggunakan bahasa tingkat tinggi tujuan umum. Penekanan diberikan pada kemudahan penggunaan, kinerja, dokumentasi, dan konsistensi API. Ini memiliki ketergantungan minimal dan didistribusikan dibawah lisensi BSDyang disederhanakan, mendorong penggunaannya baik dalam aturan akademis dan komesial.(Pedregosa Fabiann, 2011). Library dibangun diatas SciPy (Scientific Python) yang harus diinstal sebelum menggunakan scikit –learn.

Tumpukan ini meliputi :

1.Numpy: Paket array n –dimensi dasar 2.Scipy: Pustaka dasar untuk komputasi ilmiah

3.Matplotlib: Komprehensif 2D / 3D 4.Ipyton: Peningkatan konsol interaktif5.Sympy: Matematika

simbolik

6.Pandas: Struktur dan analisis data.

Ekstensi atau modul untuk Scipy care secara konvensional diberinama Scikits. Modul ini menyediakan algoritma pembelajaran dan diberi nama scikit –learn

**TFIDV x *CountVectorizer***

Dalam rangka membangun model vektor, perlu dilakukan proses pembobotan. atau pengklasifikasian vektor yang dimana proses klasifikasi vektor yang akan dilakukan menggunakan klasifikasi CountVectorizer dan TF –IDF Vectorizer.Praproses data ini dilakukan dengan menggunakan aplikasi Jupyter Notebook.

TfidfVectorizer dan CountVectorizer keduanya adalah metode untuk mengubah data teks menjadi vektor karena model hanya dapat memproses data numerik.

Algoritma **CountVectorizer** yang dapat mengubah fitur teks menjadi sebuah representasi vector.

Jelasin cara ngitungnya :

Pembobotan **TF-IDF**

ini digunakan karena efisien, mudah, dan memiliki hasil yang akurat (Robertson, 2004).Term frequency(TF) didefinisikan sebagai jumlah kemunculan suatu kata/istilah dalam suatu dokumen. Misalnya TF pada dokumen pertama untuk kata/istilah “merge” adalah 2, karena kata/istilah tersebut muncul 2 kali dalam dokumen pertama. Pada asumsi pembobotan dibalik TF-IDF, kata-kata dengan nilai TF yang tinggi akan mendapat bobot yang tinggi kecuali jika jumlah dokumen yang mengandung kata tersebut juga tinggi yang disebut inverse document frequencyatau IDF. Misalnya kata “please” memiliki jumlah kemunculan yang tinggi tetapi jumlah dokumen yang mengandung kata “please” juga tinggi, sehingga kata tersebut akan memiliki bobot yang rendah.

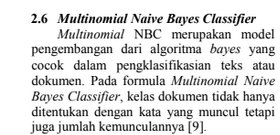
Jelasin cara ngitungnya :

#INTINYA

kalo cv cara ngitungnya menghitung kemunculan tiap katanya ada berapa, kalo tfidf dia memberikan bobot tfidf pada setiap kata

***SVM x Multinominal Naïve Bayes***

Proses klasifikas teks pada umumnya melibatkan algoritma data mining. Dua algoritma data mining yang sering digunakan untuk klasifikasi teks adalah adalah *Naive Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM).



Jelasin cara ngitungnya :

Sedangkan ***Support Vector Machine (SVM)***SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan dengan *linear* maupun *non linear*. *SVM* digunakan untuk mencari *hyperplane* terbaik dengan teknik-teknik klasifikasi lainnya. *SVM* juga dapat mengatasi masalah klasifikasi dan regresi memaksimalkan jarak antar kelas.

Jelasin cara ngitungnya :

**H*yperplane*** adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas. Dalam 2-D fungsi yang digunakan untuk klasifikasi antar kelas disebut sebagai *line whereas*, fungsi yang digunakan untuk klasifikasi antas kelas dalam 3-D disebut *plane similarly*, sedangan fungsi yang digunakan untuk klasifikasi di dalam ruang kelas dimensi yang lebih tinggi di sebut *hyperplane*.

Kesimpulan : SVM memiliki tingkat akurasi yang lebih baik apabila dibandingkan dengan NBC ketika digunakan untuk melakukan klasifikasi pada data teks.