**Laporan Akhir Tugas Besar NLP**

**Anggota:**

1. **Sari Rahmawati 1103130236**
2. **Ade Saepul Mugni 1103130240**
3. **Hafidz Sohihudin Ahmad 1103131309**

**Article Recomendations for News Feed**

**Deskripsi**

Sistem yang dibangun bertujuan untuk mengelompokkan artikel-artikel berita untuk dijadikan rekomendasi bagi pengguna, dilihat dari artikel yang pernah diaksesnya. Sistem dibangun dengan mencoba beberapa metode clusteringuntuk mengelompokkan sekumpulan artikel ke dalam topik yang sesuai dengan artikel-artikel tersebut. Hasil dari pengelompokkan tersebut kemudian dijadikan acuan sebagai prediksi rekomendasi artikel. Diketahui sekumpulan artikel yang sudah dibaca oleh pengguna, sistem merekomendasikan artikel-artikel yang terkait atau yang mungkin pengguna tertarik untuk membacanya. Pembuatan sistem mengacu pada tugas akhir [1].

**Dataset**

Data yang digunakan untuk membangun sistem ini diperoleh dari dataset BBC yang dipublikasikan secara online oleh UC Dublin Machine Learning Group [2]. Data ini terdiri dari 2225 dokumen yang dikelompokkan menjadi 5 topik dari tahun 2004-2005. Kelima topik tersebut yaitu bisnis, entertainment, politik, olahraga, dan teknologi. Format dokumen dataset ini adalah .txt yang berisi tulisan artikel sesuai dengan topiknya.

**Tahapan Pengerjaan**

1. Preprocessing

Bertujuan untuk merepresentasikan data artikel yang semula berbentuk tulisan menjadi bentuk representasi angka yang akan digunakan sebagai fitur pada artikel tersebut. Di bawah ini beberapa tahapan preprocessing yang dilakukan.

a) Tokenization dan cleaning

Pemisahan kata-kata dari setiap teks, serta penghapusan kata-kata yang tidak perlu. Menggunakan perintah split() dan strip() pada python.

b) POS tagging dan Penghapusan stopword

Menandai setiap kata dengan tag yang bersesuaian. Menggunakan NLTK.pos\_tag.

c) Stemming

Mencari kata yang merupakan kata dasar (contoh: dimakan, memakan -> makan). Menggunakan NLTK Porter Stemmer.

d) Konversi ke Vektor

Membuat matriks kemunculan fitur di setiap data/teks “[example, feature] = frequency”. Menggunakan sklearn vectorizer.

e) Low Frequency Removal

Kata-kata yang kemunculannya kurang dari tiga kali dihapus.

f) Term Frequency and Inverse Document Frequency (TF-IDF)

Mencari kepentingan suatu kata berdasarkan frekuensi kemunculan kata tersebut.

2. Clustering

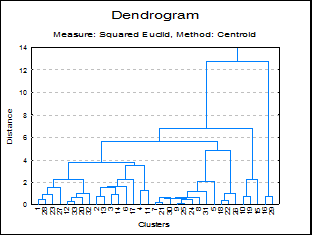
Dilakukan untuk mengelompokkan artikel-artikel ke dalam suatu kluster tertentu. Metode clustering yang digunakan adalah sebagai berikut.

a) K-Means

K means merupakan algoritma clustering dimana pada algoritma ini memisahkan data berdasarkan jumlah k cluster. K mean berusaha untuk membentuk kelompok berdasarkan karakteristik yang sama antara satu sama lain. K means mencoba mencari samples data yang memiliki karakteristik yang sama dengan centroidnya (centroid merupakan nilai tengah pada masing masing cluster).

b) Hierarchical clustering

Merupakan algoritma clustering dimana pembentukan cluster dengan melakukan penggabungan atau pemisahan secara berurutan. Bentuk dari cluster ini berupa pohon atau dendogram dimana akar dari pohon ini merupakan cluster yang unik yang menggabungkan sample-sample dan daun yang merupakan cluster yang hanya berisi satu sample. Penggabungan yang dilakukan pada tugas besar ini menggunakan kriteria Ward yang meminimalkan varian dari cluster yang digabungkan..



c) Affinity Propagation

Berbeda dengan metode clustering yang menginisialisasi centroid sejumlah kluster secara acak, affinity secara simultan mempertimbangkan semua data point memiliki potensi sebagai sebuah kluster atau eksemplar. Metode ini secara rekursif mengirimkan pesan di antara data point hingga eksemplar yang mengindikasikan klusternya terbentuk dengan baik atau fungsi objektiknya terpenuhi. Pesan yang dikirm antar data point terdiri dari availability dan responsibility. Sedangkan fungsi objektif dari metode ini adalah net similarity di suatu kluster.

3. Evaluasi clustering

Evaluasi ini bertujuan untuk memilih metode clustering yang cocok diterapkan. Di bawah ini adalah beberapa pengukuran yang digunakan untuk mengevaluasi metode clustering.

a) Homogeneity

Homegeneity adalah kesamaan antar anggota pada cluster. Hasil yang baik pada homogenty adalah jika setiap label pada suatu cluster memiliki kelas yang sama.

b) Completeness

Completeness adalah ketika semua sample dengan kelas yang sama terkonsentrasi pada satu cluster.

c) V-Measure

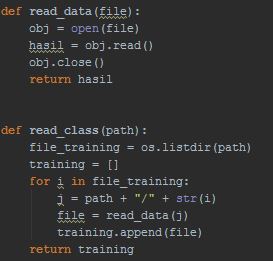
V Measure menyatakan nilai harmonik antara homogenity dan completeness pada clustering. Ve measure bisa didapatkan dengan rumus dibawah ini:

**v = 2 \* (homogeneity \* completeness) / (homogeneity + completeness)**

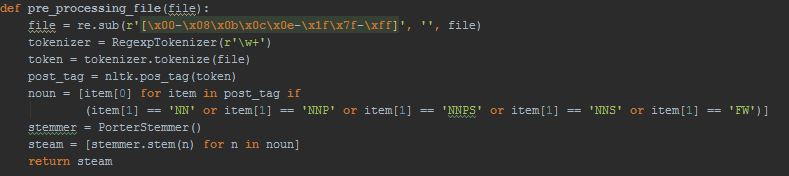
4. Testing

**Langkah-Langkah Program**

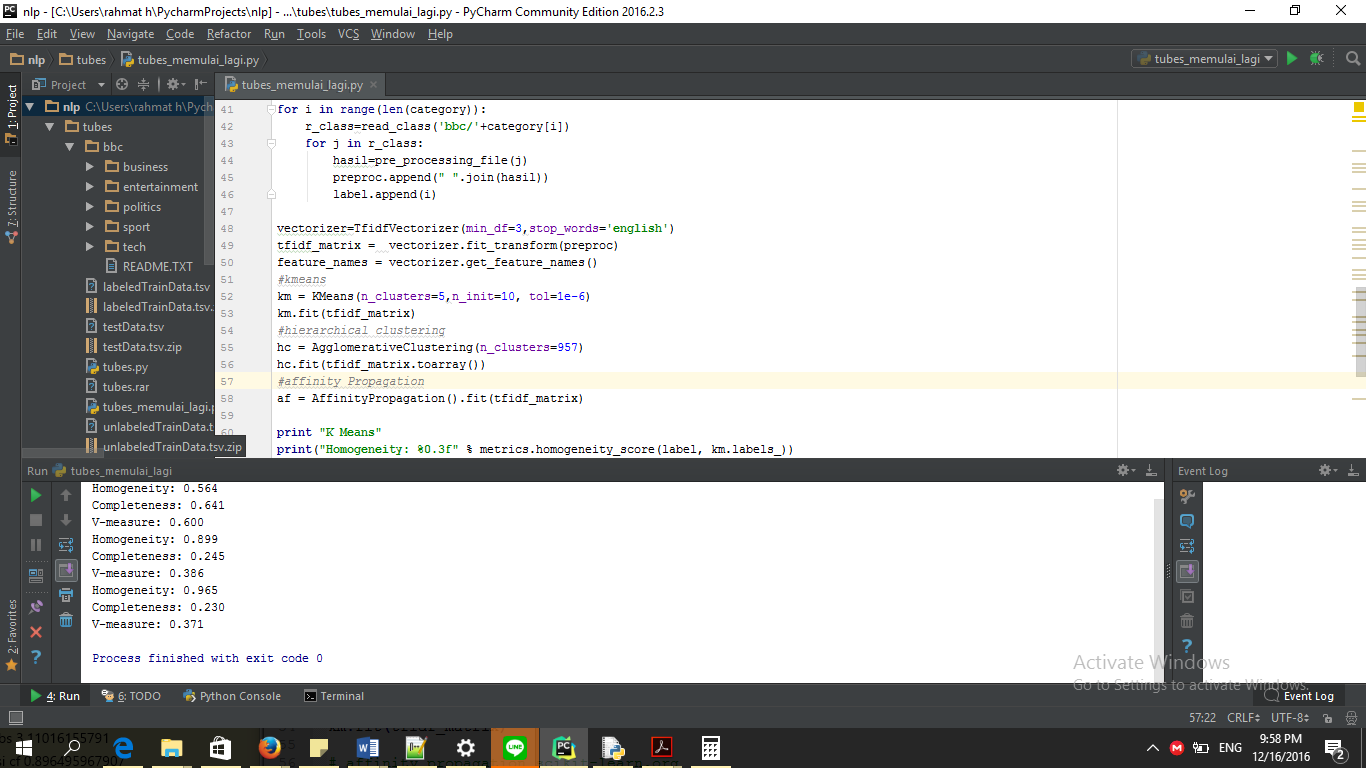
· Load data



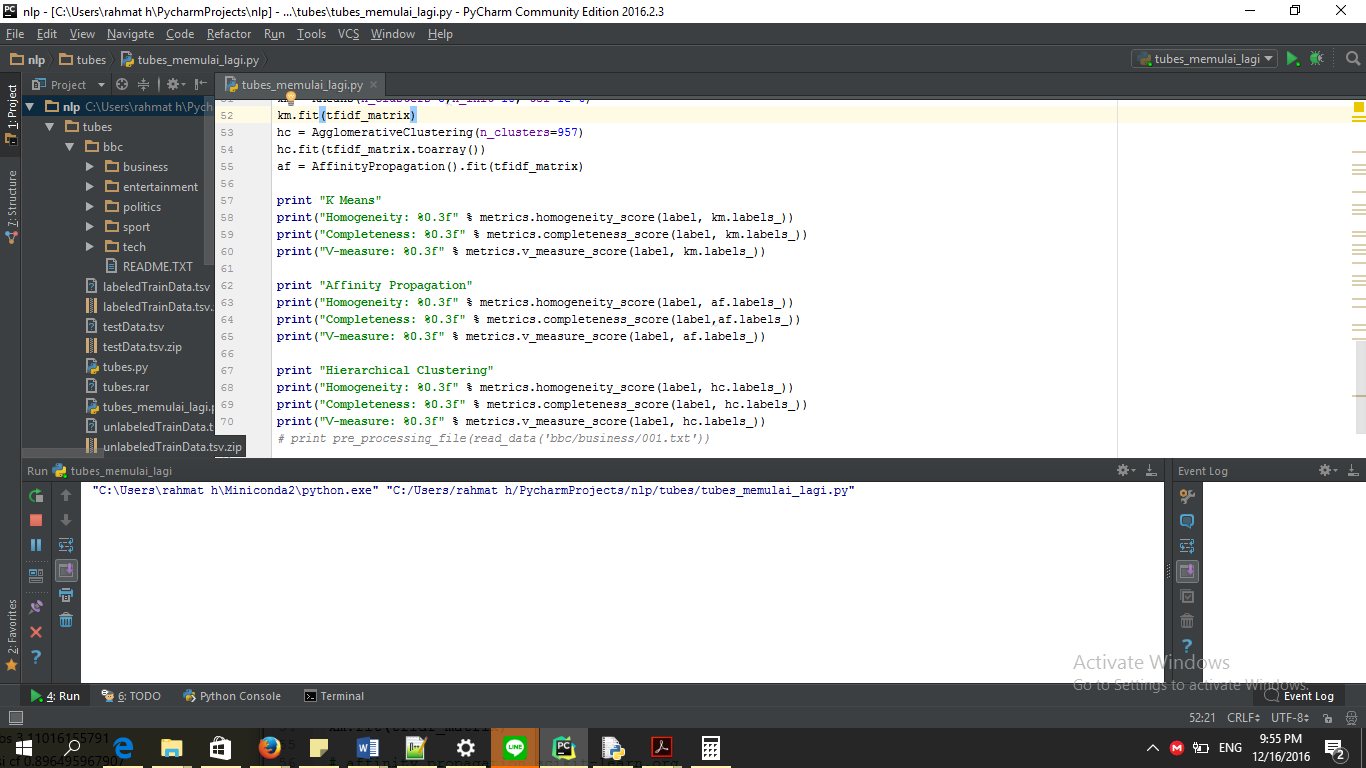
· Preprocessing



· Clustering



· Testing



**Hasil**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Homogeneity** | **Completeness** | **V-Measure** | **Clusters** |
| **K-Means** | 0.607 | 0.664 | 0.635 | 5 |
| **Hierarchical Clustering** | 0.965 | 0.230 | 0.371 | 957 |
| **Affinity Propagation** | 0.899 | 0.245 | 0.386 | 413 |

**Analisis dan Kesimpulan (?)**

Dilihat dr v measure diperoleh hasil tertinggi menggunakan kmeans dengan nilai 0.635.Sementara apabila dilihat dr masing-masing yaitu homogenity dan completeness, pada homogenity yang paling baik adalah menggunakan hierarchical clustering. Meskipun kmeans memiliki nilai v- measure yang lebih baik dibanding 2 metope lain, K mean memiliki kelemahan yaitu jumlah cluster harus ditentukan sebelum melakukan prediksi, karena itu kita tidak bisa tau berapa jumlah cluster yang cocok untuk data tersebut.

**Referensi**

[1] Shen, Minghan. “Article Recomendations for News Feed”. CS224N Final Project.

[2] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.