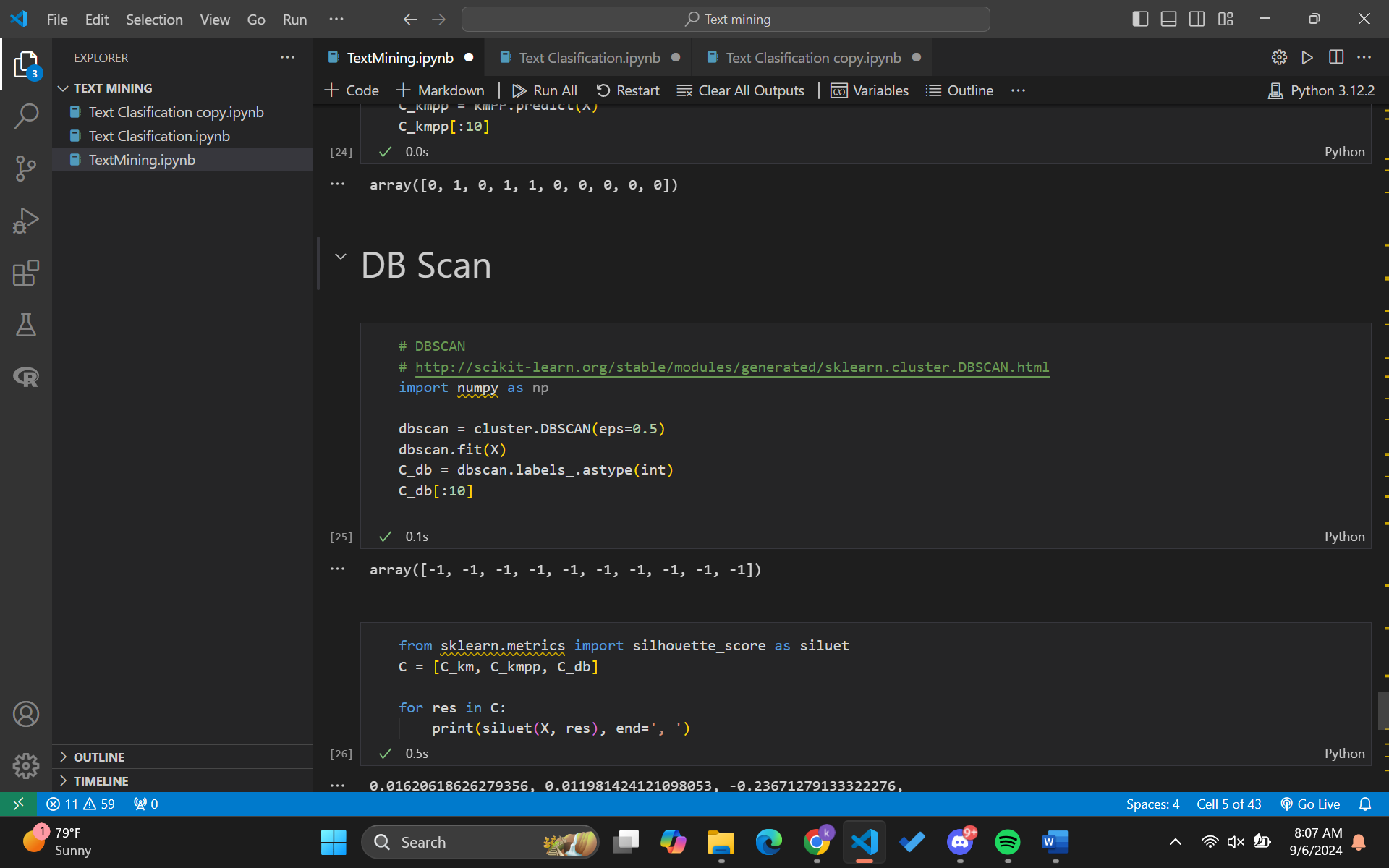
1. Lakukan seluruh percobaan pada modul ini dan berikan analisis yang kalian temukan



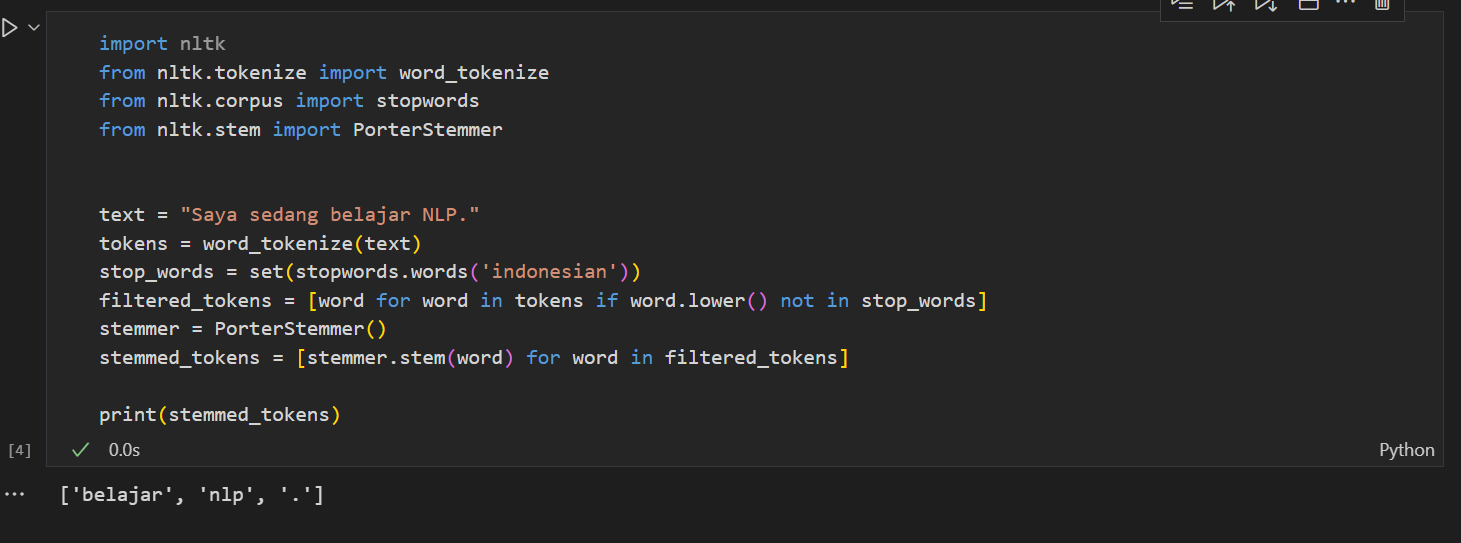
Percobaan telah saya lakukan pada kelas teori sebelumnya.

1. Jelaskan perbedaan hasil dari Preprocessing menggunakan NLTK, TextBlob dan Sastrawi dan berikan contohnya.
   1. NLTK (Natural Language Toolkit):

Fitur:

* Tokenisasi: Memecah teks menjadi kata atau kalimat.
* Stemming: Mengubah kata ke bentuk dasarnya (stem).
* Stopword Removal: Menghapus kata-kata umum yang tidak memberikan makna.

Contoh :

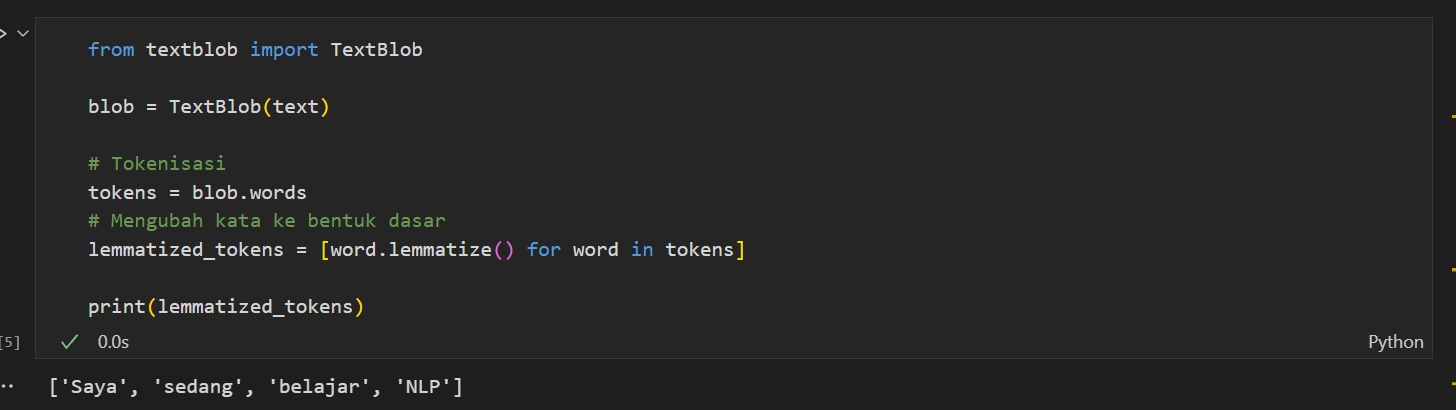


1. TextBlob

Fitur:

* Tokenisasi: Memudahkan pemisahan teks.
* Stemming dan Lemmatization: Mengubah kata ke bentuk dasarnya.

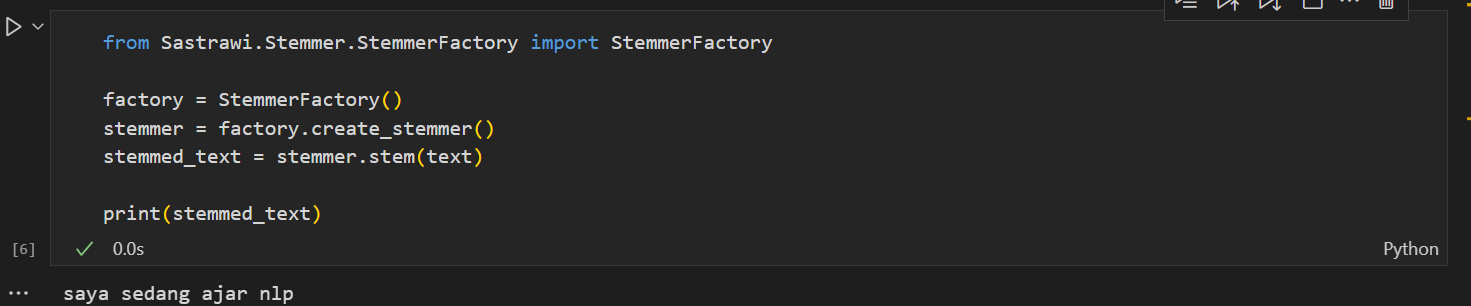
Contoh :



1. Sastrawi  
   Fitur:

* Tokenisasi: Memecah teks menjadi kata.
* Stemming: Khusus untuk bahasa Indonesia.

Contoh :

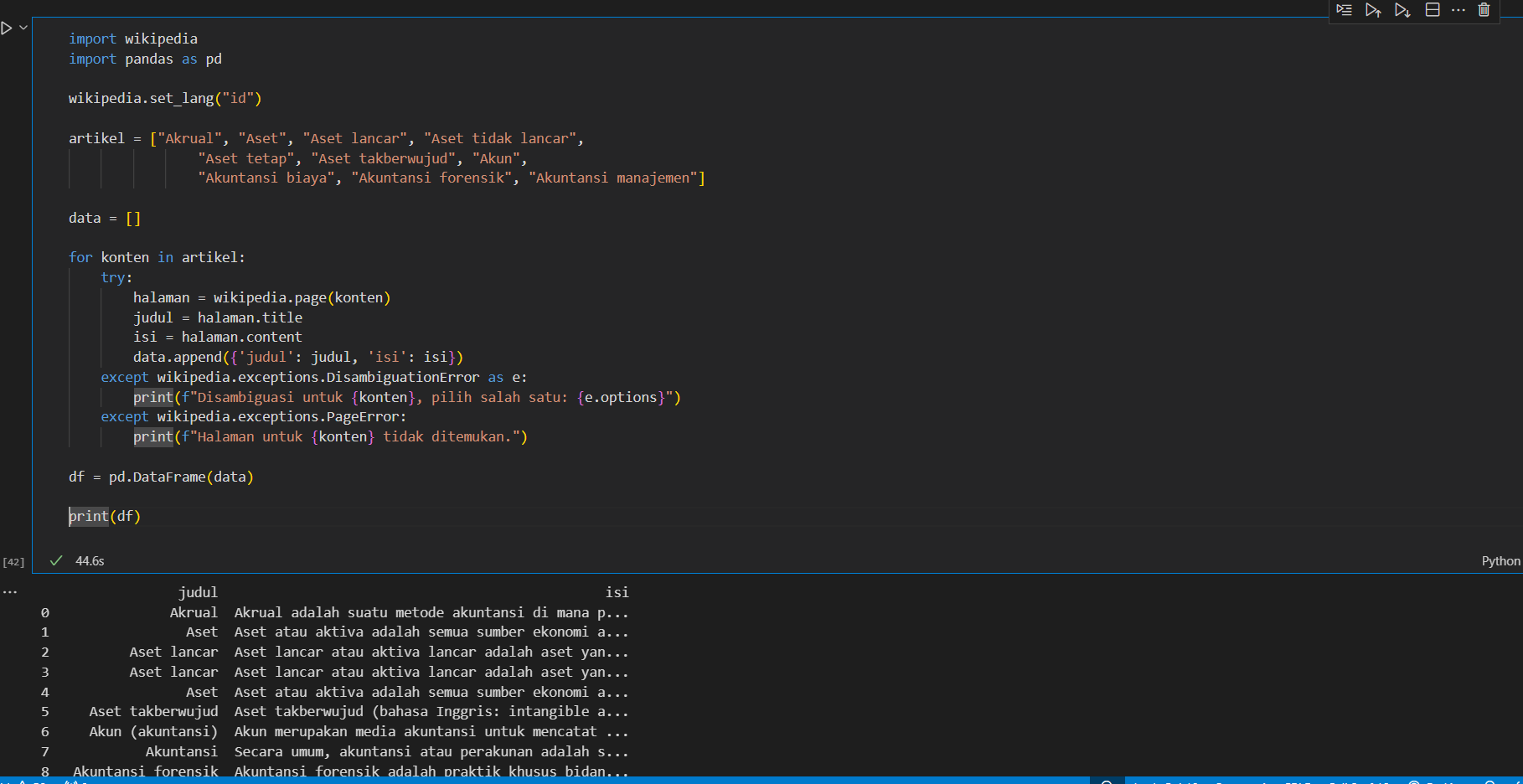


**Kesimpulan**

* **NLTK**: Memfokuskan pada penghapusan stopwords dan stemming, sehingga menghasilkan token yang lebih ringkas.
* **TextBlob**: Tidak menghapus stopwords secara otomatis, tetapi lebih memudahan dalam analisis sentimen dan lemmatization.
* **Sastrawi**: Khusus untuk bahasa Indonesia, memberikan hasil stemming yang relevan dengan konteks bahasa indonewsia.

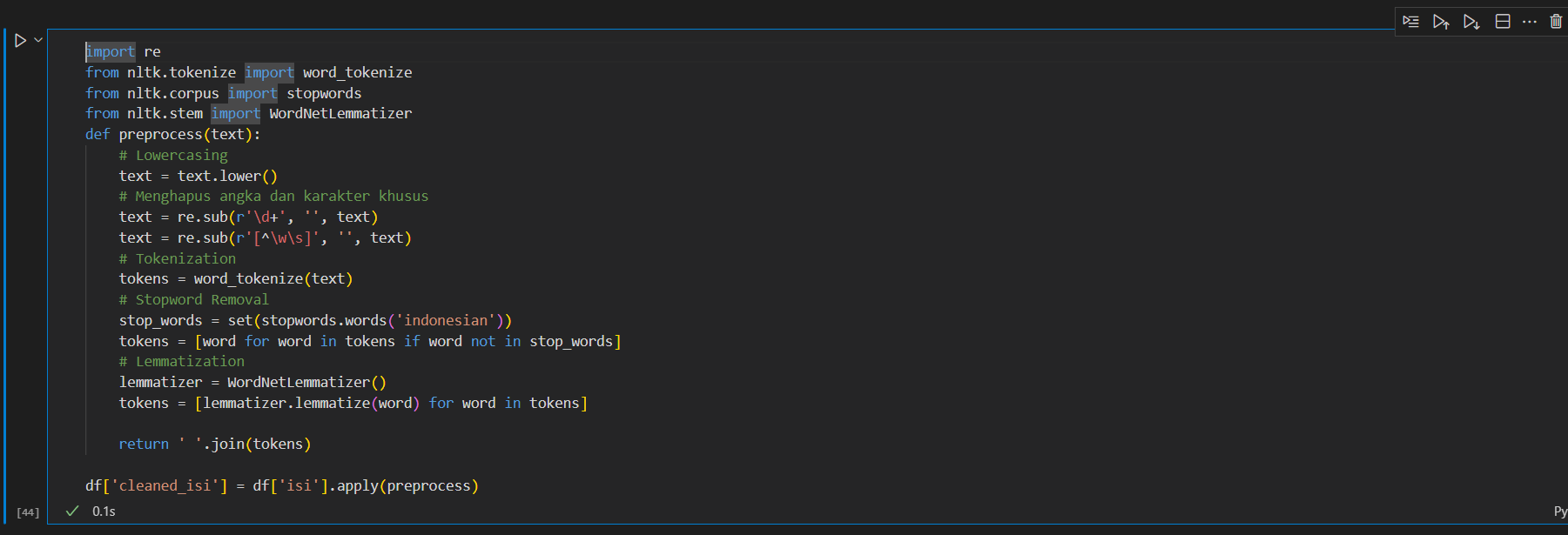
Jadi pemilihan library yang tepat, tergantung pada kebutuhan spesifik suatu data, seperti bahasa teks dan kompleksitas preprocessing yang diinginkan.

1. Crawling dataset dengan total 10 pada berbagai judul artikel Wikipedia berdasarkan daftar topik sesuai dengan akhiran nim dan wajib dalam topik yang sama. Akutansi: https://id.wikipedia.org/wiki/Daftar\_istilah\_akuntansi



Crawling data dilakukan melalui Wikipedia, untuk melakukan crawling data wikipedia kita menggunakan library Wikipedia. Sesuai soal saya mengdapatkan artikel mengenai akuntansi, kemudian siapkan 10 judul yang akan di ambil datanya. Data kemudiann diambil dan data disimpan dalam data frame, berisi judul dan isi dari artikel yang ada pada wikipedia.

1. Lakukan preprocessing yang sudah diajarkan pada modul ini (menggunakan salah satu library saja).



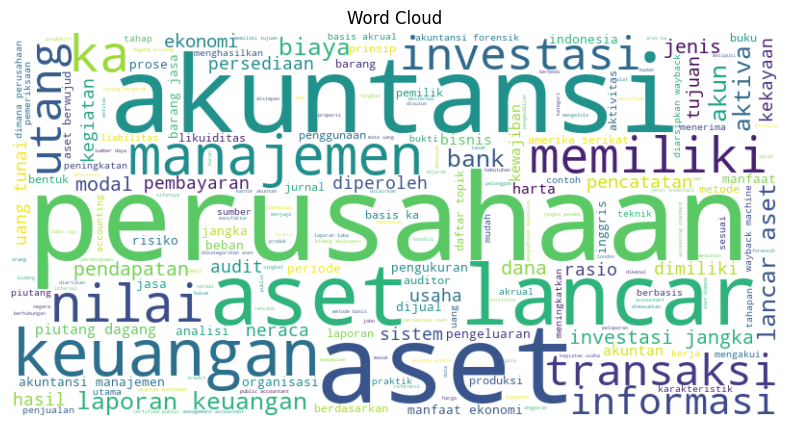
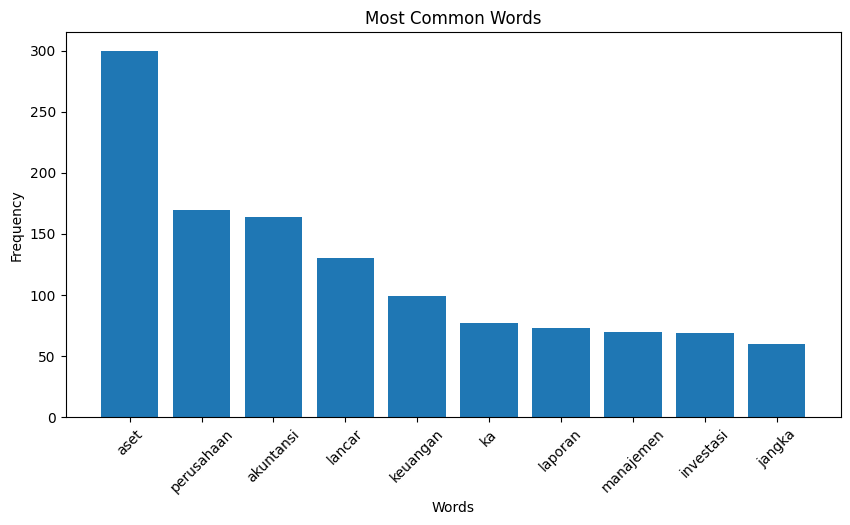
Kemudian saya membuat fungsi preprocess yang menggunakan library NLTK, fungsi berisikan beberapa langkah penting dalam membersihkan dan mengelolah teks untuk analisis *Text Mining* lebih lanjut. Berikut beberapa proses yang dilakukan dalam fungsi preprocess :

* Mengubah teks menjadi huruf kecil dengan text.lower(), hal ini dilakukan untuk memastikan konsistensi dan menghindari perbedaan yang disebabkan oleh kapitalisasi.
* Menghapus angka dan karakter khusus menggunakan RegeEx dengan re.sub(), yang membantu membersihkan teks dari elemen yang tidak relevan untuk analisis kata.
* Setelah pembersihan awal, teks di-tokenisasi menggunakan word\_tokenize() dari NLTK, yang membagi teks menjadi kata-kata individual atau token.
* Melakukan stopwords, yaitu menghapus kata-kata umum yang sering kali tidak memberikan informasi berarti (seperti "dan" atau "di"), dengan menggunakan daftar stopwords bahasa Indonesia yang disediakan oleh NLTK.
* Lemmatization, kata-kata yang telah dibersihkan kemudian diubah ke bentuk dasarnya menggunakan WordNetLemmatizer(). Hal ini membantu mengurangi variasi kata yang berbeda namun memiliki makna yang sama menjadi satu bentuk standar.

Sehingga hasil dari peprocessing menggunakan fungsi ini adalah string teks yang telah siap digunakan untuk analisis, yang disimpan dalam kolom baru bernama ‘cleaned\_isi’ pada DataFrame.

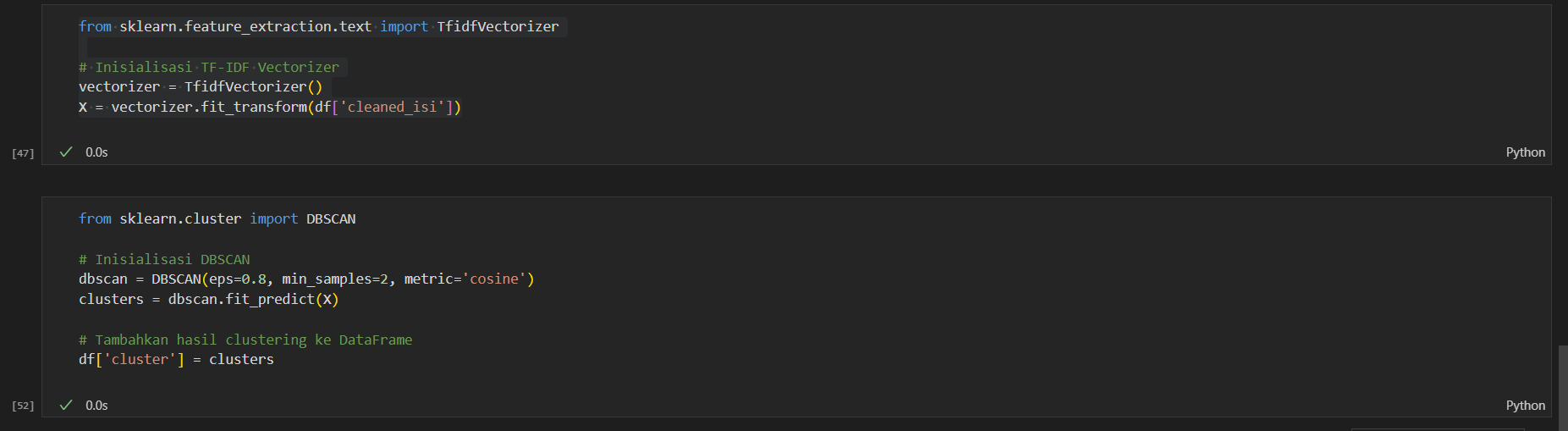
1. Buatlah wordcloud dan most common word barplot, interpretasikan hasilnya



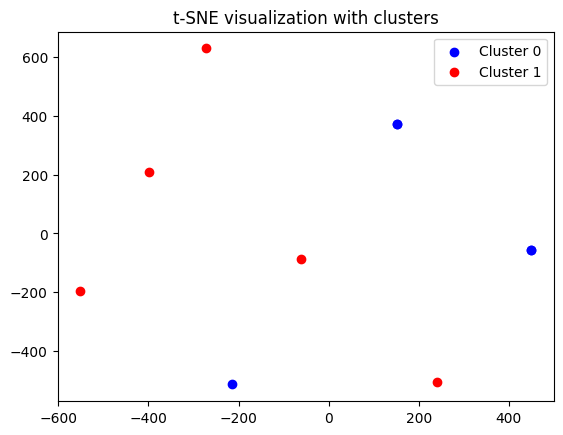
Kedua visualisasi ini untuk memperlihatkan kata yang paling sering muncul dalam artikel-artikel yang telah dikumpulkan, kata pertama yang paling banyak muncul yaitu ‘aset’, hal ini dapat terjadi dikarenakan beberapa artikel yang digunakan memang membahas mengenai asset.

1. Lakukan clustering dengan menggunakan fitur TF-IDF

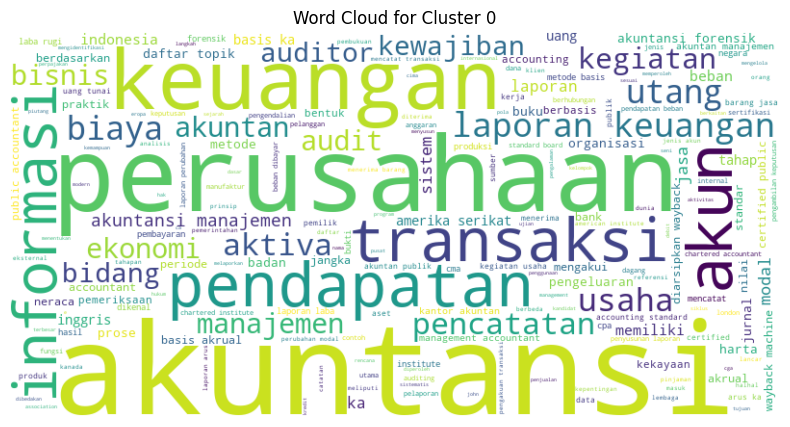


Setelah dilakukan preprocessing, kemudian dilakukan konversi teks ke dalam bentuk matriks menggunakan fitur TF-IDF, dan melanjutkan proses clustering menggunakan metode DBSCAN. DBSCAN diinisialisasi dengan parameter eps (epsilon) untuk jarak maksimum antar titik dalam satu cluster dan min\_samples untuk jumlah minimum titik yang diperlukan untuk membentuk cluster, serta menggunakan metrik cosine untuk mengukur kesamaan antar dokumen. Hasil clustering, yang memberikan label cluster untuk setiap dokumen yang kemudian ditambahkane ke dalam DataFrame pada kolom baru ‘cluster’, memungkinkan analisis lebih lanjut berdasarkan kelompok data yang serupa.

1. Buat visualisasi clusternya dan lakukan interpretasi terhadap hasil tersebut



Berdasarkan visualisasi diatas dapat dilihat bahwa cluster terbagi menjadi 2 cluster yaitu cluster 0 dan cluster 1.



Ini merupakan hasil worldcloud untuk cluster 0, diaman kata-kata yang banyak digunakan diantara lain seperti ‘perusahaan’, ‘akuntansi’, ‘keuangan’, dan istilah akuntansi lainnya.



Sedangkan ini, merupakan word cloud dari cluster 1. Dapat dilihat bahwapada cluster ini kata yang banyak mucul berisi mengenai asset, Dimana memang beberapa artikel yang saya ambil mmebahas mengenai macam-macam dari asset.

1. Gunakan validasi menggunakan salah satu Davies-Bouldin index atau Silhouette score

* validasi saya lakukan menggunakan Silhouette untuk mengukur seberapa baik data dalam sebuah cluster dibandingkan dengan cluster lain. Skor yang saya dapat adalah 0.323 menunjukkan bahwa cluster yang terbentuk suda cukup baik, tapi masih ada beberapa tumpang tindih atau masih tidak terlalu terpisah.

