Word Clustering with Virtual Threads

This Java application uses the K-Means technique to cluster word embeddings. It makes use of virtual threads for effective parallel processing and makes clustering parameter setting simple. The objective is to classify related words according to their embeddings, which is helpful for applications involving natural language processing (NLP), such as recommendation systems and semantic analysis.

Main Features

- Support for Virtual Threads: Makes use of Java's virtual threads (Java 17+) to manage memory effectively and analyse large datasets more quickly.
- Using the Euclidean distance to allocate words to the closest centroids, K-Means Clustering groups words with similar semantic meanings based on their vector representations.
- Configurable Parameters:
  - Threads: To improve resource usage, adjust the number of virtual threads.
  - Clusters: Specify how many clusters will be used to group words
- Word Embedding Loading: Provides versatility by loading word embeddings from a file in a variety of formats.
- Output Results: Following the clustering process, an output file containing the words in each cluster and the centroids is generated.

Configuration:

The following options can be configured by the application:

- **Threads**: Number of virtual threads for processing.
- **Clusters**: Number of clusters to create during the K-Means clustering.

Clustering Process

- The K-Means algorithm groups words by their embeddings.
- The Euclidean distance between word vectors assigns them to the nearest centroid.
- After clustering, the results are written to an output file, including the words in each cluster and the centroids.

Key Components

- **Runner**: Main class for user input and application flow.
- **ApplicationState**: Stores the app's state, including word embeddings and configuration.
- **KMeansClustering**: Performs the clustering algorithm.
- **WordEmbeddingLoader**: Loads word embeddings from files.
- **SearchResult**: Stores the result of a word search.

Dependencies

- **Java 17+**: Required to support virtual threads and other features.