Design algorithm

1. Delete special characters

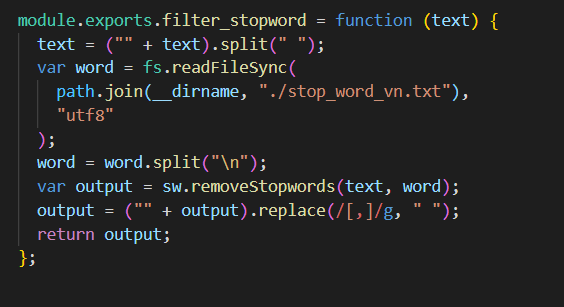


//Hình: Function clear\_special\_chars

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * Text: String | * Text (has filtered out special characters): String | * Used to delete the special characters of an passed string |

// Bảng: Description function clear\_special\_chars

1. Filter stopword

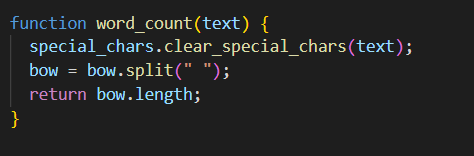


// Hình: Function filter\_stopword

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * Text: String | * output: String | * Used to filter words is stopword |

//Bảng: Description function filter\_stopword

1. TF-IDF algorithm

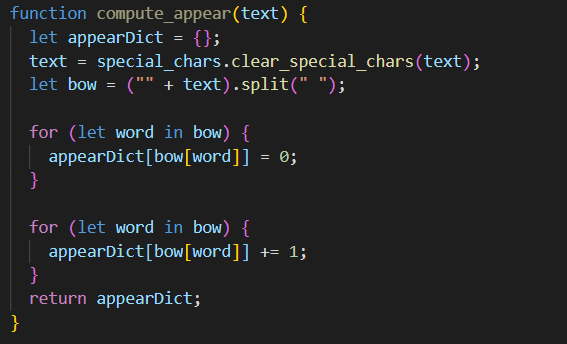
Count the Number of words in a document

// Hình: Function word\_count

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * Text: String | * Bow.length: Number | * Used to calculate the Number of words in a paragraph |

// Bảng: Description function word\_count

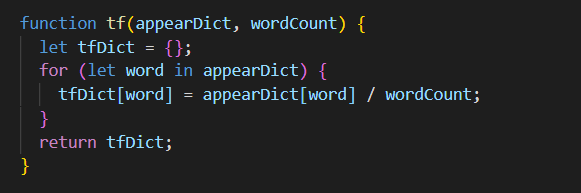
* Counts the Number of occurrences of words in a document



// Hình: Function compute\_appear

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * Text: String | * appearDict: List | * Used to calculate the Number of occurrences of words |

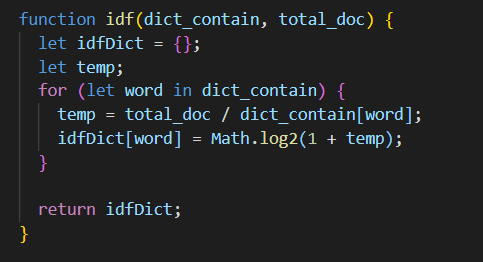
* Compute term frequency (TF)



// Hình: Function tf

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * appearDict: List * wordCount: Number | * tfDict: List | * Used to calculate term frequency |

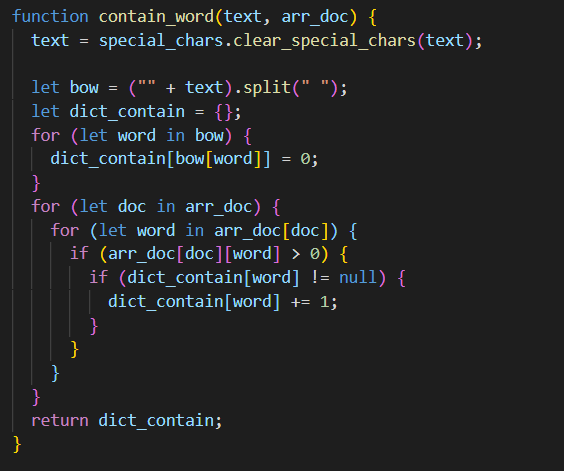
* Compute inverse document frequency (IDF)



//HÌnh: Function idf

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * Dict\_contain: List * Total\_doc: Number | * idfDict: List | * Used to calculate inverse document frequency |

* Calculates the Number of documents containing words

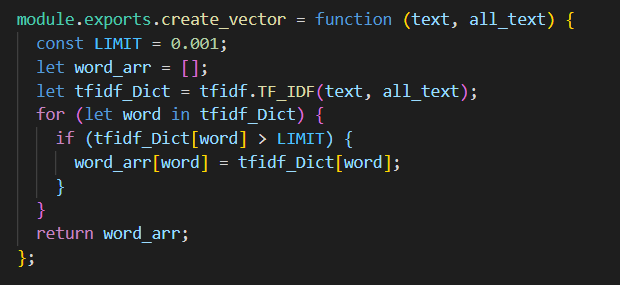


// Hình: Function contain\_word

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * text: String * arr\_doc: Array | * dict\_contain: List | * Used to calculate the Number of documents containing words |

//Bảng: Description function contain\_word

1. Create vector

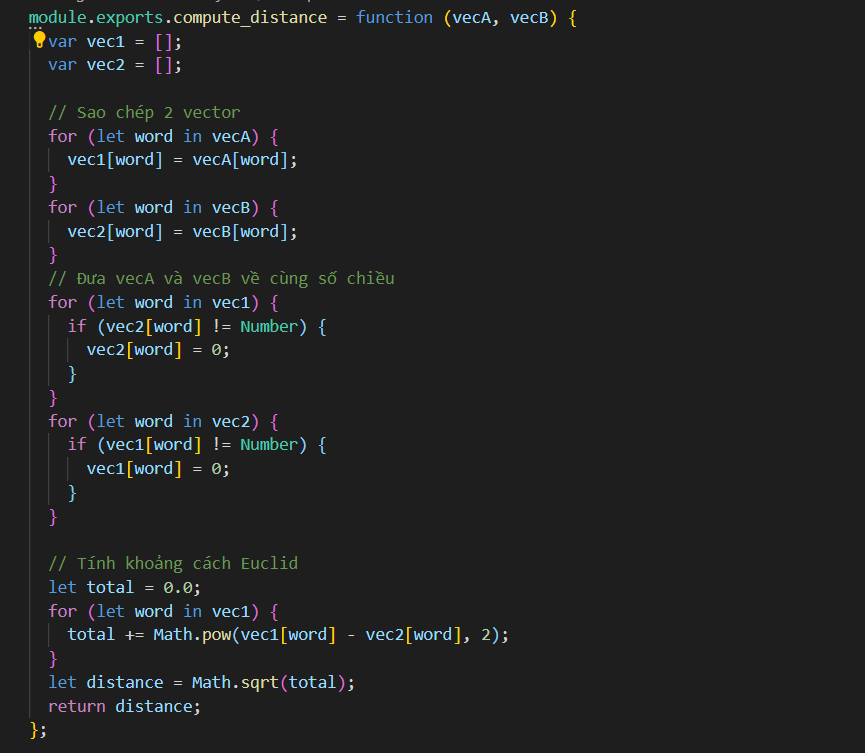


// Hình: Function create\_vector

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * text: String * arr\_text: Array | * word\_arr: Array | * Used to create a vector of a document |

// Bảng: Description function create\_vector

1. Calculate the distance between two vectors using euclidean algorithm



// Hình: Function compute\_distance

|  |  |  |
| --- | --- | --- |
| Input | Output | Usage |
| * vecA: Array * VecB: Array | * distance: Number | * Used to calculate the distance between 2 vectors |