

Tutorial 5 – TFIDF

Consider the 5 text files provided to you. Each text file has text about a particular subject, as follows:

1.txt is about weather

2.txt is about animals

3.txt is about earthquakes

4.txt is about music

5.txt is about sports

Our aim is to create a Python script that reads the contents of these text files, and then uses a TFIDF Vectorizer to find out the most important token of each file. The following steps will help you:

1. Read the contents of the text files into an array of string, where each textfile will occupy one location of the array

2. Create a TfidfVectorizer, removing the stop words at the same time, as follows:

```
vectorizer = TfidfVectorizer(stop_words="english")
```

3. Generate the TFIDF matrix using the array of strings and vectoriser

4. For each textfile, print the most important token in that file, i.e. the token with the highest TFIDF value