Discovering Similar Text Files

Given a collection of text files find the “N” most similar text files based on below mentioned similarity measure.

- Find top 200 most frequent words (normalized) in each text file. Convert all alphabetic characters to uppercase.
- Normalize the word count in each file by dividing by the total number of words.
- Exclude the common stop words, using nltk package (from nltk.corpus import stopwords), when counting frequent words, total word count, and calculating normalized frequency.
- Calculate the similarity index between two files as the sum of normalized values for all common frequent words.
- Report the N most similar pairs of textbooks.

*N is the user-provided input.