ЛАБОРАТОРНАЯ РАБОТА.

Работа с ключевыми словами.

Для нахождения ключевых слов выполните следующие действия

1) **Tokenize (получите токены документа):**

```
tokenizer = nltk.data.load('tokenizers/punkt/english.pickle')
tokens = tokenizer.tokenize(documents[curr_doc_index])
```

Lemmatize (учитывает токены в различных вариантах записи):

```
lemmatizer = WordNetLemmatizer()
tokens = [lemmatizer.lemmatize(token) for token in tokens]
```

3) **Удаление спецсимволов:**

```
stopwords = stopwords.words('english')
tokens = [token for token in tokens if token not in stopwords]
```

4) Подсчет частоты токена в тексте (ниже tf[i] представляет i-ый документ в массиве:

```
tf[curr_doc_index] = Counter(tokens)
```

5) Для каждого токена подсчитывается инверсная частота:

```
idf[t] = math.log(len(documents) / len([doc_index in range(len(documents))
if tf[doc_index][t] > 0]))
```

6) Для каждого токена и документа подсчитывается коэффициент **tf-idf**:

```
tfidf[t] = tf[curr_doc_index][t] * idf[t]
```

7) Выбираем **k слов с максимальным tf-idf** :

```
terms_sorted_tfidf_desc = sorted(tfidf.items(), key=lambda x: -x[1])
terms, scores = zip(*terms_sorted_tfidf_desc)
keywords = terms[:k]
```

Теперь попробуйте все это реализовать самостоятельно, предварительно сформировав массив путей к текстовым документам.