Поиск плагиата без организации постоянной коллекции документов.

По заданному документу найти набор документов, которые являются источниками частей текста исходного документа.

Общую актуальную картину
http://ltrc.iiit.ac.in/icon_archives/ICON2010/11Dec2010/Tutorial3-First-document.pdf
http://www.ijart.org/2011/ijart013.pdf
http://www.uni-weimar.de/medien/webis/publications/papers/stein_2012t.pdf

Этапы решения задачи.

1. Парсинг документа, удаление стоп-слов.
2. Выделение набора ключевых слов, характеризующих документ (http://www.jucs.org/jucs_13_10/machine_learning_based_keywords)
3. Генерация запроса поисковой системе по набору ключевых слов
4. Выгрузка документов релевантных запросу
5. Вычисление сходства на основе string subsequence kernel сходства документов с исходным (http://jmlr.csail.mit.edu/papers/volume2/lodhi02a/lodhi02a.pdf)
6. Выделение заимствованных частей текста. Объединение результатов сравнения наиболее релевантных документов.

Дополнительные задачи:

1. Построение аннотации документа, выполнение запросов к поисковой системе по предложениям из аннотации
2. Удаление из списка релевантных документов документы того же автора
3. Добавление в набор анализируемых документов текстов, соответствующих спискам из библиографии

Рекомендуемая книга по python:
Марк Саммерфилд. Программирование на Python 3.

Используемые средства:
NLTK (nltk.org), PDFMinerp (www.unixuser.org/~euske/python/pdfminer/index.html)