Академическая тошнота текста

Академическая тошнота текста - это насыщенность текста ключевыми словами. По тошноте текста можно судить о его натуральности и оптимизации под поисковые запросы. Академическая тошнота измеряется в процентах и вычисляется как отношение числа повторов пяти самых частых слов к общему числу слов в тексте. Важно, что при расчёте (в т.ч. и общего количества слов) не учитываются “стоп-слова”: предлоги, союзы, местоимения и т.д., а все остальные учитываются независимо от формы употребления.

Для того, чтобы обмануть простые системы анализа текстов, прибегают к замене русских букв на аналогичные по написанию английские (и наоборот). Использование данного подхода будем называть **мошенничеством**. Для расчёта корректного значения академической тошноты необходимо привести слова к нормальному варианту написания.

Задача: реализовать программу на языке Python, которая для заданных файлов рассчитывает показатель академической тошноты содержащихся в них текстов
Входные данные: набор файлов с текстами
Выходные данные: таблица в базе данных, содержащая информацию о каждом проверенном файле: показатель академической тошноты и флаг наличия мошенничества в тексте

Желательно, чтобы предложенная реализация:

- содержала тесты
- использовала параллелизм

Рекомендуется использовать:

- модуль nltk для стемминга и списка стоп-слов
- sqlite3 в качестве СУБД