Рокфеллер БД

1. Первично было введено 31 ссылка.
2. Рассмотрено 3073 страницы с данными. Пропущены ссылки для документов, книг и архивов. Но ссылки на них учтены.
3. Сохранено 1000 страниц с исходными данными. Страницы, которые открылись в браузере.
4. Получено 2064270 сырых слов. Слова все начинаются с большой буквы. Указана их связь в предложении. Имеется огромное число повторов.
5. После очистки от простейшего мусор осталось 2033314 слов. (нужно совершенствовать словарь ненужных слов).
6. После очистки от предложени, которые содержали 1 одинарное слово осталось 1786212 строк с разным количеством слов.
7. Удалили те строки где содержались одиночные слова.
8. Слова которые были двойными и тройными переписали в отдельную таблицу.
9. Таких слов оказалось 135367.
10. Что уже можно было анализировать.
11. После выделения 2-х и 3-х словных конструкций, было проведено разделение словосочетаний с Рокфеллером и без таких значений.
12. Соответственно были получены таблицы с количеством записей 17584 для Рокфеллер и 117783 для не Рокфеллеров.
13. И эти таблицы требовали дальнейшего совершенствования.
14. Можно выделить ключевые слова, например «Bank» и по ним выделить организации.
15. Убрать повторы из этих таблиц.
16. Необходимы таблицы связи партнеров с партнерами и других людей с другими людьми.
17. Выделить признаки в таблицах, по которым задействовать их при поиске совпадений входящих слов (организация поиска по всей базе данных).