ДЗ по дисциплине "Информационный поиск и анализ данных"  
  
Дз состоит из 3-ех частей (выполняется на Java):  
  
1 Часть. Написать робот для сбора публикаций c (crawler) сайта:  
- сайт выбираете произвольно, можно RSS ленты;  
- должны скачиваться html/xml страницы и обрабатываться коды возрата сервера;  
- со скаченных страниц должны парсится следующие данные:  
— текст публикации (новости)  
— заголовок  
— временем публикации  
— автор (если есть)  
— ссылка на страницу с публикацией  
  
2 Часть. Сохранение данных в базу:  
- использовать базу Elastic search  
- создать индекс для хранения документов с соответствующими полями (заголовок, время публикации, текст, ссылка, автор)  
- сохранять документы с идентификаторами, вычисленными как хэш от текста документа  
- Дополнительно:  
— выполнять поиск по ключевым словам и другим атрибутам документа - хор.  
— выполнять агрегации (например по авторам) - отл.  
  
3 Часть. Анализ текстов  
Любой из следующих алгоритмов:  
- поиск неполных дублей  
— TF,TF\*IDF, Long Sent, Heavy Sent – удовл.  
— TF\*RIDF, Lex Rand, Megashingles, – хор.  
— Shingles+MinHash, Word2Vec+MinHash и др. на базе MinHash – отл.  
- кластеризация – отл;  
- классификация – отл;  
- выявление сущностей – отл.