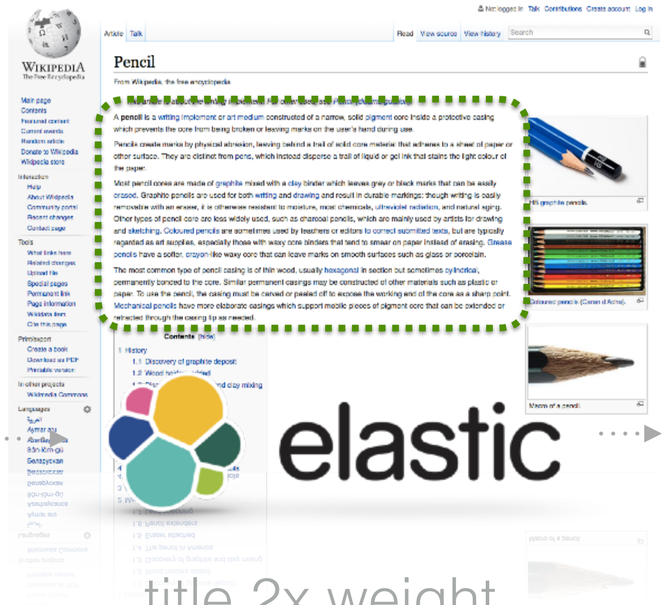
```
/* lazily load dump into a dataframe*/
val df = sqlContext.read.format("com.databricks.spark.xml").option("rowTag", “page").
   load("s3a://wiki-xml-dump/enwiki-20160407.xml")



/* extract introductory paragraphs + title */
val index_rdd = df.select("title", "revision").map(article =>
  Map("title" -> article(0), "body_text" -> article(1).toString.slice(0, 2000)))



/* write to elastic search */
index_rdd.saveToEs("wiki_index/article")
```

wood penicl



title 2x weight

+

Levenshtein distance

~~wood~~ Pencil

neo4j

```
/* lazily load dump into a dataframe*/
val df = sqlContext.read.format("com.databricks.spark.xml").option("rowTag", “page").
   load("s3a://wiki-xml-dump/enwiki-20160407.xml")



/* extract introductory paragraphs + title */
val index_rdd = df.select("title", "revision").map(article =>
  Map("title" -> article(0), "body_text" -> article(1).toString.slice(0, 2000)))



/* write to elastic search */
index_rdd.saveToEs("wiki_index/article")
```