概述

在Dblp官网上找到，dblp.xml包括以下几种类型：

- article – An article from a journal or magazine.//杂志期刊上的文章
- inproceedings – A paper in a conference or workshop proceedings.//会议论文
- proceedings – The proceedings volume of a conference or workshop.//会议记录合集
- book – An authored monograph or an edited collection of articles.
- incollection – A part or chapter in a monograph.//部分专著
- phdthesis – A PhD thesis.//博士论文
- mastersthesis – A Master's thesis. There are only very few Master's theses in dblp.//硕士论文
- www – A web page. There are only very few web pages in dblp.

对于hw3，只考虑article 和inproceedings，其中article包含journals/pvldb，inproceedings包含conf/sigmod和conf/icde.

对于hw4，考虑article、inproceedings、phdthesis、mastersthesis。

预处理

为了避免多次加载dblp.xml，对文本进行预处理，select出需要的字段和记录，保存为列式文件。

```
        val sourceFile = base + "dblp.xml"
        val processedFile = base + "dblp1"
        val parquetFile = base + "par"

        var t = sc.textFile(sourceFile)
            .map(_.replaceAll("</inproceedings>|</mastersthesis>|</phdthesis>", "</article>").replaceAll("<inproceedings|<mastersthesis|<phdthesis", "<article"))
        t.saveAsTextFile(processedFile)

        val articleDf = sqlContext.read
            .format("com.databricks.spark.xml")
            .option("rowTag", "article")
            .option("excludeAttribute", true)
            .option("treatEmptyValuesAsNulls", true)
            .load(processedFile)

        val articlesDf = articleDf.select("title", "author", "year").filter(row => !row.isNullAt(0) && !row.isNullAt(1) && !row.isNullAt(2))

        articlesDf.write.parquet(parquetFile)
        print("test2")
```