CountNaiveDistinctJournals

```
/**
 * Naive count of distinct journals in a set of years
 * (distinction made on title-ISSN, title without affiliation (aff))
 * - total number of journals in one year
 * - number of distinct journals
 * - journals with no ISSN
 * - number of distinct journals with ISSN
 */
```

```
Uri: 2009
Records in 2009: 35840
Total number of journal records: 28377 (79.18%)
Number of ~distinct journals: 1118
Journals with no ISSN (distinct title): 215 (13 recoverable by title)
Number of distinct journals with ISSN: 916
```

ExtractDistinctJournalsPerYear

```
/**
 * Create a set of distinct journals for one year, distinction on the title.
 * If two journals have the same title, they must have also the same ISSN
 * There is no cleaning of ISSN and wrong titles.
 * Extraction of country from title.
 * File: distinctJournalsYEAR.xml
 * <journal><title></title><issn></issn><country></country></journal>
 */
```

```
Uri: 2009
Records in 2009: 35840
Number of journal records: 28377
Number of ~distinct journals: 1118
```

ExtractDistinctIssnPerYear

```
/**
 * Given a set of distinct journals for one year (journals names and ISSNs),
 * creates a map of "ISSN->journals with this ISSN".
 * Generation of distinctISSN2009.xml and cleanDistinctISSN2009.xml.
 * ISSN are both cleaned and not cleaned (there is a naive cleaning method).
 * Also NULL for journals without ISSN
 * <issn><v></v><journal></journal></issn>
 */
```

```
Uri: 2009
Records in 2009: 35840
Number of journals: 28377
Number of distinct ISSN: 856 (262 merged: same ISSN, different title)
Number of journals without ISSN: 215
Number of distinct cleaned ISSN: 845 (11 merged)
Number of journals without cleaned ISSN: 215
```

Wrong Format ISSN:

The right format is YYYY-YYYZ, where Y is a number and Z a number or the character X. Java pattern is:
[0-9][0-9][0-9][0-9][/-][0-9][0-9][0-9][0-9[X]]

1.	ISSN into brackets:	(0048-6019)
2.	Followed by *(Print)* or *(online)*:	0014-2336 (Print)
3.	Ending with space or dot	
4.	With middle underscore:	0019_638X
5.	With spaces:	0048 - 3754
6.	With two dashes, but 8 numbers:	0115-8857-7
7.	Without dash:	03653439
8.	Two ISSNs:	1583-2023; 1842-8991
9.	XXXX-XXX	
10.	XXXX	
11.	XXXX (online)	
12.	X	
13.	X-XXXXX-XXX-X	
14.	XX-XXX-XXXX-X	

Naïve Cleaning: covers point 1, 2, 3, 4, 5, and 6 of wrong formats

```
Uri: 2009
Records in 2009: 35840
~distinct ISSN: 856
--number wrong ISSN: 40
~distinct cleaned ISSN: 845
--number wrong cleaned ISSN: 13
```

Problems:

- Some distinct journals haven't ISSN: **215 journals**
- Some journals have different titles but the same ISSN: they are the same journal
- Some journals have different titles because come characters are Latin instead of Cyrillic
- Some journals have same titles, but not all have ISSN: they are the same journal
- 6 records have ISSN (3 distinct) but empty journal name

```
2073-4948      1028-0308   0871-0287
```

Possible strategy:

- Step 1: find all distinct journals with ISSN (916) and without (215, 13 recoverable)
- Step 2: put on top of the list only distinct journals with a correct ISSN. Here two journals can have the same ISSN and different title: build a chain. Build a list of journals with bad format ISSN
- Step 3: clean each bad formed ISSN, trying to add to the top of the list or at the end of the chain. Build a list of journals with no cleanable ISSN
- Step 4: for each journal with no cleanable ISSN, try to find the title in the top of the list or in the chains
- Step 5: for each journal without ISSN, try to find the title in the top of the list or in the chains
- *For remaining ISSN with a bad format, introduce probability on string matching, repeating step 4 (TODO)*
- *Match ISSN and journal titles with a govern list (TODO)*

```
//////////////////////////////////////////////////////
// STEP1: ALL DISTINCT JOURNALS, WITH SAME TITLE/ISSN
//////////////////////////////////////////////////////
~Distinct Journals with no cleaned ISSN: 916
~Distinct Journals without ISSN: 215

//////////////////////////////////////////////////////
// STEP 2: build structure
//////////////////////////////////////////////////////
~Distinct Top ISSN: 816
Related ISSN: 59
Bad format ISSN: 41

//////////////////////////////////////////////////////
// STEP 3: naive cleaning of malformed ISSN
//////////////////////////////////////////////////////
~ Distinct Top ISSN: 833 (+17)
Related ISSN: 70 (+11)
Wrong ISSN: 13 (28 corrected)

//////////////////////////////////////////////////////
// STEP 4: check journals names of wrong ISSN
//////////////////////////////////////////////////////
~Distinct Top ISSN: 833
Related ISSN: 73 (+3)
~Distinct Journals without ISSN: 215
Wrong ISSN: 10 (3 corrected)

//////////////////////////////////////////////////////
// STEP 5: check journals names of journals without ISSN
//////////////////////////////////////////////////////
~Distinct Top ISSN: 833
Related ISSN: 82 (+9)
~Distinct Journals without ISSN: 203 (12 corrected, 1 in wrong)
Wrong ISSN: 10
```