Number of Numbers document:

Before processing digits, we put content of txt become a list of words.

1.

use class StringUtils methods getDigits to get numbers from each word. This method wooks in this way:

An empty ("") String will be returned if no digits found in str.

StringUtils.getDigits(null) = null

StringUtils.getDigits("") = ""

StringUtils.getDigits("abc") = ""

StringUtils.getDigits("1000$") = "1000"

StringUtils.getDigits("1123~45") = "12345"

StringUtils.getDigits("(541) 754-3010") = "5417543010"

StringUtils.getDigits("१२३") = "१२३"

2.

If the value of getDigits(word) is not equal to “” or null, then we extract the word, to be precise, we extract this word, and words which before and after this words.

3.

There are conditions that not belongs to numbers, and it should be delete from the total numbers.

1). Telephone((541) 754-3010);

page ( 20-f, (20) );

Solution: words which contain “-“ or “(“ “)” with distance smaller than 5 are deleted.

2). Numbers which is year, 2016 should be delete.

Solution: Check if the number is year(return true), delete this number.

String text\_Number = text.replaceAll("[^0-9]", "#");

String[] arr = text\_Number.split("#");

for (String s : arr) {

if (s.matches("^[0-9]{4}$")) {

hasYear= true;

}

}

3). Numbers form is MM/DD/YYYY: June,01,2016

Solution: Check if the first word belongs to from “january” to “december”, or from “jan”to “dec”, and YYYY is belongs to year, then delete all of those.

4). Numbers of item and section, ex: item 15, section 9

Solution: Check if words before number is “item” or “section”, then delete those numbers.

4.

Count numbers and write to excel.

WritableSheet sheet = workbook.createSheet("First Sheet", 0);

sheet.addCell(new Label(0, fileNum, f.getName()));

sheet.addCell(new Label(1, fileNum, String.valueOf(count\_Num)));

Of course, the exception has many other cases, and we will still work on improve algorithm to make sure the the numbers counted are useful.