

HW3 Little Research Report

[repository \(https://gitlab.com/krikun/ir-gg\)](https://gitlab.com/krikun/ir-gg)

0 Structure

```
project/          # little research project
project/docs/     # test document set
project/src/main/scala  # research scala src

project/tasks/    # this reports from .md
```

build with sbt

```
# git clone ssh or https ..
cd project
sbt
...
> run
```

1 Doc set

placed in project/docs/ directory

docId	terms	tokens
docs/arrays	502	1948
docs/mutable-and-immutable-collections	362	1048
docs/collection-sets	453	1854
docs/collections-trait-traversable	434	1633
docs/collections-trait-iterable	278	713
docs/maps	447	1810
docs/scala-for-java-programmers	962	4983
docs/classes	227	430
docs/sets	453	1854

1.1 Total tokens

```
tokens : 16273
terms  : 1709
```

tokens split by Punct and normalized to terms just by lowercase

```
val raw = io.Source.fromFile(file).mkString
// tokenization
val tokens = raw.split("[\\p{Punct}\\s]+")
val words = tokens.map(word => word.toLowerCase())
val roughly = words.toSet
```

1.2 Stop words

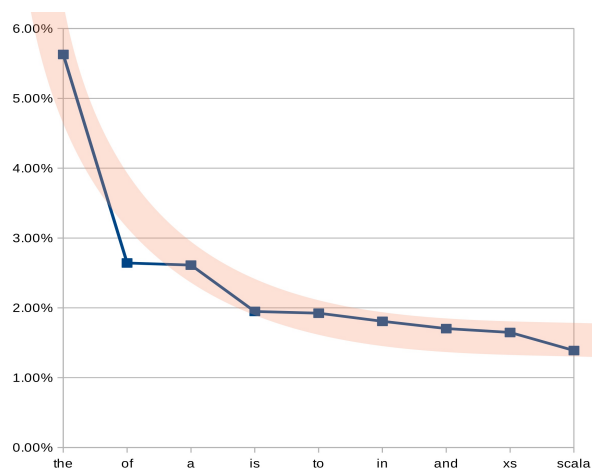
First 9 by occurrence:

the	of	a	is	to	in	and	xs	scala
916	430	425	317	313	294	277	268	226

Total: 3466 out of 16273 tokens

*As we can see, this is logarithmic distribution.
xs and **scala** appears because of test doc set:
[src \(http://docs.scala-lang.org/tutorials/\)](http://docs.scala-lang.org/tutorials/)*

or, if you wanna, by percentage to tokens:



2 Boolean search

Query : **define**

*200th most frequent
occurrence : 15
5 of 9 documents*

Result docId:

```
docs/arrays
docs/mutable-and-immutable-collections
docs/collections-trait-traversable
docs/maps
docs/scala-for-java-programmers
```

2.1 Edit distance: less or equal 1

Query : **define**

Matched words:
define, defined, defines

*Recall increase,
7 of 9 documents*

```
docs/scala-for-java-programmers
docs/arrays
docs/classes
docs/maps
docs/collections-trait-iterable
docs/collections-trait-traversable
docs/mutable-and-immutable-collections
```