HW3 Little Research Report

repository (https://gitlab.com/krikun/ir-qq)

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	roughly	tokens
docs/arrays	502	558	1948
docs/mutable-and-immutable-collections	362	396	1048
docs/collection-sets	453	506	1854
docs/collections-trait-traversable	434	487	1633
docs/collections-trait-iterable	278	307	713
docs/maps	447	494	1810
docs/scala-for-java-programmers	962	1055	4983
docs/classes	227	240	430
docs/sets	453	506	1854

2.1 Total tokens

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

tokens split by Punct and normalized to terms just by lowercase

```
val tokens = value.split("[\\p{Punct}\\s]+")
val words = tokens.map(word => word.toLowerCase())
val roughly = tokens.toSet
```

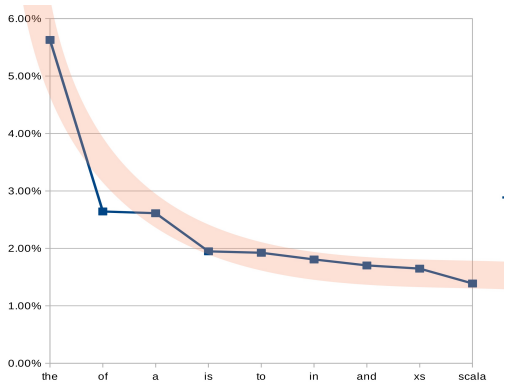
2.2 Stop words

First 7 by occurrence:

the	**of**	**a**	**is**	**to**	**in**	**and**	**scala**	***total***
916	430	425	317	313	294	277	226	*2972*

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

or, if you wanna, by percentage:



2.3