INFS7410 Assignment 1

Roy Portas - 43560846

Code Description

Function Name	Description
`App.main`	The main entrypoint of the program, instantiates the document loader and the tokenizer
`CranfieldDocument.parseFile`	Parses the cranfield document and stores it into the CranfieldDocument instance
`CranfieldDocument.getSectionText`	Gets a section of the document
`CranfieldDocument.getSectionNames`	Gets the names of all the sections in a cranfield document
`Engine.addCranfieldDocument`	Adds a cranfield document to the engine
`Engine.tokenize`	Tokens the corpus of a cranfield document
`Engine.addWord`	Adds a word into the engine corpus
`Engine.removeStopwords`	Removes a list of stopwords from the document
`Engine.printFrequency`	Prints the frequency of every word in order
`Engine.printCorpus`	Prints the summary of the words in the document
`FileLoader.loadStopWords`	Loads the stopwords file into the program
`FileLoad.loadCranfieldDocuments`	Loads a directory of cranfield documents into the program

Rules for Tokenizer and Stopword Removal

The following steps were applied to the tokenizer:

- Remove all SGML (XML) tokens from the document
- Iterate through each character in the text, storing each alphanumeric character into the string and when a non-alphanumeric character is found add the string into the word list and clear the string.
- All words are added into a hashmap with the keys being the word and the occurances being the values

The following steps apply to the stopword removal

- Each word in the stopword list is iterated through
- That word is removed from the hashmap

Statistics 1 Answers

Total number of words: 241166
Total number of unique words: 10392

the: 19455
of: 12717
and: 6678
a: 6246
in: 4645
to: 4563
is: 4114
for: 3493
are: 2429
with: 2265
on: 1944
flow: 1849
at: 1835
by: 1755
that: 1570
an: 1389
be: 1271
pressure: 1207
boundary: 1156
from: 1116
as: 1113
this: 1081
layer: 1002
which: 975
number: 973
j: 892
results: 885
it: 856

mach: 824
theory: 789
shock: 712
was: 698
method: 687
heat: 629
two: 618
been: 590
surface: 586
were: 583
wing: 550
body: 545
obtained: 539
1: 525
r: 520
given: 520
temperature: 518
effects: 511
velocity: 501
these: 500
solution: 498
or: 486

Statistics 2 Answers

Total number of words: 130642
Total number of unique words: 9984

flow: 1849
pressure: 1207
boundary: 1156
layer: 1002
number: 973
results: 885
mach: 824
theory: 789
shock: 712
method: 687
heat: 629
surface: 586
wing: 550
body: 545

obtained: 539
1: 525
temperature: 518
effects: 511
velocity: 501
solution: 498
equations: 477
transfer: 476
supersonic: 467
ratio: 465
0: 450
made: 449
laminar: 428
presented: 425
jet: 425
experimental: 423
found: 422
effect: 411
conditions: 411
plate: 393
analysis: 386
distribution: 385
reynolds: 384
range: 381
numbers: 373
case: 365
free: 364
problem: 364
stream: 357
hypersonic: 355
solutions: 351
lift: 348
shown: 347
ae: 344
air: 343
scs: 340