University of CINCINNATI

BIG DATA ANALYTICS

CS 7070

ASSIGNMENT -2

Submitted on:

03/04/2019

Submitted by

Siva Sai Krishna Paladugu

Problem Statement:

1. Design and execute a MapReduce program to computer the TFIDF for terms in set of given documents and submit the following items:

a) TFIDF for all terms in each document, sorted alphabetically by words' letters, and formatted for easy readability.

Code:

- Used same code provided in the zip and ran all phases.
- Used below standalone code to calculate TFIDF based on output files of above step.

```
import java.io.BufferedReader;
import java.io.FileReader;
import java.io.IOException;
import java.util.Comparator;
import java.util.HashMap;
import java.util.Map;
import java.util.StringTokenizer;
import java.util.TreeMap;

public class CalTdIdf {
        //Method to sort Map by VALUE
        public static <K, V extends Comparable<V>> Map<K, V>
sortByValues(final Map<K, V> map) {
        Comparator<K> valueComparator = new Comparator<K>() {
                public int compare(K k1, K k2) {
                        int compare =map.get(k2).compareTo(map.get(k1));
                                if (compare == 0)
                                        return 1;
                                else
                                        return compare;
                }
        };

        Map<K, V> sortedByValues = new TreeMap<K, V>(valueComparator);
        sortedByValues.putAll(map);
        return sortedByValues;
    }
        //Method to sort Map by KEY
        public static void sortbykey(Map<String, TreeMap<String, Double>> map)
         {
                TreeMap<String, TreeMap<String, Double>> sorted = new
                                                                 TreeMap<>();
                sorted.putAll(map);
                for (Map.Entry<String, TreeMap<String, Double>> entry:
                                                           sorted.entrySet())
                        for (Map.Entry<String, Double> child:
                                                 entry.getValue().entrySet())
                                System.out.println(entry.getKey() + "  " +
                                     child.getKey() + "  " + child.getValue());
         }
        //Method to print top 15 words with highest tfidf in each doc
        public static void getTopFifteenSorted(Map<String, Map<String,
Double>> map) {
                TreeMap<String, Map<String, Double>> sorted = new TreeMap<>();
                sorted.putAll(map);
                for (Map.Entry<String, Map<String, Double>> entry:
                                                          sorted.entrySet()) {
                        int counter = 0;
                        System.out.println("The top Fifteen words with Highest
TF*IDF value in document: " + entry.getKey());
                   for (Map.Entry<String, Double> child:
                                                entry.getValue().entrySet()) {
                                counter = counter + 1;
                                if (counter <= 15) {
                                        System.out.println(child.getKey() + "
" + child.getValue());
                                }
                        }
                }
        }
//Method to change "term : (docName, TFIDF)" format to docName : (term,TFIDF)
        public static void interChangeDocNameTerm(Map<String, TreeMap<String,
Double>> map) {
                Map<String, TreeMap<String, Double>> fb = new HashMap<>();
                for (Map.Entry<String, TreeMap<String, Double>> entry:
                                                             map.entrySet()) {
                        for (Map.Entry<String, Double> child:
entry.getValue().entrySet()) {

                                TreeMap<String, Double> temp;
                                if (fb.containsKey(child.getKey())) {
                                        temp = fb.get(child.getKey());
                                        temp.put(entry.getKey(),
child.getValue());
                                } else {
                                        temp = new TreeMap<>();
                                        temp.put(entry.getKey(),
child.getValue());

                                        fb.put(child.getKey(), temp);
                                }
                        }
                }

                Map<String, Map<String, Double>> fbb = new HashMap<>();
                for (Map.Entry<String, TreeMap<String, Double>> entry :
fb.entrySet()) {
                        Map<String, Double> child =
sortByValues(entry.getValue());
                        fbb.put(entry.getKey(), child);
                }
                getTopFifteenSorted(fbb);
        }
        public static void main(String[] args) {
                BufferedReader br1 = null;
                BufferedReader br2 = null;
                BufferedReader br3 = null;

                Map<String, TreeMap<String, Double>> termtfIdf = new
HashMap<>();
                Map<String, Integer> docNameTotalTerms = new HashMap<>();
                Map<String, Double> termIdf = new HashMap<>();
                try {
                //Change path to output files of 3 phases for ques 1&2
                        br1 = new BufferedReader(new
FileReader("./Phase1OP/part-r-00000"));
                        br2 = new BufferedReader(new
FileReader("./Phase2OP/part-r-00000"));
                        br3 = new BufferedReader(new
FileReader("./Phase3OP/part-r-00000"));
                        String line1;
                        String line2;
                        String line3;

                        StringTokenizer word_list;
                        int totalDocs = 0;
                        //Calculating total documents
                        while ((line2 = br2.readLine()) != null) {
                                word_list = new StringTokenizer(line2);
                                String docName = word_list.nextToken();
                                int totalTermInDoc =
Integer.parseInt(word_list.nextToken());
                                docNameTotalTerms.put(docName, totalTermInDoc);
                                totalDocs += 1;
                        }
                        //Calculating IDF for each word
                        while ((line3 = br3.readLine()) != null) {
                                word_list = new StringTokenizer(line3);
                                String term = word_list.nextToken();
                                int docWithTerm =
Integer.parseInt(word_list.nextToken());
                                double tm = (double) totalDocs / docWithTerm;
                                double idf = Math.log(tm);
                                termIdf.put(term, idf);
                        }
//Calculating TF for each word in each doc and then TF*IDF
                        while ((line1 = br1.readLine()) != null) {
                                word_list = new StringTokenizer(line1);
                                String keyPhase1 = word_list.nextToken();
                                String[] termDocName = keyPhase1.split(",");
                                String term = termDocName[0];
                                String docName = termDocName[1];
                                int tFrequency =
Integer.parseInt(word_list.nextToken());

                                TreeMap<String, Double> temp;
                                if (termtfIdf.containsKey(term)) {
                                        temp = termtfIdf.get(term);
                                        int totalTerms =
docNameTotalTerms.get(docName);
                                        double tf = (double) tFrequency /
totalTerms;
                                        double tfIdf = (double) tf *
termIdf.get(term);
                                        temp.put(docName, tfIdf);
                                } else {
                                        temp = new TreeMap<>();
                                        int totalTerms =
docNameTotalTerms.get(docName);
                                        double tf = (double) tFrequency /
totalTerms;
                                        double tfIdf = (double) tf *
termIdf.get(term);
                                        temp.put(docName, tfIdf);
                                        termtfIdf.put(term, temp);
                                }

                        }

                        System.out.println("All words TF*IDF values sorted by
terms of all documents");
                        sortbykey(termtfIdf);
                        System.out.println("Top Fifteen words TF*IDF values
sorted by tf*idf values in each document");

                        interChangeDocNameTerm(termtfIdf);
                        br1.close();
                        br2.close();
                        br3.close();
                } catch (IOException e) {
                        // TODO Auto-generated catch block
                        System.out.println("Erro Occured"+e.printStackTrace());
                }
        }
}
```

Formulae Used:

totalDocs (Total Number of documents) = *number of lines in phase2 output file*.
docWithTerm (Number of documents with term t in it) = *from phase 3 output file*.
Calculating IDF:
IDF = log(totalDocs / docWithTerm)
tFrequency (Number of times term t appears in a document) = *from phase1 output file*.
totalTerms (Total number of terms in the given document) = *from phase2 output file*.
Calculating TF:
TF = tFrequency / totalTerms
Calculating TF-IDF:
*tfidf = TF * IDF*

Results:

All words TF*IDF values sorted by terms of all documents.

Term	Doc Name	TF*IDF
11032	0002	0.01703274866152108
11032	0003	0.023374729546129996
11032	0009	0.01703274866152108
11032	0010	0.023374729546129996
11033	0002	0.00851637433076054
11033	0003	0.023374729546129996
11033	0009	0.00851637433076054
11033	0010	0.023374729546129996
1500s	0001	0.012072662512836372
1500s	0003	0.023374729546129996
1500s	0008	0.012072662512836372
1500s	0010	0.023374729546129996
1914	0003	0.03812254189846925
1914	0010	0.03812254189846925
1960s	0001	0.019689664497011594
1960s	0008	0.019689664497011594
200	0005	0.01480792949775252
200	0012	0.01480792949775252
2000	0002	0.013889608288589574
2000	0009	0.013889608288589574
45	0002	0.027779216577179147
45	0009	0.027779216577179147
Aldus	0001	0.019689664497011594
Aldus	0008	0.019689664497011594
Aliquam	0006	0.013371339322597426
Aliquam	0007	0.018098580497253082
All	0005	0.01480792949775252
All	0012	0.01480792949775252
BC	0002	0.027779216577179147
BC	0009	0.027779216577179147
Bonorum	0002	0.00851637433076054
Bonorum	0003	0.023374729546129996
Bonorum	0009	0.00851637433076054
Bonorum	0010	0.023374729546129996
Cicero	0002	0.00851637433076054
Cicero	0003	0.023374729546129996
Cicero	0009	0.00851637433076054
Cicero	0010	0.023374729546129996
College	0002	0.013889608288589574
College	0009	0.013889608288589574
Content	0004	0.017228456434885145
Content	0011	0.017228456434885145
Contrary	0002	0.013889608288589574
Contrary	0009	0.013889608288589574
Cras	0006	0.04011401796779227
Cras	0007	0.036197160994506164

Curabitur	0007	0.025100067169575763
Curae	0006	0.018544079476029852
Donec	0006	0.02674267864519485
Donec	0007	0.036197160994506164
Duis	0006	0.037088158952059705
English	0003	0.023374729546129996
English	0004	0.010563579698731826
English	0010	0.023374729546129996
English	0011	0.010563579698731826
Evil	0002	0.013889608288589574
Evil	0009	0.013889608288589574
Extremes	0002	0.013889608288589574
Extremes	0009	0.013889608288589574
Finibus	0002	0.00851637433076054
Finibus	0003	0.023374729546129996
Finibus	0009	0.00851637433076054
Finibus	0010	0.023374729546129996
Good	0002	0.013889608288589574
Good	0009	0.013889608288589574
H	0003	0.03812254189846925
H	0010	0.03812254189846925
HampdenSydney	0002	0.013889608288589574
HampdenSydney	0009	0.013889608288589574
If	0005	0.01480792949775252
If	0012	0.01480792949775252
Internet	0005	0.02961585899550504
Internet	0012	0.02961585899550504
Ipsum	0001	0.008014134364569433
Ipsum	0002	0.0056533816060140955
Ipsum	0003	0.0038791820594458425
Ipsum	0004	0.003506183784499127
Ipsum	0005	0.007533948627849363
Ipsum	0008	0.008014134364569433
Ipsum	0009	0.0056533816060140955
Ipsum	0010	0.0038791820594458425
Ipsum	0011	0.003506183784499127
Ipsum	0012	0.007533948627849363
It	0001	0.008911321057322294
It	0002	0.003143140372931507
It	0004	0.003898702962578504
It	0005	0.0033509513066790446
It	0008	0.008911321057322294
It	0009	0.003143140372931507
It	0011	0.003898702962578504
It	0012	0.0033509513066790446
Latin	0002	0.025549122992281622
Latin	0005	0.009079440402215784
Latin	0009	0.025549122992281622
Latin	0012	0.009079440402215784
Letraset	0001	0.019689664497011594
Letraset	0008	0.019689664497011594
Lorem	0001	0.003824675911632075
Lorem	0002	0.0033725339918461122

Lorem	0003	0.0018513058933963765
Lorem	0004	0.0016732957113390328
Lorem	0005	0.003595511445852467
Lorem	0006	6.493386342509679E-4
Lorem	0008	0.003824675911632075
Lorem	0009	0.0033725339918461122
Lorem	0010	0.0018513058933963765
Lorem	0011	0.0016732957113390328
Lorem	0012	0.003595511445852467
Maecenas	0007	0.025100067169575763
Malorum	0002	0.00851637433076054
Malorum	0003	0.023374729546129996
Malorum	0009	0.00851637433076054
Malorum	0010	0.023374729546129996
Many	0004	0.017228456434885145
Many	0011	0.017228456434885145
Mauris	0007	0.025100067169575763
McClintock	0002	0.013889608288589574
McClintock	0009	0.013889608288589574
Nulla	0006	0.013371339322597426
Nulla	0007	0.018098580497253082
Nunc	0007	0.025100067169575763
PageMaker	0001	0.019689664497011594
PageMaker	0008	0.019689664497011594
Pellentesque	0006	0.018544079476029852
Praesent	0006	0.018544079476029852
Rackham	0003	0.03812254189846925
Rackham	0010	0.03812254189846925
Renaissance	0002	0.013889608288589574
Renaissance	0009	0.013889608288589574
Richard	0002	0.013889608288589574
Richard	0009	0.013889608288589574
Sections	0003	0.03812254189846925
Sections	0010	0.03812254189846925
Sed	0007	0.025100067169575763
Suspendisse	0006	0.013371339322597426
Suspendisse	0007	0.018098580497253082
The	0002	0.006286280745863014
The	0003	0.008626917193790731
The	0004	0.003898702962578504
The	0005	0.0033509513066790446
The	0009	0.006286280745863014
The	0010	0.008626917193790731
The	0011	0.003898702962578504
The	0012	0.0033509513066790446
There	0005	0.01480792949775252
There	0012	0.01480792949775252
This	0002	0.013889608288589574
This	0009	0.013889608288589574
Ut	0006	0.018544079476029852
Various	0004	0.017228456434885145
Various	0011	0.017228456434885145
Vestibulum	0006	0.037088158952059705

Virginia	0002	0.013889608288589574
Virginia	0009	0.013889608288589574
a	0001	0.008911321057322294
a	0002	0.015715701864657535
a	0004	0.01949351481289252
a	0005	0.010052853920037134
a	0008	0.008911321057322294
a	0009	0.015715701864657535
a	0011	0.01949351481289252
a	0012	0.010052853920037134
ac	0006	0.02674267864519485
ac	0007	0.018098580497253082
accident	0004	0.017228456434885145
accident	0011	0.017228456434885145
accompanied	0003	0.03812254189846925
accompanied	0010	0.03812254189846925
adipiscing	0006	0.018544079476029852
aliquet	0006	0.02674267864519485
aliquet	0007	0.018098580497253082
also	0001	0.012072662512836372
also	0003	0.023374729546129996
also	0008	0.012072662512836372
also	0010	0.023374729546129996
alteration	0005	0.01480792949775252
alteration	0012	0.01480792949775252
always	0005	0.01480792949775252
always	0012	0.01480792949775252
amet	0002	0.00851637433076054
amet	0006	0.049191595014989986
amet	0007	0.011097093824930402
amet	0009	0.00851637433076054
an	0001	0.019689664497011594
an	0008	0.019689664497011594
and	0001	0.013366981585983442
and	0002	0.00942942111879452
and	0003	0.008626917193790731
and	0004	0.011696108887735512
and	0008	0.013366981585983442
and	0009	0.00942942111879452
and	0010	0.008626917193790731
and	0011	0.011696108887735512
ante	0006	0.02674267864519485
ante	0007	0.018098580497253082
anything	0005	0.01480792949775252
anything	0012	0.01480792949775252
arcu	0006	0.013371339322597426
arcu	0007	0.018098580497253082
are	0003	0.023374729546129996
are	0005	0.01815888080443157
are	0010	0.023374729546129996
are	0012	0.01815888080443157
as	0004	0.021127159397463652
as	0005	0.009079440402215784

as	0011	0.021127159397463652
as	0012	0.009079440402215784
at	0002	0.006786579359332557
at	0004	0.008417968628402883
at	0006	0.006533348786223133
at	0009	0.006786579359332557
at	0011	0.008417968628402883
augue	0006	0.013371339322597426
augue	0007	0.054295741491759246
available	0005	0.01480792949775252
available	0012	0.01480792949775252
be	0004	0.010563579698731826
be	0005	0.009079440402215784
be	0011	0.010563579698731826
be	0012	0.009079440402215784
been	0001	0.019689664497011594
been	0008	0.019689664497011594
belief	0002	0.013889608288589574
belief	0009	0.013889608288589574
believable	0005	0.01480792949775252
believable	0012	0.01480792949775252
below	0003	0.03812254189846925
below	0010	0.03812254189846925
bibendum	0006	0.018544079476029852
book	0001	0.012072662512836372
book	0002	0.00851637433076054
book	0008	0.012072662512836372
book	0009	0.00851637433076054
but	0001	0.012072662512836372
but	0005	0.009079440402215784
but	0008	0.012072662512836372
but	0012	0.009079440402215784
by	0002	0.003143140372931507
by	0003	0.025880751581372194
by	0004	0.007797405925157008
by	0005	0.0033509513066790446
by	0009	0.003143140372931507
by	0010	0.025880751581372194
by	0011	0.007797405925157008
by	0012	0.0033509513066790446
centuries	0001	0.019689664497011594
centuries	0008	0.019689664497011594
chunk	0003	0.03812254189846925
chunk	0010	0.03812254189846925
chunks	0005	0.01480792949775252
chunks	0012	0.01480792949775252
cites	0002	0.013889608288589574
cites	0009	0.013889608288589574
classical	0002	0.027779216577179147
classical	0009	0.027779216577179147
combined	0005	0.01480792949775252
combined	0012	0.01480792949775252
comes	0002	0.027779216577179147

comes	0009	0.027779216577179147
commodo	0006	0.018544079476029852
condimentum	0007	0.025100067169575763
congue	0007	0.025100067169575763
consectetur	0002	0.010746467915658066
consectetur	0006	0.02069096061372971
consectetur	0009	0.010746467915658066
consequat	0006	0.018544079476029852
containing	0001	0.019689664497011594
containing	0008	0.019689664497011594
content	0004	0.03445691286977029
content	0011	0.03445691286977029
convallis	0006	0.037088158952059705
cubilia	0006	0.018544079476029852
cursus	0006	0.018544079476029852
dapibus	0006	0.013371339322597426
dapibus	0007	0.018098580497253082
de	0002	0.00851637433076054
de	0003	0.023374729546129996
de	0009	0.00851637433076054
de	0010	0.023374729546129996
default	0004	0.017228456434885145
default	0011	0.017228456434885145
desktop	0001	0.012072662512836372
desktop	0004	0.010563579698731826
desktop	0008	0.012072662512836372
desktop	0011	0.010563579698731826
diam	0006	0.013371339322597426
diam	0007	0.054295741491759246
dictionary	0005	0.01480792949775252
dictionary	0012	0.01480792949775252
dictum	0006	0.018544079476029852
discovered	0002	0.013889608288589574
discovered	0009	0.013889608288589574
distracted	0004	0.017228456434885145
distracted	0011	0.017228456434885145
distribution	0004	0.017228456434885145
distribution	0011	0.017228456434885145
dolor	0002	0.00851637433076054
dolor	0006	0.008198599169164999
dolor	0007	0.011097093824930402
dolor	0009	0.00851637433076054
dont	0005	0.01480792949775252
dont	0012	0.01480792949775252
dui	0006	0.018544079476029852
dummy	0001	0.03937932899402319
dummy	0008	0.03937932899402319
during	0002	0.013889608288589574
during	0009	0.013889608288589574
editors	0004	0.017228456434885145
editors	0011	0.017228456434885145
electronic	0001	0.019689664497011594
electronic	0008	0.019689664497011594

elementum	0007	0.025100067169575763
elit	0006	0.037088158952059705
embarrassing	0005	0.01480792949775252
embarrassing	0012	0.01480792949775252
eros	0007	0.025100067169575763
essentially	0001	0.019689664497011594
essentially	0008	0.019689664497011594
est	0007	0.025100067169575763
established	0004	0.017228456434885145
established	0011	0.017228456434885145
et	0002	0.005373233957829033
et	0003	0.014747812352339261
et	0006	0.015518220460297282
et	0007	0.014002973344645361
et	0009	0.005373233957829033
et	0010	0.014747812352339261
etc	0005	0.01480792949775252
etc	0012	0.01480792949775252
ethics	0002	0.013889608288589574
ethics	0009	0.013889608288589574
eu	0006	0.02674267864519485
eu	0007	0.018098580497253082
euismod	0007	0.025100067169575763
even	0005	0.01480792949775252
even	0012	0.01480792949775252
ever	0001	0.019689664497011594
ever	0008	0.019689664497011594
evolved	0004	0.017228456434885145
evolved	0011	0.017228456434885145
ex	0007	0.025100067169575763
exact	0003	0.03812254189846925
exact	0010	0.03812254189846925
facilisi	0007	0.025100067169575763
facilisis	0007	0.025100067169575763
fact	0004	0.017228456434885145
fact	0011	0.017228456434885145
faucibus	0006	0.013371339322597426
faucibus	0007	0.036197160994506164
fermentum	0007	0.05020013433915153
feugiat	0006	0.02674267864519485
feugiat	0007	0.018098580497253082
first	0002	0.00851637433076054
first	0005	0.009079440402215784
first	0009	0.00851637433076054
first	0012	0.009079440402215784
five	0001	0.019689664497011594
five	0008	0.019689664497011594
for	0003	0.023374729546129996
for	0004	0.010563579698731826
for	0010	0.023374729546129996
for	0011	0.010563579698731826
form	0003	0.023374729546129996
form	0005	0.009079440402215784

form	0010	0.023374729546129996
form	0012	0.009079440402215784
free	0005	0.01480792949775252
free	0012	0.01480792949775252
fringilla	0006	0.02674267864519485
fringilla	0007	0.018098580497253082
from	0002	0.02149293583131613
from	0003	0.029495624704678522
from	0005	0.005728489095536738
from	0009	0.02149293583131613
from	0010	0.029495624704678522
from	0012	0.005728489095536738
galley	0001	0.019689664497011594
galley	0008	0.019689664497011594
generate	0005	0.01480792949775252
generate	0012	0.01480792949775252
generated	0005	0.01480792949775252
generated	0012	0.01480792949775252
generator	0005	0.01480792949775252
generator	0012	0.01480792949775252
generators	0005	0.01480792949775252
generators	0012	0.01480792949775252
going	0002	0.00851637433076054
going	0005	0.009079440402215784
going	0009	0.00851637433076054
going	0012	0.009079440402215784
gravida	0006	0.02674267864519485
gravida	0007	0.036197160994506164
handful	0005	0.01480792949775252
handful	0012	0.01480792949775252
has	0001	0.015234003968350448
has	0002	0.005373233957829033
has	0004	0.006664876736153321
has	0008	0.015234003968350448
has	0009	0.005373233957829033
has	0011	0.006664876736153321
have	0004	0.010563579698731826
have	0005	0.009079440402215784
have	0011	0.010563579698731826
have	0012	0.009079440402215784
here	0004	0.03445691286977029
here	0011	0.03445691286977029
hidden	0005	0.01480792949775252
hidden	0012	0.01480792949775252
humour	0004	0.010563579698731826
humour	0005	0.01815888080443157
humour	0011	0.010563579698731826
humour	0012	0.01815888080443157
id	0007	0.025100067169575763
imperdiet	0006	0.013371339322597426
imperdiet	0007	0.036197160994506164
in	0001	9.561689779080187E-4
in	0002	0.0033725339918461122

in	0003	0.0018513058933963765
in	0004	8.366478556695164E-4
in	0005	0.0014382045783409869
in	0006	0.0019480159027529037
in	0008	9.561689779080187E-4
in	0009	0.0033725339918461122
in	0010	0.0018513058933963765
in	0011	8.366478556695164E-4
in	0012	0.0014382045783409869
including	0001	0.019689664497011594
including	0008	0.019689664497011594
industry	0001	0.019689664497011594
industry	0008	0.019689664497011594
industrys	0001	0.019689664497011594
industrys	0008	0.019689664497011594
infancy	0004	0.017228456434885145
infancy	0011	0.017228456434885145
injected	0004	0.010563579698731826
injected	0005	0.01815888080443157
injected	0011	0.010563579698731826
injected	0012	0.01815888080443157
interested	0003	0.03812254189846925
interested	0010	0.03812254189846925
into	0001	0.019689664497011594
into	0008	0.019689664497011594
ipsum	0002	0.005373233957829033
ipsum	0004	0.006664876736153321
ipsum	0006	0.010345480306864855
ipsum	0007	0.0070014866723226805
ipsum	0009	0.005373233957829033
ipsum	0011	0.006664876736153321
is	0001	0.002003533591142358
is	0002	0.0028266908030070478
is	0003	0.0038791820594458425
is	0004	0.003506183784499127
is	0005	0.0015067897255698728
is	0008	0.002003533591142358
is	0009	0.0028266908030070478
is	0010	0.0038791820594458425
is	0011	0.003506183784499127
is	0012	0.0015067897255698728
isnt	0005	0.01480792949775252
isnt	0012	0.01480792949775252
it	0001	0.007617001984175224
it	0002	0.005373233957829033
it	0004	0.013329753472306641
it	0008	0.007617001984175224
it	0009	0.005373233957829033
it	0011	0.013329753472306641
its	0004	0.017228456434885145
its	0011	0.017228456434885145
justo	0006	0.013371339322597426
justo	0007	0.018098580497253082

lacinia	0006	0.013371339322597426
lacinia	0007	0.018098580497253082
lacus	0007	0.025100067169575763
layout	0004	0.017228456434885145
layout	0011	0.017228456434885145
leap	0001	0.019689664497011594
leap	0008	0.019689664497011594
leo	0006	0.013371339322597426
leo	0007	0.018098580497253082
letters	0004	0.017228456434885145
letters	0011	0.017228456434885145
libero	0006	0.018544079476029852
ligula	0007	0.025100067169575763
like	0001	0.012072662512836372
like	0004	0.021127159397463652
like	0008	0.012072662512836372
like	0011	0.021127159397463652
line	0002	0.027779216577179147
line	0009	0.027779216577179147
literature	0002	0.027779216577179147
literature	0009	0.027779216577179147
long	0004	0.017228456434885145
long	0011	0.017228456434885145
look	0004	0.010563579698731826
look	0005	0.009079440402215784
look	0011	0.010563579698731826
look	0012	0.009079440402215784
looked	0002	0.013889608288589574
looked	0009	0.013889608288589574
looking	0004	0.017228456434885145
looking	0011	0.017228456434885145
looks	0005	0.01480792949775252
looks	0012	0.01480792949775252
lorem	0004	0.017228456434885145
lorem	0011	0.017228456434885145
luctus	0006	0.037088158952059705
magna	0006	0.04011401796779227
magna	0007	0.018098580497253082
majority	0005	0.01480792949775252
majority	0012	0.01480792949775252
make	0001	0.019689664497011594
make	0008	0.019689664497011594
making	0002	0.005373233957829033
making	0004	0.006664876736153321
making	0005	0.005728489095536738
making	0009	0.005373233957829033
making	0011	0.006664876736153321
making	0012	0.005728489095536738
malesuada	0006	0.018544079476029852
many	0004	0.010563579698731826
many	0005	0.009079440402215784
many	0011	0.010563579698731826
many	0012	0.009079440402215784

massa	0006	0.013371339322597426
massa	0007	0.018098580497253082
mattis	0007	0.025100067169575763
mauris	0006	0.013371339322597426
mauris	0007	0.018098580497253082
metus	0006	0.018544079476029852
mi	0006	0.013371339322597426
mi	0007	0.036197160994506164
middle	0005	0.01480792949775252
middle	0012	0.01480792949775252
model	0004	0.010563579698731826
model	0005	0.009079440402215784
model	0011	0.010563579698731826
model	0012	0.009079440402215784
mollis	0006	0.018544079476029852
more	0001	0.012072662512836372
more	0002	0.00851637433076054
more	0008	0.012072662512836372
more	0009	0.00851637433076054
moreorless	0004	0.017228456434885145
moreorless	0011	0.017228456434885145
nec	0007	0.025100067169575763
necessary	0005	0.01480792949775252
necessary	0012	0.01480792949775252
need	0005	0.01480792949775252
need	0012	0.01480792949775252
nibh	0006	0.013371339322597426
nibh	0007	0.018098580497253082
nisi	0006	0.0534853572903897
nisi	0007	0.036197160994506164
nisl	0006	0.013371339322597426
nisl	0007	0.036197160994506164
non	0006	0.013371339322597426
non	0007	0.036197160994506164
noncharacteristic	0005	0.01480792949775252
noncharacteristic	0012	0.01480792949775252
normal	0004	0.017228456434885145
normal	0011	0.017228456434885145
not	0001	0.012072662512836372
not	0002	0.00851637433076054
not	0008	0.012072662512836372
not	0009	0.00851637433076054
now	0004	0.017228456434885145
now	0011	0.017228456434885145
nulla	0006	0.013371339322597426
nulla	0007	0.018098580497253082
nunc	0006	0.037088158952059705
obscure	0002	0.013889608288589574
obscure	0009	0.013889608288589574
of	0001	0.008014134364569433
of	0002	0.009893417810524668
of	0003	0.0038791820594458425
of	0004	0.005259275676748691

of	0005	0.009040738353419236
of	0008	0.008014134364569433
of	0009	0.009893417810524668
of	0010	0.0038791820594458425
of	0011	0.005259275676748691
of	0012	0.009040738353419236
old	0002	0.013889608288589574
old	0009	0.013889608288589574
on	0002	0.005373233957829033
on	0004	0.006664876736153321
on	0005	0.011456978191073476
on	0009	0.005373233957829033
on	0011	0.006664876736153321
on	0012	0.011456978191073476
one	0002	0.013889608288589574
one	0009	0.013889608288589574
only	0001	0.019689664497011594
only	0008	0.019689664497011594
opposed	0004	0.017228456434885145
opposed	0011	0.017228456434885145
or	0005	0.02961585899550504
or	0012	0.02961585899550504
orci	0006	0.037088158952059705
original	0003	0.03812254189846925
original	0010	0.03812254189846925
ornare	0006	0.018544079476029852
over	0002	0.005373233957829033
over	0004	0.006664876736153321
over	0005	0.005728489095536738
over	0009	0.005373233957829033
over	0011	0.006664876736153321
over	0012	0.005728489095536738
packages	0004	0.017228456434885145
packages	0011	0.017228456434885145
page	0004	0.03445691286977029
page	0011	0.03445691286977029
passage	0002	0.00851637433076054
passage	0005	0.009079440402215784
passage	0009	0.00851637433076054
passage	0012	0.009079440402215784
passages	0001	0.012072662512836372
passages	0005	0.009079440402215784
passages	0008	0.012072662512836372
passages	0012	0.009079440402215784
pellentesque	0006	0.018544079476029852
pharetra	0006	0.018544079476029852
piece	0002	0.013889608288589574
piece	0009	0.013889608288589574
placerat	0007	0.025100067169575763
point	0004	0.017228456434885145
point	0011	0.017228456434885145
popular	0002	0.027779216577179147
popular	0009	0.027779216577179147

popularised	0001	0.019689664497011594
popularised	0008	0.019689664497011594
porta	0007	0.025100067169575763
porttitor	0007	0.025100067169575763
posuere	0006	0.037088158952059705
predefined	0005	0.01480792949775252
predefined	0012	0.01480792949775252
pretium	0007	0.025100067169575763
primis	0006	0.018544079476029852
printer	0001	0.019689664497011594
printer	0008	0.019689664497011594
printing	0001	0.019689664497011594
printing	0008	0.019689664497011594
professor	0002	0.013889608288589574
professor	0009	0.013889608288589574
publishing	0001	0.012072662512836372
publishing	0004	0.010563579698731826
publishing	0008	0.012072662512836372
publishing	0011	0.010563579698731826
pulvinar	0007	0.025100067169575763
purpose	0004	0.017228456434885145
purpose	0011	0.017228456434885145
purus	0007	0.025100067169575763
quam	0006	0.018544079476029852
quis	0006	0.02674267864519485
quis	0007	0.018098580497253082
random	0002	0.013889608288589574
random	0009	0.013889608288589574
randomised	0005	0.01480792949775252
randomised	0012	0.01480792949775252
readable	0004	0.03445691286977029
readable	0011	0.03445691286977029
reader	0004	0.017228456434885145
reader	0011	0.017228456434885145
reasonable	0005	0.01480792949775252
reasonable	0012	0.01480792949775252
recently	0001	0.019689664497011594
recently	0008	0.019689664497011594
release	0001	0.019689664497011594
release	0008	0.019689664497011594
remaining	0001	0.019689664497011594
remaining	0008	0.019689664497011594
repeat	0005	0.01480792949775252
repeat	0012	0.01480792949775252
repetition	0005	0.01480792949775252
repetition	0012	0.01480792949775252
reproduced	0003	0.0762450837969385
reproduced	0010	0.0762450837969385
rhoncus	0007	0.025100067169575763
roots	0002	0.013889608288589574
roots	0009	0.013889608288589574
rutrum	0006	0.018544079476029852
sapien	0006	0.018544079476029852

scelerisque	0007	0.025100067169575763
scrambled	0001	0.019689664497011594
scrambled	0008	0.019689664497011594
search	0004	0.017228456434885145
search	0011	0.017228456434885145
section	0002	0.013889608288589574
section	0009	0.013889608288589574
sections	0002	0.013889608288589574
sections	0009	0.013889608288589574
sed	0006	0.013371339322597426
sed	0007	0.036197160994506164
sem	0006	0.018544079476029852
semper	0006	0.05563223842808956
sentence	0005	0.01480792949775252
sentence	0012	0.01480792949775252
sheets	0001	0.019689664497011594
sheets	0008	0.019689664497011594
simply	0001	0.012072662512836372
simply	0002	0.00851637433076054
simply	0008	0.012072662512836372
simply	0009	0.00851637433076054
since	0001	0.012072662512836372
since	0003	0.023374729546129996
since	0008	0.012072662512836372
since	0010	0.023374729546129996
sit	0002	0.00851637433076054
sit	0006	0.049191595014989986
sit	0007	0.011097093824930402
sit	0009	0.00851637433076054
sites	0004	0.017228456434885145
sites	0011	0.017228456434885145
slightly	0005	0.01480792949775252
slightly	0012	0.01480792949775252
sodales	0007	0.025100067169575763
software	0001	0.019689664497011594
software	0008	0.019689664497011594
sollicitudin	0007	0.05020013433915153
some	0005	0.01480792949775252
some	0012	0.01480792949775252
sometimes	0004	0.03445691286977029
sometimes	0011	0.03445691286977029
source	0002	0.013889608288589574
source	0009	0.013889608288589574
specimen	0001	0.019689664497011594
specimen	0008	0.019689664497011594
standard	0001	0.012072662512836372
standard	0003	0.023374729546129996
standard	0008	0.012072662512836372
standard	0010	0.023374729546129996
still	0004	0.017228456434885145
still	0011	0.017228456434885145
structures	0005	0.01480792949775252
structures	0012	0.01480792949775252

suffered	0005	0.01480792949775252
suffered	0012	0.01480792949775252
sure	0005	0.01480792949775252
sure	0012	0.01480792949775252
survived	0001	0.019689664497011594
survived	0008	0.019689664497011594
suscipit	0007	0.025100067169575763
tempor	0006	0.018544079476029852
tempus	0006	0.018544079476029852
tend	0005	0.01480792949775252
tend	0012	0.01480792949775252
text	0001	0.008911321057322294
text	0002	0.003143140372931507
text	0004	0.003898702962578504
text	0005	0.0033509513066790446
text	0008	0.008911321057322294
text	0009	0.003143140372931507
text	0011	0.003898702962578504
text	0012	0.0033509513066790446
that	0004	0.03445691286977029
that	0011	0.03445691286977029
the	0001	0.01202120154685415
the	0002	0.008480072409021143
the	0003	0.007758364118891685
the	0004	0.005259275676748691
the	0005	0.009040738353419236
the	0008	0.01202120154685415
the	0009	0.008480072409021143
the	0010	0.007758364118891685
the	0011	0.005259275676748691
the	0012	0.009040738353419236
their	0003	0.023374729546129996
their	0004	0.021127159397463652
their	0010	0.023374729546129996
their	0011	0.021127159397463652
theory	0002	0.013889608288589574
theory	0009	0.013889608288589574
there	0005	0.01480792949775252
there	0012	0.01480792949775252
therefore	0005	0.01480792949775252
therefore	0012	0.01480792949775252
this	0005	0.01480792949775252
this	0012	0.01480792949775252
those	0003	0.03812254189846925
those	0010	0.03812254189846925
through	0002	0.013889608288589574
through	0009	0.013889608288589574
tincidunt	0006	0.02674267864519485
tincidunt	0007	0.018098580497253082
to	0001	0.004455660528661147
to	0002	0.003143140372931507
to	0004	0.003898702962578504
to	0005	0.013403805226716178

to	0008	0.004455660528661147
to	0009	0.003143140372931507
to	0011	0.003898702962578504
to	0012	0.013403805226716178
took	0001	0.019689664497011594
took	0008	0.019689664497011594
tortor	0006	0.013371339322597426
tortor	0007	0.018098580497253082
translation	0003	0.03812254189846925
translation	0010	0.03812254189846925
treatise	0002	0.013889608288589574
treatise	0009	0.013889608288589574
true	0005	0.01480792949775252
true	0012	0.01480792949775252
turpis	0006	0.018544079476029852
type	0001	0.03937932899402319
type	0008	0.03937932899402319
typesetting	0001	0.03937932899402319
typesetting	0008	0.03937932899402319
ullamcorper	0006	0.018544079476029852
ultrices	0006	0.037088158952059705
ultricies	0006	0.018544079476029852
unchanged	0001	0.019689664497011594
unchanged	0008	0.019689664497011594
uncover	0004	0.017228456434885145
uncover	0011	0.017228456434885145
undoubtable	0002	0.013889608288589574
undoubtable	0009	0.013889608288589574
unknown	0001	0.019689664497011594
unknown	0008	0.019689664497011594
up	0002	0.013889608288589574
up	0009	0.013889608288589574
use	0004	0.010563579698731826
use	0005	0.009079440402215784
use	0011	0.010563579698731826
use	0012	0.009079440402215784
used	0003	0.03812254189846925
used	0010	0.03812254189846925
uses	0005	0.01480792949775252
uses	0012	0.01480792949775252
using	0004	0.03445691286977029
using	0011	0.03445691286977029
ut	0007	0.05020013433915153
variations	0005	0.01480792949775252
variations	0012	0.01480792949775252
varius	0007	0.025100067169575763
vehicula	0006	0.013371339322597426
vehicula	0007	0.036197160994506164
vel	0006	0.013371339322597426
vel	0007	0.018098580497253082
velit	0006	0.013371339322597426
velit	0007	0.018098580497253082
venenatis	0006	0.013371339322597426

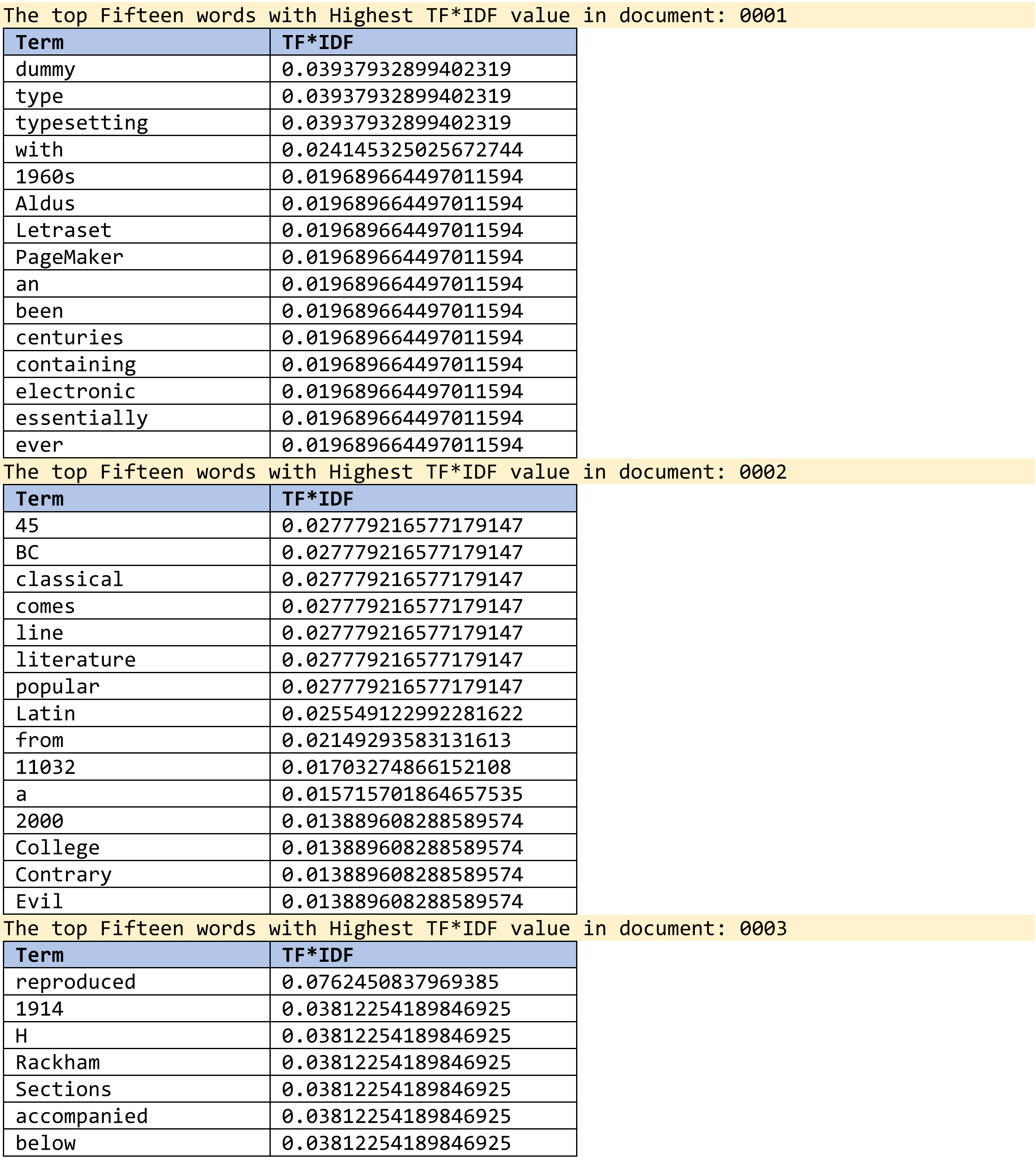
venenatis	0007	0.018098580497253082
versions	0001	0.007617001984175224
versions	0003	0.014747812352339261
versions	0004	0.006664876736153321
versions	0008	0.007617001984175224
versions	0010	0.014747812352339261
versions	0011	0.006664876736153321
very	0002	0.013889608288589574
very	0009	0.013889608288589574
vitae	0006	0.013371339322597426
vitae	0007	0.018098580497253082
viverra	0006	0.02674267864519485
viverra	0007	0.036197160994506164
volutpat	0007	0.025100067169575763
vulputate	0006	0.018544079476029852
was	0001	0.019689664497011594
was	0008	0.019689664497011594
web	0004	0.03445691286977029
web	0011	0.03445691286977029
when	0001	0.012072662512836372
when	0004	0.010563579698731826
when	0008	0.012072662512836372
when	0011	0.010563579698731826
which	0005	0.02961585899550504
which	0012	0.02961585899550504
will	0004	0.03445691286977029
will	0011	0.03445691286977029
with	0001	0.024145325025672744
with	0005	0.009079440402215784
with	0008	0.024145325025672744
with	0012	0.009079440402215784
word	0002	0.013889608288589574
word	0009	0.013889608288589574
words	0002	0.00851637433076054
words	0005	0.02723832120664735
words	0009	0.00851637433076054
words	0012	0.02723832120664735
written	0002	0.013889608288589574
written	0009	0.013889608288589574
years	0002	0.00851637433076054
years	0004	0.010563579698731826
years	0009	0.00851637433076054
years	0011	0.010563579698731826
you	0005	0.02961585899550504
you	0012	0.02961585899550504

b.) List of top fifteen words from each document having the highest TFIDF value. This selection of top 15 words may be done manually or by a sequential program

Code:

- Used same standalone program ("interChangeDocNameTerm" and "getTopFifteenSorted" are two main methods used to compute top 15 words in each doc having the highest tfidf)

Results:



The top Fifteen words with Highest TF*IDF value in document: 0001

Term	TF*IDF
dummy	0.03937932899402319
type	0.03937932899402319
typesetting	0.03937932899402319
with	0.024145325025672744
1960s	0.019689664497011594
Aldus	0.019689664497011594
Letraset	0.019689664497011594
PageMaker	0.019689664497011594
an	0.019689664497011594
been	0.019689664497011594
centuries	0.019689664497011594
containing	0.019689664497011594
electronic	0.019689664497011594
essentially	0.019689664497011594
ever	0.019689664497011594

The top Fifteen words with Highest TF*IDF value in document: 0002

Term	TF*IDF
45	0.027779216577179147
BC	0.027779216577179147
classical	0.027779216577179147
comes	0.027779216577179147
line	0.027779216577179147
literature	0.027779216577179147
popular	0.027779216577179147
Latin	0.025549122992281622
from	0.02149293583131613
11032	0.01703274866152108
a	0.015715701864657535
2000	0.013889608288589574
College	0.013889608288589574
Contrary	0.013889608288589574
Evil	0.013889608288589574

The top Fifteen words with Highest TF*IDF value in document: 0003

Term	TF*IDF
reproduced	0.0762450837969385
1914	0.03812254189846925
H	0.03812254189846925
Rackham	0.03812254189846925
Sections	0.03812254189846925
accompanied	0.03812254189846925
below	0.03812254189846925

chunk	0.03812254189846925
exact	0.03812254189846925
interested	0.03812254189846925
original	0.03812254189846925
those	0.03812254189846925
translation	0.03812254189846925
used	0.03812254189846925
from	0.029495624704678522

The top Fifteen words with Highest TF*IDF value in document: 0004

Term	TF*IDF
content	0.03445691286977029
here	0.03445691286977029
page	0.03445691286977029
readable	0.03445691286977029
sometimes	0.03445691286977029
that	0.03445691286977029
using	0.03445691286977029
web	0.03445691286977029
will	0.03445691286977029
as	0.021127159397463652
like	0.021127159397463652
their	0.021127159397463652
a	0.01949351481289252
Content	0.017228456434885145
Many	0.017228456434885145

The top Fifteen words with Highest TF*IDF value in document: 0005

Term	TF*IDF
Internet	0.02961585899550504
or	0.02961585899550504
which	0.02961585899550504
you	0.02961585899550504
words	0.02723832120664735
are	0.01815888080443157
humour	0.01815888080443157
injected	0.01815888080443157
200	0.01480792949775252
All	0.01480792949775252
If	0.01480792949775252
There	0.01480792949775252
alteration	0.01480792949775252
always	0.01480792949775252
anything	0.01480792949775252

The top Fifteen words with Highest TF*IDF value in document: 0006

Term	TF*IDF
semper	0.05563223842808956
nisi	0.0534853572903897
amet	0.049191595014989986
sit	0.049191595014989986
Cras	0.04011401796779227
magna	0.04011401796779227
Duis	0.037088158952059705
Vestibulum	0.037088158952059705
convallis	0.037088158952059705

elit	0.037088158952059705
luctus	0.037088158952059705
nunc	0.037088158952059705
orci	0.037088158952059705
posuere	0.037088158952059705
ultrices	0.037088158952059705

The top Fifteen words with Highest TF*IDF value in document: 0007

Term	TF*IDF
augue	0.054295741491759246
diam	0.054295741491759246
fermentum	0.05020013433915153
sollicitudin	0.05020013433915153
ut	0.05020013433915153
Cras	0.036197160994506164
Donec	0.036197160994506164
faucibus	0.036197160994506164
gravida	0.036197160994506164
imperdiet	0.036197160994506164
mi	0.036197160994506164
nisi	0.036197160994506164
nisl	0.036197160994506164
non	0.036197160994506164
sed	0.036197160994506164

The top Fifteen words with Highest TF*IDF value in document: 0008

Term	TF*IDF
dummy	0.03937932899402319
type	0.03937932899402319
typesetting	0.03937932899402319
with	0.024145325025672744
1960s	0.019689664497011594
Aldus	0.019689664497011594
Letraset	0.019689664497011594
PageMaker	0.019689664497011594
an	0.019689664497011594
been	0.019689664497011594
centuries	0.019689664497011594
containing	0.019689664497011594
electronic	0.019689664497011594
essentially	0.019689664497011594
ever	0.019689664497011594

The top Fifteen words with Highest TF*IDF value in document: 0009

Term	TF*IDF
45	0.027779216577179147
BC	0.027779216577179147
classical	0.027779216577179147
comes	0.027779216577179147
line	0.027779216577179147
literature	0.027779216577179147
popular	0.027779216577179147
Latin	0.025549122992281622
from	0.02149293583131613

11032	0.01703274866152108
a	0.015715701864657535
2000	0.013889608288589574
College	0.013889608288589574
Contrary	0.013889608288589574
Evil	0.013889608288589574

The top Fifteen words with Highest TF*IDF value in document: 0010

Term	TF*IDF
reproduced	0.0762450837969385
1914	0.03812254189846925
H	0.03812254189846925
Rackham	0.03812254189846925
Sections	0.03812254189846925
accompanied	0.03812254189846925
below	0.03812254189846925
chunk	0.03812254189846925
exact	0.03812254189846925
interested	0.03812254189846925
original	0.03812254189846925
those	0.03812254189846925
translation	0.03812254189846925
used	0.03812254189846925
from	0.029495624704678522

The top Fifteen words with Highest TF*IDF value in document: 0011

Term	TF*IDF
content	0.03445691286977029
here	0.03445691286977029
page	0.03445691286977029
readable	0.03445691286977029
sometimes	0.03445691286977029
that	0.03445691286977029
using	0.03445691286977029
web	0.03445691286977029
will	0.03445691286977029
as	0.021127159397463652
like	0.021127159397463652
their	0.021127159397463652
a	0.01949351481289252
Content	0.017228456434885145
Many	0.017228456434885145

The top Fifteen words with Highest TF*IDF value in document: 0012

Term	TF*IDF
Internet	0.02961585899550504
or	0.02961585899550504
which	0.02961585899550504
you	0.02961585899550504
words	0.02723832120664735
are	0.01815888080443157
humour	0.01815888080443157
injected	0.01815888080443157
200	0.01480792949775252
All	0.01480792949775252

If	0.01480792949775252
There	0.01480792949775252
alteration	0.01480792949775252
always	0.01480792949775252
anything	0.01480792949775252

Problem Statement:

2. Modify the programs to remove from consideration all those words that occur only once in each document. Repeat the tasks of Q1 above. Comment on any changes in the results of part (b).

Code:

- Just changed Reducer function in Phase1 mapreduce program from given code to remove from consideration all those words that occur only once as shown below.
- Used same standalone program as in question 1 to calculate TFIDF of terms by changing path to new output files.

```
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable>
    {

      public void reduce(Text key, Iterable<IntWritable> values, Context context)
          throws IOException, InterruptedException
          {
            int sum = 0;
            for (IntWritable val : values)
            {
                sum += val.get();
            }
            //Added condition to consider that occur more than once
            if (sum > 1) {
                    context.write(key, new IntWritable(sum));
            }

          }
    }
```

Results:

All words TF*IDF values sorted by terms that occur more than once in each doc for all documents.

Term	Doc Name	TF*IDF value
11032	0002	0.05779869255574371
11032	0009	0.05779869255574371
45	0002	0.05779869255574371
45	0009	0.05779869255574371
BC	0002	0.05779869255574371
BC	0009	0.05779869255574371
Cras	0006	0.0716703787691222
Cras	0007	0.0943031299593713
Donec	0006	0.04778025251274814

Donec	0007	0.0943031299593713
Duis	0006	0.06626417732768002
Internet	0005	0.0716703787691222
Internet	0012	0.0716703787691222
Ipsum	0001	0.04383406574142318
Ipsum	0002	0.026159039232784797
Ipsum	0004	0.016894379504506847
Ipsum	0005	0.04054651081081644
Ipsum	0008	0.04383406574142318
Ipsum	0009	0.026159039232784797
Ipsum	0011	0.016894379504506847
Ipsum	0012	0.04054651081081644
It	0001	0.0968518632015165
It	0008	0.0968518632015165
Latin	0002	0.08669803883361556
Latin	0009	0.08669803883361556
Lorem	0001	0.04383406574142318
Lorem	0002	0.032698799040981
Lorem	0004	0.016894379504506847
Lorem	0005	0.04054651081081644
Lorem	0008	0.04383406574142318
Lorem	0009	0.032698799040981
Lorem	0011	0.016894379504506847
Lorem	0012	0.04054651081081644
The	0002	0.05779869255574371
The	0009	0.05779869255574371
Vestibulum	0006	0.06626417732768002
a	0001	0.02191703287071159
a	0002	0.032698799040981
a	0004	0.04223594876126713
a	0005	0.024327906486489862
a	0008	0.02191703287071159
a	0009	0.032698799040981
a	0011	0.04223594876126713
a	0012	0.024327906486489862
ac	0006	0.06626417732768002
aliquet	0006	0.06626417732768002
amet	0006	0.19879253198304003
and	0001	0.056201122748103675
and	0002	0.03353937970451348
and	0004	0.04332169878499658
and	0008	0.056201122748103675
and	0009	0.03353937970451348
and	0011	0.04332169878499658
ante	0006	0.06626417732768002
are	0005	0.0716703787691222
are	0012	0.0716703787691222
as	0004	0.07465664455116895
as	0011	0.07465664455116895
augue	0007	0.19617684077273687
by	0003	0.3662040962227032
by	0004	0.0457755120278379
by	0010	0.3662040962227032

by	0011	0.0457755120278379
classical	0002	0.05779869255574371
classical	0009	0.05779869255574371
comes	0002	0.05779869255574371
comes	0009	0.05779869255574371
consectetur	0006	0.06626417732768002
content	0004	0.07465664455116895
content	0011	0.07465664455116895
convallis	0006	0.06626417732768002
diam	0007	0.19617684077273687
dummy	0001	0.0968518632015165
dummy	0008	0.0968518632015165
elit	0006	0.06626417732768002
et	0006	0.0716703787691222
et	0007	0.0943031299593713
eu	0006	0.06626417732768002
faucibus	0007	0.1307845605151579
fermentum	0007	0.1307845605151579
feugiat	0006	0.06626417732768002
fringilla	0006	0.06626417732768002
from	0002	0.0708782121721361
from	0003	0.24413606414846883
from	0009	0.0708782121721361
from	0010	0.24413606414846883
gravida	0006	0.04778025251274814
gravida	0007	0.0943031299593713
has	0001	0.0968518632015165
has	0008	0.0968518632015165
here	0004	0.07465664455116895
here	0011	0.07465664455116895
humour	0005	0.0716703787691222
humour	0012	0.0716703787691222
imperdiet	0007	0.1307845605151579
in	0002	0.0706023175285403
in	0005	0.035018749494155996
in	0006	0.035018749494155996
in	0009	0.0706023175285403
in	0012	0.035018749494155996
injected	0005	0.0716703787691222
injected	0012	0.0716703787691222
ipsum	0006	0.06626417732768002
is	0002	0.03543910608606805
is	0004	0.0457755120278379
is	0009	0.03543910608606805
is	0011	0.0457755120278379
it	0004	0.07465664455116895
it	0011	0.07465664455116895
like	0004	0.07465664455116895
like	0011	0.07465664455116895
line	0002	0.05779869255574371
line	0009	0.05779869255574371
literature	0002	0.05779869255574371
literature	0009	0.05779869255574371

luctus	0006	0.06626417732768002
magna	0006	0.09939626599152002
mi	0007	0.1307845605151579
nisi	0006	0.09556050502549628
nisi	0007	0.0943031299593713
nisl	0007	0.1307845605151579
non	0007	0.1307845605151579
nunc	0006	0.06626417732768002
of	0001	0.04383406574142318
of	0002	0.0457783186573734
of	0004	0.025341569256760274
of	0005	0.048655812972979724
of	0008	0.04383406574142318
of	0009	0.0457783186573734
of	0011	0.025341569256760274
of	0012	0.048655812972979724
on	0005	0.0716703787691222
on	0012	0.0716703787691222
or	0005	0.0716703787691222
or	0012	0.0716703787691222
orci	0006	0.06626417732768002
page	0004	0.07465664455116895
page	0011	0.07465664455116895
popular	0002	0.05779869255574371
popular	0009	0.05779869255574371
posuere	0006	0.06626417732768002
quis	0006	0.06626417732768002
readable	0004	0.07465664455116895
readable	0011	0.07465664455116895
reproduced	0003	0.39816877093956776
reproduced	0010	0.39816877093956776
sed	0007	0.1307845605151579
semper	0006	0.09939626599152002
sit	0006	0.19879253198304003
sollicitudin	0007	0.1307845605151579
sometimes	0004	0.07465664455116895
sometimes	0011	0.07465664455116895
text	0001	0.0968518632015165
text	0008	0.0968518632015165
that	0004	0.07465664455116895
that	0011	0.07465664455116895
the	0001	0.029565657858479123
the	0002	0.017644021625221412
the	0003	0.040515901509767686
the	0004	0.011395097299622162
the	0005	0.02187858681527455
the	0008	0.029565657858479123
the	0009	0.017644021625221412
the	0010	0.040515901509767686
the	0011	0.011395097299622162
the	0012	0.02187858681527455
their	0004	0.07465664455116895
their	0011	0.07465664455116895

tincidunt	0006	0.06626417732768002
to	0005	0.1433407575382444
to	0012	0.1433407575382444
type	0001	0.0968518632015165
type	0008	0.0968518632015165
typesetting	0001	0.0968518632015165
typesetting	0008	0.0968518632015165
ultrices	0006	0.06626417732768002
using	0004	0.07465664455116895
using	0011	0.07465664455116895
ut	0007	0.1307845605151579
vehicula	0007	0.1307845605151579
viverra	0006	0.04778025251274814
viverra	0007	0.0943031299593713
web	0004	0.07465664455116895
web	0011	0.07465664455116895
which	0005	0.0716703787691222
which	0012	0.0716703787691222
will	0004	0.07465664455116895
will	0011	0.07465664455116895
with	0001	0.0968518632015165
with	0008	0.0968518632015165
words	0005	0.1075055681536833
words	0012	0.1075055681536833
you	0005	0.0716703787691222
you	0012	0.0716703787691222

Results for part-b:

The top Fifteen words with Highest TF*IDF value in document: 0001

Term	TF*IDF
It	0.0968518632015165
dummy	0.0968518632015165
has	0.0968518632015165
text	0.0968518632015165
type	0.0968518632015165
typesetting	0.0968518632015165
with	0.0968518632015165
and	0.056201122748103675
Ipsum	0.04383406574142318
Lorem	0.04383406574142318
of	0.04383406574142318
the	0.029565657858479123
a	0.02191703287071159

The top Fifteen words with Highest TF*IDF value in document: 0002

Term	TF*IDF
Latin	0.08669803883361556
from	0.0708782121721361
in	0.0706023175285403
11032	0.05779869255574371
45	0.05779869255574371

BC	0.05779869255574371
The	0.05779869255574371
classical	0.05779869255574371
comes	0.05779869255574371
line	0.05779869255574371
literature	0.05779869255574371
popular	0.05779869255574371
of	0.0457783186573734
is	0.03543910608606805
and	0.03353937970451348

The top Fifteen words with Highest TF*IDF value in document: 0003

Term	**TF*IDF**
reproduced	0.39816877093956776
by	0.3662040962227032
from	0.24413606414846883
the	0.040515901509767686

The top Fifteen words with Highest TF*IDF value in document: 0004

Term	**TF*IDF**
as	0.07465664455116895
content	0.07465664455116895
here	0.07465664455116895
it	0.07465664455116895
like	0.07465664455116895
page	0.07465664455116895
readable	0.07465664455116895
sometimes	0.07465664455116895
that	0.07465664455116895
their	0.07465664455116895
using	0.07465664455116895
web	0.07465664455116895
will	0.07465664455116895
by	0.0457755120278379
is	0.0457755120278379

The top Fifteen words with Highest TF*IDF value in document: 0005

Term	**TF*IDF**
to	0.1433407575382444
words	0.1075055681536833
Internet	0.0716703787691222
are	0.0716703787691222
humour	0.0716703787691222
injected	0.0716703787691222
on	0.0716703787691222
or	0.0716703787691222
which	0.0716703787691222
you	0.0716703787691222
of	0.048655812972979724
Ipsum	0.04054651081081644
Lorem	0.04054651081081644
in	0.035018749494155996
a	0.024327906486489862

The top Fifteen words with Highest TF*IDF value in document: 0006

Term	**TF*IDF**

amet	0.19879253198304003
sit	0.19879253198304003
magna	0.09939626599152002
semper	0.09939626599152002
nisi	0.09556050502549628
Cras	0.0716703787691222
et	0.0716703787691222
Duis	0.06626417732768002
Vestibulum	0.06626417732768002
ac	0.06626417732768002
aliquet	0.06626417732768002
ante	0.06626417732768002
consectetur	0.06626417732768002
convallis	0.06626417732768002
elit	0.06626417732768002

The top Fifteen words with Highest TF*IDF value in document: 0007

Term	TF*IDF
augue	0.19617684077273687
diam	0.19617684077273687
faucibus	0.1307845605151579
fermentum	0.1307845605151579
imperdiet	0.1307845605151579
mi	0.1307845605151579
nisl	0.1307845605151579
non	0.1307845605151579
sed	0.1307845605151579
sollicitudin	0.1307845605151579
ut	0.1307845605151579
vehicula	0.1307845605151579
Cras	0.0943031299593713
Donec	0.0943031299593713
et	0.0943031299593713

The top Fifteen words with Highest TF*IDF value in document: 0008

Term	TF*IDF
It	0.0968518632015165
dummy	0.0968518632015165
has	0.0968518632015165
text	0.0968518632015165
type	0.0968518632015165
typesetting	0.0968518632015165
with	0.0968518632015165
and	0.056201122748103675
Ipsum	0.04383406574142318
Lorem	0.04383406574142318
of	0.04383406574142318
the	0.029565657858479123
a	0.02191703287071159

The top Fifteen words with Highest TF*IDF value in document: 0009

Term	TF*IDF
Latin	0.08669803883361556
from	0.0708782121721361
in	0.0706023175285403

11032	0.05779869255574371
45	0.05779869255574371
BC	0.05779869255574371
The	0.05779869255574371
classical	0.05779869255574371
comes	0.05779869255574371
line	0.05779869255574371
literature	0.05779869255574371
popular	0.05779869255574371
of	0.0457783186573734
is	0.03543910608606805
and	0.03353937970451348

The top Fifteen words with Highest TF*IDF value in document: 0010

Term	TF*IDF
reproduced	0.39816877093956776
by	0.3662040962227032
from	0.24413606414846883
the	0.040515901509767686

The top Fifteen words with Highest TF*IDF value in document: 0011

Term	TF*IDF
as	0.07465664455116895
content	0.07465664455116895
here	0.07465664455116895
it	0.07465664455116895
like	0.07465664455116895
page	0.07465664455116895
readable	0.07465664455116895
sometimes	0.07465664455116895
that	0.07465664455116895
their	0.07465664455116895
using	0.07465664455116895
web	0.07465664455116895
will	0.07465664455116895
by	0.0457755120278379
is	0.0457755120278379

The top Fifteen words with Highest TF*IDF value in document: 0012

Term	TF*IDF
to	0.1433407575382444
words	0.1075055681536833
Internet	0.0716703787691222
are	0.0716703787691222
humour	0.0716703787691222
injected	0.0716703787691222
on	0.0716703787691222
or	0.0716703787691222
which	0.0716703787691222
you	0.0716703787691222
of	0.048655812972979724
Ipsum	0.04054651081081644
Lorem	0.04054651081081644
in	0.035018749494155996
a	0.024327906486489862

Comments:

- TF*IDF values are increased for same words → (So if we want to rank with decent threshold we are increasing TFIDF values to same words by ignoring rare words/ noise words.)
- Words which does not appear in output (top 15 in each doc) of ques-1 but are actually important to consider came up in output (top 15 in each doc) of ques-2.
- Observed that removing rare words is important because they're so rare, the association between them and other words is dominated by noise**.** Here as documents size are small, we have considered Threshold -1 i.e., removing words that occur only once.

Note:

- Code files are attached as zip for reference.