N-GRAM IMPLEMENTATION REPORT

MUSTAFA ÇAĞATAY ÖNAL

2016510082

INTRODUCTION

1.1. Problem Definition

This project aims to implement the n-gram algorithms into different program languages (java and python) and extract some useful information from a part of Turkish Noval Corpus, which consist of 5 different novels and compare this test result with each other.

N-grams of texts are extensively used in text mining and natural language processing tasks. They are basically a set of co-occurring words within a given text and when computing the n-grams you typically move n word forward. You can think of an N-gram as the sequence of N words, by that notion, a 2-gram (or bigram) is a two-word sequence of words like "please turn", "turn your", or "your homework", and a 3-gram (or trigram) is a three-word sequence of words like "please turn your", or "turn your homework".

1.2. Project Scope

In this project, 1-gram, 2-gram and 3-gram found given novels as smooth as possible. First, the project was implemented in java and obtained top 100 items in each n-gram and running time of the program for each n-gram. Then, the same algorithm is implemented in python and their test result compared.

1.3. Implementation

Instead of taking it from the internet, I used my implementation and code. First, the file_read function reads 5 files one by one. This function takes the file name as String and with "ISO-8859-9" code, reads turkish characters correctly. Read file line by line and store data in String variable and return this variable at the end. In the main, all files data stores in the variable called text with file_read function.

1. file read function

Then, the split_word function used for split data correctly to use for n-gram algorithms. This function first organizes data before splitting it. Because splitting correctly and doing this process as smoothly as possible is very important for further process and obtaining correct results. All punctuations are removed from data and data parsed according to the spaces. This function return array contains words, in the main data stored in a string array variable called words.

```
public static String[] split_words(String data) {
    data = data.replaceAll("[^a-zA-Zzçĕöṣüıîâ1-9 ]", "");
    data = data.replaceAll(" ", " ");
    data = data.replaceAll(" ", " ");
    return data.split(" ");
}
```

2. split word function

Then the find_n_gram function is called and taken as an input this string array that stores splitted data and this function takes number n as an input which is 1,2 and 3 used in this project. This function first creates an arraylist from a class n_gram_array which is created for this project. This class consists of a string which holds n_gram and integer value of this n_gram which is how many times this n_gram consists in this project. List created from this class will hold all n-gram information. Then compare each word with each other and count their values and store this data in an array list. Finally, sort n-grams based on their values in descending order and print op 100 items with their values.

```
lic static void find_n_grams(String[] words, int n) {
   System.out.println("Top 100 items in " + n + "-gram");
   List<n_gram_array> n_grams = new ArrayList<n_gram_array>();
   for(int i = 0; i < (words.length - n + 1); i++) {
      boolean flag = true;
      String n_gram = "";
      if(n = -1);</pre>
      if(n == 1) {
      n_gram = String.join(" ", words[i], words[i+1], words[i+2]);
      for(int j = 0; j < n_grams.size(); j++) {
    if(n_grams.get(j).getN_gram().equals(n_gram)) {
        n_grams.get(j).setValue(n_grams.get(j).getValue() + 1);</pre>
                   flag = false;
            }
if(!flag) {
      if(flag) {
           n_grams.add(new n_gram_array(n_gram, 1));
for(int k = 0; k < 100; k++) {
       int max_value = 0;
      int max_id = 0;
      for(int j = 0; j < n_grams.size(); j++) {</pre>
            if(n_grams.get(j).getValue() > max_value) {
    max_value = n_grams.get(j).getValue();
                  max_id = j;
       .
System.out.println(k+1 + ": " + n_grams.get(max_id).getN_gram() + " - " + n_grams.get(max_id).getValue());
      n_grams.remove(max_id);
```

3. find n gram function

```
public class n_gram_array {
    private String n_gram;
    private int value;

    public n_gram_array(String n_gram, int value) {
        super();
        this.n_gram = n_gram;
        this.value = value;
    }

    public String getN_gram() {
        return n_gram;
    }

    public void setN_gram(String n_gram) {
        this.n_gram = n_gram;
    }

    public int getValue() {
        return value;
    }

    public void setValue(int value) {
        this.value = value;
    }
}
```

4. n gram array class

In the main running times calculated for reading data and splitting words and for each n-grams.

```
ublic static void main(String[] args) throws FileNotFoundException, IOException {
  long start1 = System.currentTimeMillis();
  String text = file_read("UNUTULMUŞ DİYARLAR.txt").toLowerCase() + file_read("BİLİM İŞ BAŞINDA.txt").toLowerCase()
          + file_read("BOZKIRDA.txt").toLowerCase() + file_read("DEĞİŞİM.txt").toLowerCase() + file_read("DENEMELER.txt").toLowerCase();
  String[] words = split_words(text);
  long start2 = System.currentTimeMillis();
  find n grams(words, 1);
  long stop1 = System.currentTimeMillis();
  find n grams(words, 2);
  long stop2 = System.currentTimeMillis();
  find_n_grams(words, 3);
  long stop3 = System.currentTimeMillis();
  System.out.println("Running time for 1-gram calculation: " + (stop1 - start2) / 1000 + " second\n"
          + "Running time for 2-gram calculation: " + (stop2 - stop1) / 1000 + " second\n"
          + "Running time for 3-gram calculation: " + (stop3 - stop1) / 1000 + " second");
  System.out.println("Runtime for reading data and spliting words: " + (start2 - start1) / 1000 + " second");
  System.out.println("Total runtime: " + (stop3 - start1) / 1000 + " second");
  System.out.println("Total word count: " + words.length);
```

5. main class

Then the same algorithm is implemented in python programming language.

1.4. Result

The program reads 5 txt. file and total number of words in java is 100256, total number of words in python is 101122.

```
Time for 1-gram: 24.1795158
Time for 2-gram: 74.1579864
Time for 3-gram: 67.31983100000
Total runtime: 165.6573332
Total word count: 101122

Running time for 1-gram calculation: 7 second
Running time for 2-gram calculation: 64 second
Running time for 3-gram calculation: 74 second
Running time for reading data and splitting words: 3 second
Total word count: 100256
```

6. python runtime

7. java runtime

1-gram calculation in java is 7 second and 24 second in python. So, java program calculate faster than python for 1-gram calculation. 2-gram calculation in java is 64 second

and 2-gram calculation in python is 74 second. 3-gram calculation in java is 74 second and 3-gram calculation in python is 67 second. We can clearly see that java implementation is slightly faster than python implementation, 6 second faster in total runtime. Reading data from input text files and splitting data into words took 3 second in java.

1.4.1. Unigram - Top 100

```
Top 100 items in 1-gram
                          50: az - 143
                                                             3502
                                                                                   161
                                                    bir
1: bir - 3485
                         51: var - 140
                                                2 : ve
                                                            1731
                                                                       46 : gregor
2: ve - 1726
                         52: bana - 138
                                                3 : bu
                                                            1134
3: bu - 1115
                         53: nasıl - 136
                                                4 : de
                                                            903
4: de - 900
                                                                       48 : olarak
                          54: ancak - 131
                                                            789
5: da - 786
                         55: yalnız - 129
                                                              700
6: daha - 698
                          56: olduğunu - 129
                                                6 : daha
                         57: bizim - 123
58: dedi - 122
59: aynı - 121
7: için - 677
                                                              677
                                                                       51 : bana - 146
8: ne - 618
9: gibi - 515
                                                                       53 : az - 145
10: ki - 510
                         60: olduğu - 120
                                                10 : ki - 515
                         61: fazla - 119
11: kadar - 495
                                                11 : kadar
                                                            - 497
                         62: yine - 119
12: her - 460
                                                                       55 : nasıl - 138
13: o - 458
                         63: üzerinde - 118
14: çok - 435
                         64: m1 - 118
                                                                       57 : olduğunu - 129
15: ama - 424
                         65: hem - 117
                                                14 : çok
                         66: göre - 115
67: önce - 114
16: diye - 418
                                                15 : ama
                                                                       59 : vasili - 126
17: en - 383
                                                16 : diye -
                                                               418
18: kendi - 317
                         68: benim - 112
                                                                       60 : bizim - 124
                                                17: en - 384
                         69: bunu - 112
19: sonra - 294
20: sey - 284
                                                18 : kendi -
                         70: karşı - 111
21: zaman - 264
                         71: ile - 109
                                                19 : sonra - 305
22: bütün - 255
                         72: yere - 108
                                                   : şey - 285
                                                20
                                                                       64 : olduğu -
23: insan - 252
                          73: arasında - 108
                                                21 : zaman -
                                                                264
24: hiç - 244
                          74: hep - 108
                                                22 : bütün -
                                                                258
                          75: ilk - 104
76: biri - 104
25: iyi - 234
                                                                       66 : fazla - 119
26: büyük - 234
                                                                       67: hem - 118
27: başka - 233
                          77: biraz - 102
                                                24 : hiç - 246
                                                                       68 : Üzerinde - 118
                          78: yeni - 101
28: onu - 220
                                                                       69 : benim - 117
                          79: fakat - 100
29: çünkü - 216
                                                26 : büyük - 235
                         80: gün - 100
30: ben - 215
                                                27 : başka - 234
31: değil - 213
                         81: şimdi - 100
                                                                       71 : göre -
                                                28 : ben - 223
32: öyle - 212
                         82: uzun - 99
                                                29 : onu
33: onun - 204
                         83: olur - 99
                                                                       73 : ile - 110
                                                   : çünkü - 218
34: olan - 203
                         84: yok - 98
                                                                       74 : karşı - 110
35: pek - 199
                          85: türlü - 98
                                                31 : değil -
                                                               215
36: bile - 196
                         86: bizi - 97
87: kez - 96
                                                                       75 : fakat - 110
                                                32 : öyle -
                                                               212
37: kitap - 192
                                                33 : onun
38: doğru - 186
                         88: artik - 94
                                                                       77 : yere - 108
                                                34 : olan
                                                               204
39: bölüm - 177
                         89: hemen - 94
                                                                       78 : arasında - 107
                                                35 : pek
40: böyle - 170
                         90: üstüne - 93
                                                                       79 : biraz - 104
                                                36 : bile -
                                                               197
41: hiçbir - 167
                         91: güzel - 92
                                                                       80 : ilk - 104
42: kendini - 162
                         92: üzerine - 92
                                                37 : kitap - 195
43: mi - 160
                          93: gregorun - 92
                                                                       81 : biri - 104
                                                38 : doğru -
                                                                187
44: ona - 157
                          94: kızkardeşi - 92
                                                39 : bölüm
                                                                177
45: içinde - 153
                          95: sen - 91
                                                40 : böyle
                                                                171
46: olarak - 152
                          96: adam - 88
                                                41 : hiçbir
                                                                 168
47: gregor - 150
                          97: yerde - 88
48: iki - 149
                                                42 : mi - 164
                          98: insanın - 88
49: ya - 146
                          99: şeyi - 87
                                                                       86 : artik - 100
                                                43 : kendini
                                                                  164
50: az - 143
                         100: bize - 87
                                                                       87 : olur - 99
                                                   : dedi
```

1.4.2. Digram - Top 100

			1.1	4		
	bir şey - 126	0.00	bir süre - 26	1 : bir şey - 127	7000	: ona - 161
	o kadar - 94		sonra da - 26	2 : o kadar - 94	46	: gregor - 159
	ya da - 84	0.00	yeni bir - 26	3 : ya da - 84	47	: içinde - 155
0.000	ne kadar - 76		her gün - 26	4 : ne kadar - 76	48	: olarak - 152
	büyük bir - 65	1000000	en iyi - 25	5 : büyük bir - 65	- 33.00	: iki - 151
	kitap 3 - 65		bir gün - 25	6 : kitap 3 - 65		
000	3 bölüm - 65	0.000	gibi bir - 25			: ya - 147
	başka bir - 60	D 20 0	bir başka - 25	7 : 3 bölüm – 65	51	: bana - 146
	hiç de - 58	100000000000000000000000000000000000000	olduğu için - 25	8 : başka bir - 60	52	: var - 145
1000000	1 bölüm - 56	1 M 6	bir tavırla - 24	9 : hiç de – 58	53	: az - 145
1000	2 bölüm - 55	12 mm m	diye bağırdı - 24	10 : kitap 1 - 56	54	: yakov - 145
the same of the	kitap 1 - 55	0.0	müdür bey - 24	11 : 1 bölüm - 56	2000	: nasıl - 138
	kitap 2 - 54	CO. C. C.	her şeyin - 24	12 : diye sordu - 55		
2200	bu kadar - 52	0	der ki - 24	13 : kitap 2 - 55		: ancak - 134
200000	hem de - 52		ve bir - 23	50	57	: olduğunu - 129
000000000000000000000000000000000000000	her şeyi - 50	2000	bir türlü - 23	14 : 2 bölüm - 55	58	: yalnız - 129
	böyle bir - 49	100000000000000000000000000000000000000	en büyük - 23	15 : bu kadar - 53	59	: vasili - 126
ACCORDS:	daha çok - 47	The same of	bir insan - 23	16 : hem de - 52	10000	: bizim - 124
00000	bir sesle - 46	T 100 C	güzel bir - 23	17 : her şeyi - 50		: aynı - 123
	her sey - 44	1.00	bir ses - 22	18 : böyle bir - 49	10000	
	daha fazla - 43	1000	ve daha - 22	19 : daha çok - 47		: m1 - 123
0.00	yine de - 43		pek çok - 22	20 : bir sesle - 46		: malva - 121
100000	her zaman - 40	10000	bir şeyler - 22		64	: olduğu - 120
	daha iyi - 40	0.000	ne var - 22	21 : yine de - 44	65	: yine - 120
	çok daha - 37	1000000	bir sürü - 22	22 : her şey - 44	66	: fazla - 119
0.00	ne de - 36	1000	herhangi bir - 21	23 : daha fazla - 43		: hem - 118
	iyi bir - 36	0.00	bir kez - 21	24 : daha iyi - 40		: üzerinde - 118
	diye sordu - 35	100000	diye karşılık - 21	25 : her zaman - 40		
9000	o zaman - 35	0.0000000000000000000000000000000000000	bir iş - 21	26 : çok daha - 37	1414	: benim - 117
10000	ki bu - 34	F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F 100 F	hafif bir - 20	27 : iyi bir - 36	70	: önce - 115
F 10 2 1 1 1 1 1	için de - 32	1.00	en az - 20	28 : ne de - 36	71	: göre - 115
TO 100 C TO	bir an - 32	D 25 0	kendi kendine - 20	55 F. 197 F. St. 11.	72	: bunu - 114
1011/45/00	kimi zaman - 32	B	çok kez - 20	29 : o zaman - 36	73	: ile - 110
	daha az - 31	64:	ağır ağır - 19	30 : ki bu - 34	7/	: karşı - 110
	ben de - 31		da bir - 19	31 : diye bağırdı - 34		
B000 100	en çok - 29	STATE OF	bu konuda - 19	32 : için de - 32	1.000	: fakat - 110
60600	belki de - 29		sağa sola - 19	33 : kimi zaman - 32		: hep - 110
0.000	bir biçimde - 29	V 200 / 200		34 : bir an - 32	77	: yere - 108
	o da - 29 bir tek - 29	100000000000000000000000000000000000000	bu işi - 19	35 : ben de - 32	78	: arasında - 107
D	daha da - 28	1000000	biraz daha - 19	and the state of t	79	: biraz – 104
2 (60 (7)	için bir - 28	1000	başka türlü - 19	36 : daha az - 31		: ilk - 104
100000000000000000000000000000000000000	hiçbir şey - 28		küçük bir - 18	37 : belki de - 30	9,0000	: biri - 104
100	var ki - 28	100000	tam bir - 18	38 : bir biçimde - 29		10 M 10 M 10 M 10 M 10 M 10 M 10 M 10 M
1000	anne ve - 28		yavaş yavaş - 18 de bir - 18	39 : bir süre - 29		: gün - 102
0.000	bu yüzden - 27	00.000	az sonra - 18	40 : anne ve - 29	83	: uzun - 101
	ve bu - 27	100000	her yerde - 18	41 : en çok - 29	84	: yeni - 101
	öyle bir - 27		ne diye - 18	42 : bir tek - 29	85	: şimdi - 101
A CONTRACTOR OF THE PARTY OF TH	onun için - 27	100000000000000000000000000000000000000	ama bu - 18	43 : o da - 29	235,000	: artık - 100
	bölüm 12 - 27	0.00	: bundan böyle - 17			: olur - 99
		100	Dundan boyte 17	44 : hiçbir şey - 29		ACCOUNT OF THE PARTY

1.4.3. Trigram - Top 100

```
1: kitap 3 bölüm - 65
2: kitap 1 bölüm - 55
3: kitap 2 bölüm - 54
4: başka bir şey - 18
5: 2 bölüm 12 - 17
6: ne yazık ki - 11
7: 3 bölüm 9 - 11
8: en ufak bir - 10
9: diye karşılık verdi - 10
10: ne var ki - 10
11: 1 bölüm 26 - 10
12: o kadar ki - 10
13: bir an için - 9
14: ama yine de - 9
15: anne ve babası - 9
16: baylardan orta boylusu - 9
17: 3 bölüm 13 - 9
18: 3 bölüm 12 - 9
19: bir an önce - 8
20: ben kendi hesabıma - 8
21: 3 bölüm 1 - 8
22: kısa bir süre - 7
23: bir yandan da - 7
24: pek o kadar - 7
25: 3 bölüm 2 - 7
26: 3 bölüm 5 - 7
27: böyle bir şey - 6
28: her zamanki gibi - 6
29: bana öyle geliyor - 6
30: öyle geliyor ki - 6
31: ya da bir - 6
32: daha büyük bir - 6
33: boğuk bir sesle - 6
34: sert bir sesle - 6
35: diye sesini yükseltti - 6
36:
37: ruh ve beden - 6
38: öyle sanıyorum ki - 6
39: 1 bölüm 14 - 6
40: ne kadar da - 5
41: daha iyi bir - 5
42: çok daha fazla - 5
43: dört bir yana - 5
44: bir şey mi - 5
45: kimi zaman da - 5
46: diye geçirdi içinden - 5
47: bir şey değildir - 5
48: bir o yana - 5
49: o yana bir - 5
50: yana bir bu - 5
51: bir bu yana - 5
52: 2 bölüm 2 - 5
53: 1 bölüm 2 -
   ve daha baska
```

```
55: bu yüzden de - 4
56: o kadar da - 4
57: o zamana kadar - 4
58: bir kez daha - 4
59: yine de bu - 4
60: bir şey değildi - 4
61: dahil olmak üzere - 4
62: başka herhangi bir - 4
63: işe yarar bir - 4
64: diğer bir deyişle - 4
65: de o kadar - 4
66: söz konusu olduğunda - 4
67: birkaç yıl sonra - 4
68: asık bir yüzle - 4
69: ne de olsa - 4
70: anne ve babasına - 4
71: tam bir sessizlik - 4
72: böyle bir şeyin - 4
73: anne ve babasının - 4
74: yara bere içinde - 4
75: hiçbir şey yoktur - 4
76: 2 bölüm 16 - 4
77: bu kadar büyük - 4
78: kadar büyük bir - 4
79: ve demiş ki - 4
80: ne kadar az - 4
81: o kadar az - 4
82: o güne dek - 3
83: uzun bir süre - 3
84: son bir kez - 3
85: giderek artan bir - 3
86: su dolu bir - 3
87: her iki elini - 3
88: iki elini de - 3
89: çekip çıkardı ve - 3
90: başını iki yana -
91: iki yana salladı - 3
92: o kadar çok - 3
93: öyle mi diye - 3
94: bir süre sonra - 3
95: satın almak için - 3
96: çok iyi bir - 3
97: da o kadar - 3
98: de daha fazla - 3
99: bir göz attı - 3
100: da dahil olmak -
```

```
45 : dört bir yana – 5
1 : kitap 3 bölüm
2 : kitap 1 bölüm -
                                   46 : ne kadar da - 5
3 : kitap 2 bölüm – 55
                                   47 : kimi zaman da - 5
4 : başka bir şey - 18
                                   48 : bir şey değildir - 5
5 : diye karşılık verdi - 18
                                   49 : bir o yana - 5
6 : 2 bölüm 12 - 17
                                   50 : o yana bir - 5
7 : ne yazık ki - 11
                                   51 : yana bir bu - 5
8 : ne var ki - 11
                                   52 : bir bu yana - 5
9:3 bölüm 9 - 11
                                   53 : 2 bölüm 2 - 5
10 : en ufak bir - 10
                                   54 : 1 bölüm 2 - 5
11 : 1 bölüm 26 - 10
                                   55 : ve daha başka - 5
12 : o kadar ki -
                                   56 : dahil olmak üzere - 4
                   10
13 : bir an için - 9
                                   57 : bir sey değildi - 4
14 : ama yine de -
                                   58 : başka herhangi bir - 4
15 : anne ve babası -
                                   59 : işe yarar bir - 4
16 : baylardan orta boylusu
                                   60 : diğer bir deyişle - 4
17 : 3 bölüm 13 - 9
                                   61 : de o kadar - 4
18 : 3 bölüm 12 - 9
                                   62 : söz konusu olduğunda - 4
19 : bir an önce - 8
                                   63 : o kadar da - 4
20 : ben kendi hesabıma - 8
                                   64 : birkaç yıl sonra - 4
21 : 3 bölüm 1 - 8
                                   65 : yine de bu - 4
22 : kısa bir süre - 7
                                   66 : tam bir sessizlik - 4
23 : bir yandan da - 7
                                   67 : o zamana kadar - 4
24 : pek o kadar - 7
                                   68 : böyle bir şeyin - 4
25 : sert bir sesle
                                   69 : anne ve babasının - 4
26 : 3 bölüm 2 - 7
                                   70 : yara bere içinde - 4
27 : 3 bölüm 5
                                   71 : ne de olsa - 4
28 : ya da bir - 6
                                   72 : diye söze başladı - 4
29 : daha büyük bir - 6
                                   73 : mi diye sordu - 4
30 : diye geçirdi içinden
                                   74 : diye sordu malva - 4
                                                                    89 : yaklaşık 4 yıl  -  3
31 : her zamanki gibi – 6
                                   75 : diye sözlerini sürdürdü
                                                                    90 : 4 yıl önce - 3
32 : böyle bir şey - 6
                                   76 : asık bir yüzle – 4
                                                                    91 : bir ses dalgası – 3
33 : boğuk bir sesle - 6
                                   77 : bir kahkaha attı - 4
34 : diye sesini yükseltti
                                                                    92 : ultrasonik bir ses  –  3
                                   78 : ciddi bir tavırla
35 : diye sordu yakov - 6
                                   79 : bir kez daha - 4
                                                                    93 : büyük miktarda enerji - 3
36 : ruh ve beden - 6
                                   80 : hiçbir şey yoktur - 4
                                                                    94 : bir kabın içine  -  3
37 : bana öyle geliyor - 6
                                   81 : 2 bölüm 16 - 4
                                                                    95 : için yarı geçirgen – 3
38 : öyle geliyor ki – 6
                                   82 : bu kadar büyük - 4
39 : öyle sanıyorum ki - 6
                                                                    96 : yaygın olarak kullanılan – 3
                                   83 : kadar büyük bir
40 : 1 bölüm 14 - 6
                                   84 : bu yüzden de - 4
                                                                    97 : çok yararlı bir  -  3
41 : çok daha fazla - 5
                                   85 : ve demiş ki - 4
                                                                    98 : dikkatli bir biçimde – 3
42 : daha iyi bir - 5
                                   86 : ne kadar az - 4
43 : anne ve babasına - 5
                                                                   99 : uzun bir süre  -  3
                                   87 : o kadar az - 4
44 : bir şey mi - 5
                                   88 : da dahil olmak -
                                                                    100 : bir süre önce  –
```

13. Top 100 item in python