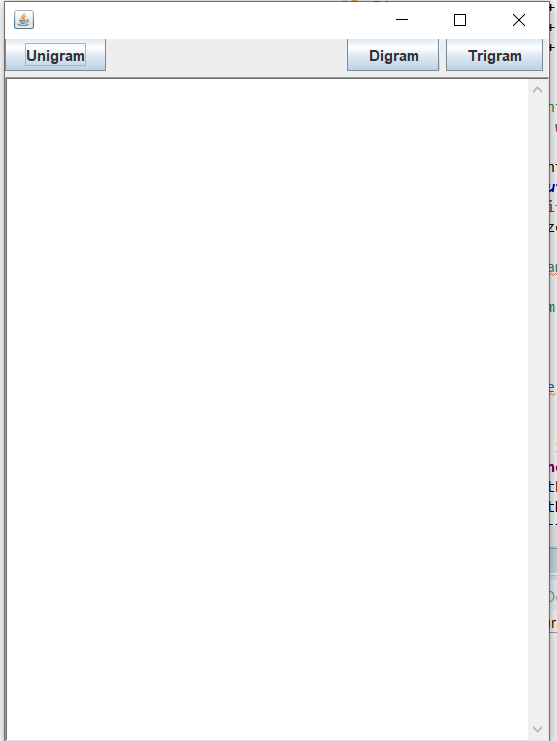
**N-GRAM PROJECT REPORT**

**Ali Kaan Alboy**

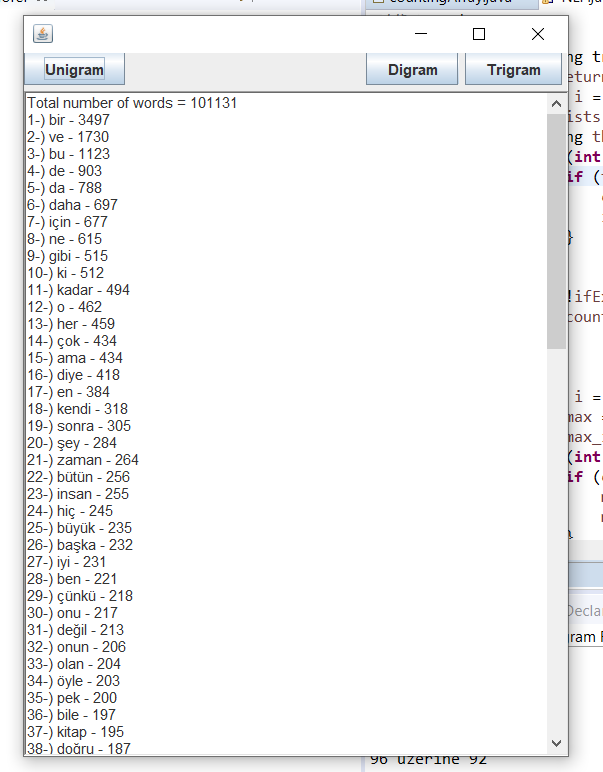
* 1. **Project Content**  
      In this project, we were asked to integrate or write a program to run the algorithms of bigram, digram and trigram from us for five book text. I wanted to develop my own algorithm instead of finding it on the internet to learn the subject better.
  2. **Project**  
      In my project, the predetermined texts are thrown into a series as soon as the program runs, and whichever algorithm you choose reduces the workload by using the same series. Then the punctuation marks are removed from the main series and the text is cleared and rendered. I also created a class called “countingArray” and used an arraylist of this type to hold the words and numbers in our main series. Each element created in this class type has a “word“ and a “count” variable.

When the program runs user can pick any n-gram type to use.



Resim 1:GUI

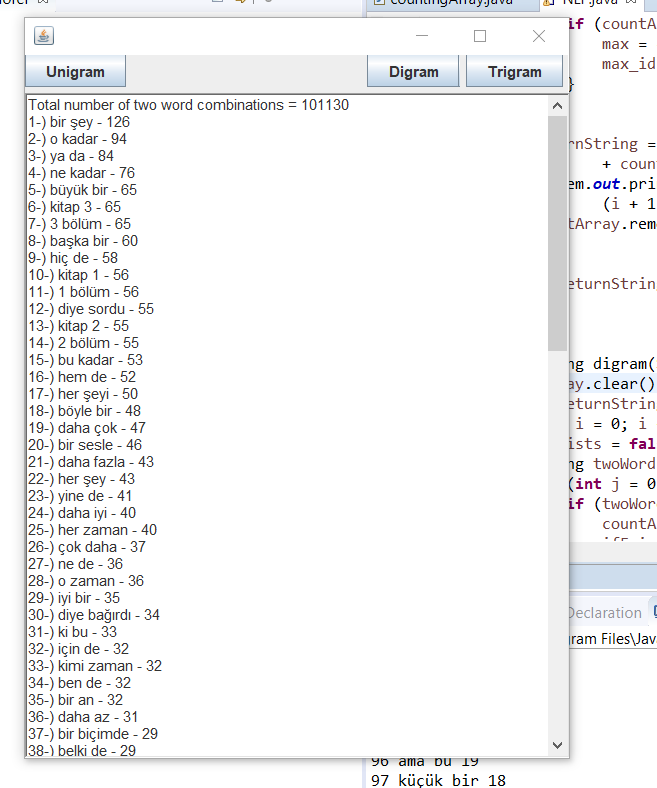
When you select unigram in the opened program, the program runs the unigram algorithm. In this algorithm, the word count from the main series is searched in the arraylist. If it is found, the number is increased by one, if not, it is added to the end of the list. After browsing all the words, the words that are used less than 70 are deleted from the list to speed up the algorithm before finding the 100 most used words. After the subtraction, the most common word is found and added to the feedback text and removed from the list and the same operation continues 100 times in total. The algorithm prints the top 100 words in the text in an average of 10.73 seconds.



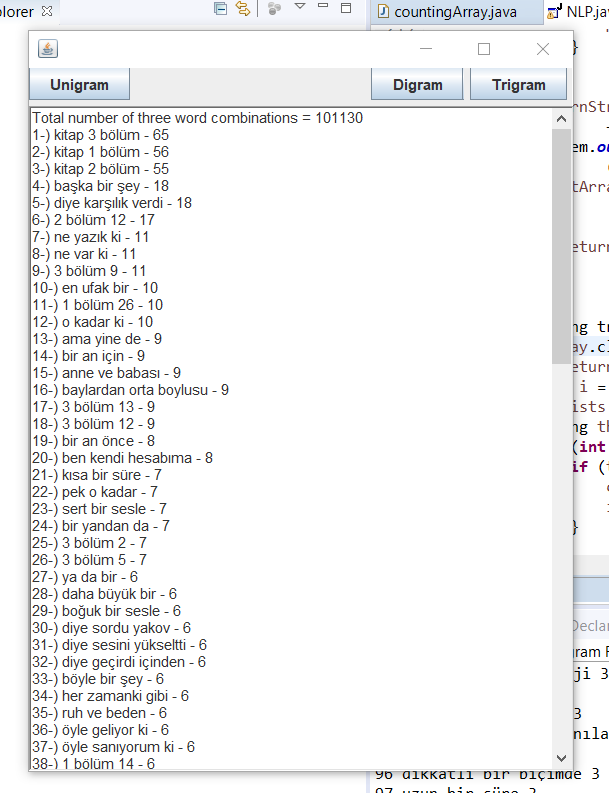
Resim 2: Unigram

When you select digram in the opened program, the program runs the digram algorithm. In this algorithm, the main series is looped into the for loop and concatenates 2 words into one string each turn. The concatenated string is being searched in the arraylist. If it is found, the number is increased by one, if not, it is added to the end of the list. Not like unigram, in digram the program does not reduce the list because the process of estimating the number of duo can cause inconsistency. After the arraylist is created, the most common duo is found and added to the feedback text and removed from the list, and the same operation continues 100 times in total. The algorithm prints up to 100 pairs of text in the text within an average of 44.65 seconds.

Resim 3: Digram

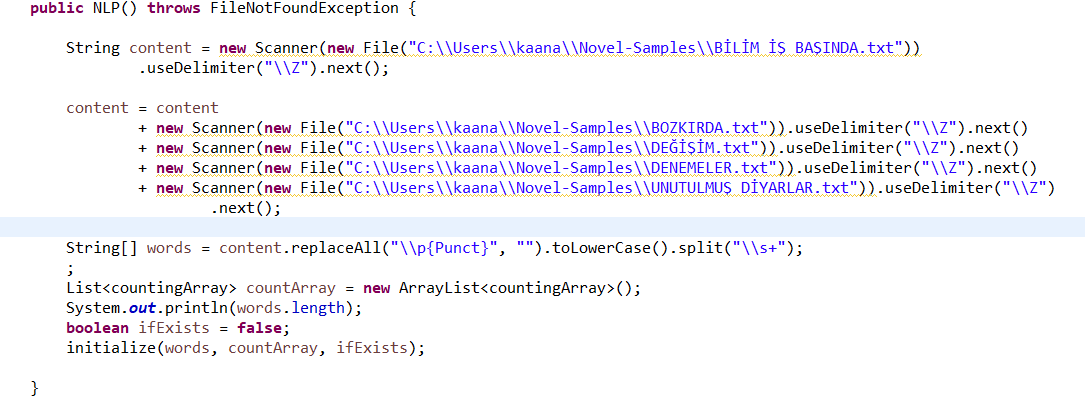


When you select trigram in the opened program, the program runs the trigram algorithm. In this algorithm, the main series is placed in the for loop and combines 3 words in a string each turn. The concatenated string is being searched in the arraylist. If it is found, the number is increased by one, if not, it is added to the end of the list. Just like digram, when we use trigram the program doesn’t subtract words from the list for the same reason. After the arraylist is created, the most occurring trio in the text is added to the feedback text and removed from the list and the same operation continues 100 times in total. The algorithm prints up top 100 trios in the text within an average of 58.33 seconds.

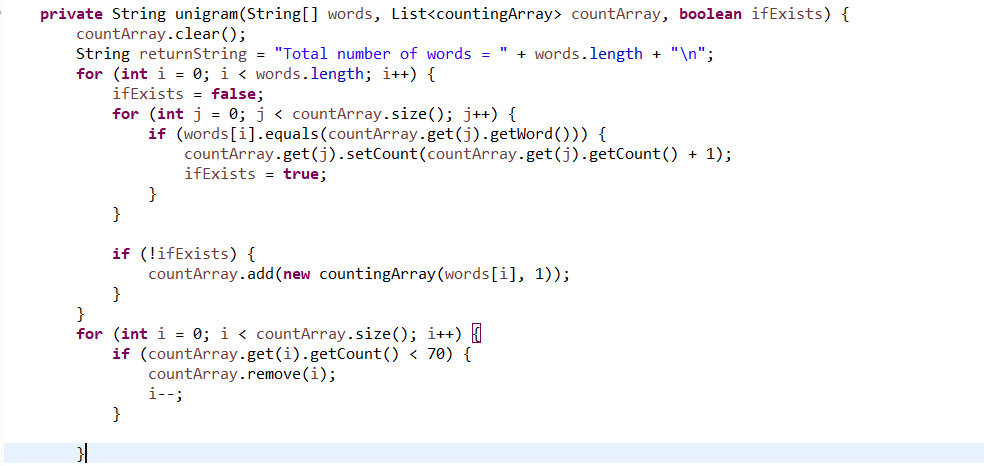


Resim 4:Trigram

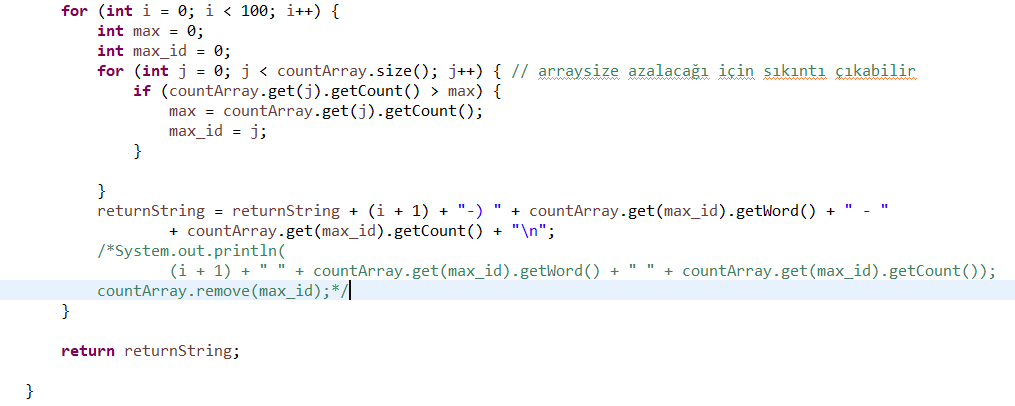
* 1. **Functions**



This function is the first function that runs when the program is opened. Here, all .txt files are read and thrown into a string. Punctuation is omitted from this string and parsed according to the spaces, and each word corresponds to an element in the words series. An arraylist of type countingArray and ifExists boolean are created for the control. The series we created and the boolean parameter are sent to the initialize function. Initialize is where the JFrame portion of the program is created. Here, user interface operations are performed for what happens when you press the button.



A similar function was used for all n-gram species. The only difference is that in the first if, it is compared to a string in which the two elements in the digram and the three elements in the trigram are combined, instead of comparing with the element of the main series words. In all functions, words (array with the main text), countArray (arraylist with each word and number used) and ifExists (a boolean that checks if the word has already been added to the arraylist) are taken as parameters. After navigating the lists, the unigram words that are less than 70 times are deleted from the list.



In order to find the most 100 words, the list is searched and the most used word is found, added to the feedback string, removed from the list and the same operation is done 100 times to create the list.

* 1. **Result**  
      As a result, the program reads 5 texts of a total of 101131 words, optionally passes unigram, digram and trigram algorithms and prints the most used 100 data on the screen in under 1 minute. Below you can find the top 100 lists that the program finds.
  2. **Unigram – 100**

1-) bir - 3497

2-) ve - 1730

3-) bu - 1123

4-) de - 903

5-) da - 788

6-) daha - 697

7-) için - 677

8-) ne - 615

9-) gibi - 515

10-) ki - 512

11-) kadar - 494

12-) o - 462

13-) her - 459

14-) çok - 434

15-) ama - 434

16-) diye - 418

17-) en - 384

18-) kendi - 318

19-) sonra - 305

20-) şey - 284

21-) zaman - 264

22-) bütün - 256

23-) insan - 255

24-) hiç - 245

25-) büyük - 235

26-) başka - 232

27-) iyi - 231

28-) ben - 221

29-) çünkü - 218

30-) onu - 217

31-) değil - 213

32-) onun - 206

33-) olan - 204

34-) öyle - 203

35-) pek - 200

36-) bile - 197

37-) kitap - 195

38-) doğru - 187

39-) bölüm - 177

40-) böyle - 168

41-) hiçbir - 166

42-) kendini - 164

43-) dedi - 163

44-) ona - 161

45-) mi - 157

46-) içinde - 155

47-) olarak - 152

48-) iki - 151

49-) ya - 147

50-) yakov - 145

51-) az - 144

52-) bana - 143

53-) gregor - 143

54-) var - 141

55-) nasıl - 140

56-) ancak - 134

57-) olduğunu - 129

58-) yalnız - 129

59-) vasili - 126

60-) bizim - 124

61-) aynı - 123

62-) malva - 121

63-) olduğu - 120

64-) mı - 120

65-) üzerinde - 118

66-) fazla - 118

67-) hem - 117

68-) yine - 116

69-) göre - 115

70-) benim - 115

71-) önce - 114

72-) bunu - 114

73-) karşı - 112

74-) ile - 110

75-) fakat - 110

76-) hep - 110

77-) arasında - 108

78-) yere - 108

79-) biri - 104

80-) biraz - 103

81-) ilk - 103

82-) gün - 102

83-) yeni - 101

84-) uzun - 100

85-) olur - 99

86-) türlü - 98

87-) artık - 97

88-) kez - 96

89-) şimdi - 96

90-) adam - 96

91-) bizi - 96

92-) yok - 95

93-) gregorun - 95

94-) üstüne - 93

95-) kızkardeşi - 93

96-) üzerine - 92

97-) güzel - 92

98-) hemen - 91

99-) sen - 89

100-) insanın – 88

* 1. **Digram­ – 100**

1-) bir şey - 126

2-) o kadar - 94

3-) ya da - 84

4-) ne kadar - 76

5-) büyük bir - 65

6-) kitap 3 - 65

7-) 3 bölüm - 65

8-) başka bir - 60

9-) hiç de - 58

10-) kitap 1 - 56

11-) 1 bölüm - 56

12-) diye sordu - 55

13-) kitap 2 - 55

14-) 2 bölüm - 55

15-) bu kadar - 53

16-) hem de - 52

17-) her şeyi - 50

18-) böyle bir - 48

19-) daha çok - 47

20-) bir sesle - 46

21-) daha fazla - 43

22-) her şey - 43

23-) yine de - 41

24-) daha iyi - 40

25-) her zaman - 40

26-) çok daha - 37

27-) ne de - 36

28-) o zaman - 36

29-) iyi bir - 35

30-) diye bağırdı - 34

31-) ki bu - 33

32-) için de - 32

33-) kimi zaman - 32

34-) ben de - 32

35-) bir an - 32

36-) daha az - 31

37-) bir biçimde - 29

38-) belki de - 29

39-) bir süre - 29

40-) o da - 29

41-) en çok - 29

42-) anne ve - 29

43-) için bir - 28

44-) bu yüzden - 28

45-) daha da - 28

46-) hiçbir şey - 28

47-) var ki - 28

48-) bir tek - 28

49-) ve bu - 27

50-) öyle bir - 27

51-) onun için - 27

52-) bölüm 12 - 27

53-) yeni bir - 26

54-) sonra da - 26

55-) bir gün - 26

56-) her gün - 26

57-) gibi bir - 25

58-) bir başka - 25

59-) en iyi - 25

60-) olduğu için - 25

61-) bir tavırla - 24

62-) her şeyin - 24

63-) der ki - 24

64-) ve bir - 23

65-) karşılık verdi - 23

66-) bir insan - 23

67-) güzel bir - 23

68-) ne var - 23

69-) bir türlü - 23

70-) en büyük - 23

71-) pek çok - 22

72-) bir şeyler - 22

73-) bir ses - 22

74-) herhangi bir - 22

75-) bir sürü - 22

76-) müdür bey - 22

77-) ve daha - 22

78-) diye karşılık - 21

79-) bir iş - 21

80-) hafif bir - 20

81-) en az - 20

82-) çok kez - 20

83-) bir kez - 20

84-) kendi kendine - 20

85-) bu işi - 19

86-) bu konuda - 19

87-) biraz daha - 19

88-) sağa sola - 19

89-) da bir - 19

90-) bir yandan - 19

91-) ağır ağır - 19

92-) bundan böyle - 19

93-) az sonra - 19

94-) ne diye - 19

95-) başka türlü - 19

96-) ama bu - 19

97-) küçük bir - 18

98-) yavaş yavaş - 18

99-) de bir - 18

100-) tam bir - 18

* 1. **Trigram – 100**

1-) kitap 3 bölüm - 65

2-) kitap 1 bölüm - 56

3-) kitap 2 bölüm - 55

4-) başka bir şey - 18

5-) diye karşılık verdi - 18

6-) 2 bölüm 12 - 17

7-) ne yazık ki - 11

8-) ne var ki - 11

9-) 3 bölüm 9 - 11

10-) en ufak bir - 10

11-) 1 bölüm 26 - 10

12-) o kadar ki - 10

13-) ama yine de - 9

14-) bir an için - 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-) kısa bir süre - 7

22-) pek o kadar - 7

23-) sert bir sesle - 7

24-) bir yandan da - 7

25-) 3 bölüm 2 - 7

26-) 3 bölüm 5 - 7

27-) ya da bir - 6

28-) daha büyük bir - 6

29-) boğuk bir sesle - 6

30-) diye sordu yakov - 6

31-) diye sesini yükseltti - 6

32-) diye geçirdi içinden - 6

33-) böyle bir şey - 6

34-) her zamanki gibi - 6

35-) ruh ve beden - 6

36-) öyle geliyor ki - 6

37-) öyle sanıyorum ki - 6

38-) 1 bölüm 14 - 6

39-) çok daha fazla - 5

40-) daha iyi bir - 5

41-) dört bir yana - 5

42-) ne kadar da - 5

43-) kimi zaman da - 5

44-) anne ve babasına - 5

45-) bir şey değildir - 5

46-) bir o yana - 5

47-) o yana bir - 5

48-) yana bir bu - 5

49-) bir bu yana - 5

50-) bana öyle geliyor - 5

51-) ve daha başka - 5

52-) dahil olmak üzere - 4

53-) bir şey değildi - 4

54-) başka herhangi bir - 4

55-) işe yarar bir - 4

56-) diğer bir deyişle - 4

57-) de o kadar - 4

58-) söz konusu olduğunda - 4

59-) o kadar da - 4

60-) birkaç yıl sonra - 4

61-) yine de bu - 4

62-) diye söze başladı - 4

63-) o zamana kadar - 4

64-) mi diye sordu - 4

65-) bir şey mi - 4

66-) diye sordu malva - 4

67-) diye sözlerini sürdürdü - 4

68-) asık bir yüzle - 4

69-) bir kahkaha attı - 4

70-) ne de olsa - 4

71-) ciddi bir tavırla - 4

72-) bir kez daha - 4

73-) tam bir sessizlik - 4

74-) böyle bir şeyin - 4

75-) anne ve babasının - 4

76-) yara bere içinde - 4

77-) hiçbir şey yoktur - 4

78-) 2 bölüm 16 - 4

79-) bu kadar büyük - 4

80-) kadar büyük bir - 4

81-) 1 bölüm 20 - 4

82-) bu yüzden de - 4

83-) ve demiş ki - 4

84-) ne kadar az - 4

85-) o kadar az - 4

86-) 3 bölüm 10 - 4

87-) 3 bölüm 1 - 4

88-) da dahil olmak - 3

89-) bir ses dalgası - 3

90-) ultrasonik bir ses - 3

91-) büyük miktarda enerji - 3

92-) bir kabın içine - 3

93-) için yarı geçirgen - 3

94-) yaygın olarak kullanılan - 3

95-) çok yararlı bir - 3

96-) dikkatli bir biçimde - 3

97-) uzun bir süre - 3

98-) bir süre önce - 3

99-) bir elektrik yükü - 3

100-) termodinamiğin ikinci yasası - 3