N-GRAM IMPLEMENTATION REPORT

Ali Şiyar Arslan
2019510017

1.1 Problem Definition

The goal of this project is to develop n-gram algorithm using the Java programming language. The project involves extracting valuable insights from a segment of the Turkish Novel Corpus, which comprises four distinct novels, along with an English text. The objective is to compare the outcomes obtained from these tests. N-grams, a fundamental concept in text mining and natural language processing, represent sequences of words that occur consecutively within a given text.

When computing n-grams, the usual practice is to advance by n words in the text. n-gram is a series of n adjacent letters (including punctuation marks and blanks), syllables, or rarely whole words found in a language dataset; or adjacent phonemes extracted from a speech-recording dataset, or adjacent base pairs extracted from a genome. They are collected from a text or speech corpus. If Latin numerical prefixes are used, then n-gram of size 1 is called a "unigram", size 2 a "bigram" (or, less commonly, a "digram") etc.

1.2. Project Scope

In this project, the goal was to seamlessly extract 1-gram, 2-gram, and 3-gram occurrences from a set of provided novels. The implementation was carried out in Java, focusing on achieving a smooth and efficient process. The outcome includes the identification and counting of the top 20 items for each n-gram, along with the recorded running time for each specific n-gram analysis.

1.3. Implementation

Instead of getting the algorithm ready-made from the internet and using it, I wrote the algorithm myself.

I created a function called readTxt to read from txt. This function is a Java method that reads a specified text file and cleans it by making various edits to the text. Reads the text file using FileInputStream and Scanner. Appends each line to a String variable as long as there is a next line in the file. A line skip character (\n) is added to the end of each line. After the file reading process is completed, the Scanner object is closed. Punctuation marks are identified and removed from the text. Spaces, numbers, line break characters, extra spaces, and other unnecessary elements are removed. All characters are converted to lowercase. These operations clean the text, making it ready for later n-gram analysis or other text-based operations.

I wrote a function called calculate_ngram_counts_probabilities to find the frequency of words and then calculate the ngram probability. This function is a Java method that counts words and word groups in a text and then calculates n-gram probabilities based on these numbers. An array is created by splitting the text based on the space character. Empty elements in the created array are cleared. Counts the number of each word in the text. It also calculates the numbers of two consecutive words (bigrams) and three words (trigrams). Unigram, bigram and trigram probabilities are calculated using N-gram numbers. This includes the probability of each word, the probability of two words occurring consecutively, and the probability of three words occurring consecutively. These operations perform basic calculations of an n-gram model used to understand grammatical structures and frequencies in text.

I created a function called ngramProcess, which contains each ngram process and contains the print operations in a collective manner. This function calculates unigram, bigram and trigram probabilities by processing a specified text file. A time start is recorded to measure the running time of the function. Reads the specified text file via the readTxt function. It calculates unigram, bigram and trigram probabilities and numbers with the calculate_ngram_counts_probabilities function. It prints the top 20 unigram, bigram and trigram numbers and their probabilities on the screen. Calculates the time elapsed after the function completes its execution time. Cleans up data structures containing calculated unigram, bigram, and trigram numbers and probabilities. These operations are used to understand the language structure in a text file and evaluate n-gram probabilities.

I created a function called sortByValues. This function is a helper method used to sort a Map structure by its values. The function takes a public key (K) and value (V) type. When comparing values, the values are expected to be a type that fits the Comparable interface. A Comparator object is defined, this object will be used to compare values. In Comparator, the values of two given keys (k1 and k2) are compared and a result is obtained for sorting. Using TreeMap, a copy of the original Map structure is created, sorted by values. This sorted copy Map object is returned.

I created a function called printTopItems. This function is a Java method designed to print a certain number of top elements of a specified Map structure to the screen. Parameters ; newMap: Structure of the Map to be printed, topCount: Number of elements to be printed at the top. A loop is started for each element in newMap. For each item, the item's key (entry.getKey()) and value (entry.getValue()) are printed on the screen in a specific format. If the element's value is a Double type, it is printed using a special format (.5g). The loop is terminated when the specified topCount number is reached.

1.4 Result

1.4.1 BİLİM İŞ BAŞINDA.txt

1.4.1.1 BİLİM İŞ BAŞINDA.txt Unigram Count and Probabilities

```
*******************************************************
top 20 unigram count

1- bir ==> 454
2- ve ==> 203
3- bu ==> 154
4- için ==> 90
5- de ==> 77
6- çok ==> 73
7- ile ==> 71
8- daha ==> 66
9- da ==> 60
10- olarak ==> 52
11- ancak ==> 50
12- elektrik ==> 47
13- su ==> 43
14- veya ==> 43
15- kadar ==> 41
16- büyük ==> 37
17- biçimde ==> 30
18- gibi ==> 30
19- enerji ==> 29
20- iki ==> 29
```

```
*******************************************************
top 20 unigram probabilities

1- bir ==> 0,048928
2- ve ==> 0,021877
3- bu ==> 0,016597
4- için ==> 0,0096993
5- de ==> 0,0082983
6- çok ==> 0,0078672
7- ile ==> 0,0076517
8- daha ==> 0,0071128
9- da ==> 0,0064662
10- olarak ==> 0,0056041
11- ancak ==> 0,0053885
12- elektrik ==> 0,0050652
13- su ==> 0,0046341
14- veya ==> 0,0046341
15- kadar ==> 0,0044186
16- büyük ==> 0,0039875
17- biçimde ==> 0,0032331
18- gibi ==> 0,0032331
19- enerji ==> 0,0031253
20- iki ==> 0,0031253
```

1.4.1.2 BİLİM İŞ BAŞINDA.txt Bigram Count and Probabilities

```
*******************************************************
top 20 bigram count

1- bir biçimde ==> 19
2- pek çok ==> 19
3- ile ilgili ==> 11
4- deniz suyu ==> 10
5- bir elektrik ==> 9
6- bu nedenle ==> 9
7- herhangi bir ==> 9
8- iyi bir ==> 9
9- uygun bir ==> 9
10- yine de ==> 9
11- bir ses ==> 8
12- elektrik yükü ==> 8
13- yıl önce ==> 8
14- büyük bir ==> 7
15- küçük bir ==> 7
16- söz konusu ==> 7
17- ancak bu ==> 6
18- bir cam ==> 6
19- bir kısmı ==> 6
20- bir şey ==> 6
```

```
*******************************************************
top 20 bigram probabilities

1- abartma eğiliminden ==> 1,0000
2- abartılmasını veya ==> 1,0000
3- acemice yapılmış ==> 1,0000
4- adla anılıyor ==> 1,0000
5- adlandırma olduğunu ==> 1,0000
6- adlandırıldığı günlerde ==> 1,0000
7- adlandırılmaya başlandı ==> 1,0000
8- adlandırılır sıcaklık ==> 1,0000
9- adlandırıyordu bizimse ==> 1,0000
10- ahşap yapı ==> 1,0000
11- akabilir fakat ==> 1,0000
12- akacağını ancak ==> 1,0000
13- akademik takvimlerine ==> 1,0000
14- akamaz yürüyen ==> 1,0000
15- akan deniz ==> 1,0000
16- akarak uzaklaştıklarından ==> 1,0000
17- akmasına izin ==> 1,0000
18- akmaya karşı ==> 1,0000
19- akmayacağını söylüyor ==> 1,0000
20- akropolisin girişinin ==> 1,0000
```

1.4.1.3 BİLİM İŞ BAŞINDA.txt Trigram Count and Probabilities

```
**************************************************
top 20 trigram count

1- yaklaşık yıl önce ==> 6
2- ne yazık ki ==> 5
3- başka herhangi bir ==> 4
4- dahil olmak üzere ==> 4
5- diğer bir deyişle ==> 4
6- söz konusu olduğunda ==> 4
7- bir elektrik yükü ==> 3
8- bir kabın içine ==> 3
9- bir ses dalgası ==> 3
10- buzun erime noktası ==> 3
11- büyük miktarda enerji ==> 3
12- da dahil olmak ==> 3
13- için yarı geçirgen ==> 3
14- işe yarar bir ==> 3
15- termodinamiğin ikinci yasası ==> 3
16- ultrasonik bir ses ==> 3
17- yaygın olarak kullanılan ==> 3
18- alçak basınç altında ==> 2
19- ani etkili damıtma ==> 2
20- arkasından burnunuzun ucuna ==> 2
```

```
**************************************************
top 20 trigram probabilities

1- a a grilfiths ==> 1,0000
2- a grilfiths yaklaşık ==> 1,0000
3- a m ramseyin ==> 1,0000
4- a rowland bir ==> 1,0000
5- abartma eğiliminden payına ==> 1,0000
6- abartılmasını veya önemsiz ==> 1,0000
7- acemice yapılmış bir ==> 1,0000
8- adamları sorunun ne ==> 1,0000
9- adamları tarafından tasarlanmış ==> 1,0000
10- adamlarının açıkladığı olgulara ==> 1,0000
11- adamlarının kitle iletişim ==> 1,0000
12- adamlarının yeni bir ==> 1,0000
13- adamı belirli maddelerin ==> 1,0000
14- adamı genellikle şöyle ==> 1,0000
15- adamı kitle iletişim ==> 1,0000
16- adamı olan charles ==> 1,0000
17- adla anılıyor işleminde ==> 1,0000
18- adlandırma olduğunu gösterdi ==> 1,0000
19- adlandırıldığı günlerde profesörlerinin ==> 1,0000
20- adlandırılmaya başlandı ancak ==> 1,0000

The ngram probability calculation of the BİLİM İŞ BAŞINDA.txt took 381 millisecond.

**************************************************
**************************************************
```

1.4.2 BOZKIRDA.txt

1.4.2.1 BOZKIRDA.txt Unigram Count and Probabilities

```
**************************************************
top 20 unigram count

1- bir ==> 623
2- diye ==> 208
3- ne ==> 195
4- de ==> 157
5- bu ==> 152
6- yakov ==> 145
7- dedi ==> 136
8- da ==> 129
9- vasili ==> 126
10- malva ==> 121
11- ve ==> 120
12- gibi ==> 96
13- ben ==> 94
14- sonra ==> 90
15- sen ==> 72
16- o ==> 71
17- daha ==> 70
18- seryojka ==> 70
19- için ==> 63
20- mi ==> 59
```

```
**************************************************
top 20 unigram probabilities

1- bir ==> 0,035955
2- diye ==> 0,012004
3- ne ==> 0,011254
4- de ==> 0,0090610
5- bu ==> 0,0087724
6- yakov ==> 0,0083684
7- dedi ==> 0,0078490
8- da ==> 0,0074450
9- vasili ==> 0,0072719
10- malva ==> 0,0069833
11- ve ==> 0,0069256
12- gibi ==> 0,0055405
13- ben ==> 0,0054251
14- sonra ==> 0,0051942
15- sen ==> 0,0041554
16- o ==> 0,0040977
17- daha ==> 0,0040399
18- seryojka ==> 0,0040399
19- için ==> 0,0036359
20- mi ==> 0,0034051
```

1.4.2.2 BOZKIRDA.txt Bigram Count and Probabilities

```
*******************************************************
top 20 bigram count

1- diye sordu ==> 45
2- bir sesle ==> 44
3- bir şey ==> 33
4- diye bağırdı ==> 28
5- bir tavırla ==> 22
6- karşılık verdi ==> 21
7- diye karşılık ==> 19
8- ağır ağır ==> 15
9- ben de ==> 15
10- hem de ==> 14
11- bir şeyler ==> 13
12- bir süre ==> 12
13- sert bir ==> 12
14- ne var ==> 11
15- o zaman ==> 11
16- ya da ==> 11
17- anladın mı ==> 10
18- belki de ==> 10
19- dedi yakov ==> 10
20- kimi zaman ==> 10
```

```
*******************************************************
top 20 bigram probabilities

1- abanarak doğruldu ==> 1,0000
2- abandı kucak ==> 1,0000
3- abanıyor ciğerlerim ==> 1,0000
4- acelen ne ==> 1,0000
5- acıktı ki ==> 1,0000
6- acıma duygusu ==> 1,0000
7- acımamıştı ama ==> 1,0000
8- acımaya başlamıştı ==> 1,0000
9- acısan neyse ==> 1,0000
10- acısıyla kederli ==> 1,0000
11- acıyacak ne ==> 1,0000
12- acıyan başını ==> 1,0000
13- acıyordu malvaya ==> 1,0000
14- acıyorum ona ==> 1,0000
15- acıyı tattıkça ==> 1,0000
16- adamdır be ==> 1,0000
17- adamlardan biri ==> 1,0000
18- adem baba ==> 1,0000
19- adı da ==> 1,0000
20- adımlık bir ==> 1,0000
```

1.4.2.3 BOZKIRDA.txt Trigram Count and Probabilities

```
*******************************************************
top 20 trigram count

1- diye karşılık verdi ==> 18
2- sert bir sesle ==> 7
3- diye sordu yakov ==> 6
4- başka bir şey ==> 5
5- boğuk bir sesle ==> 5
6- diye sesini yükseltti ==> 5
7- asık bir yüzle ==> 4
8- bir kahkaha attı ==> 4
9- ciddi bir tavırla ==> 4
10- diye sordu malva ==> 4
11- diye sözlerini sürdürdü ==> 4
12- mi diye sordu ==> 4
13- ne kadar da ==> 4
14- ama yine de ==> 3
15- belli belirsiz bir ==> 3
16- bir bardak votka ==> 3
17- bir kedi gibi ==> 3
18- bir sesle ne ==> 3
19- bir yandan da ==> 3
20- bir şey mi ==> 3
```

```
*******************************************************
top 20 trigram probabilities

1- abanarak doğruldu ulan ==> 1,0000
2- abandı kucak kucağa ==> 1,0000
3- abanıyor ciğerlerim taze ==> 1,0000
4- acaba ben de ==> 1,0000
5- acaba dövmüş olsaydı ==> 1,0000
6- acaba ha melun ==> 1,0000
7- acaba hay allah ==> 1,0000
8- acaba malva kıçta ==> 1,0000
9- acaba neyin nesi ==> 1,0000
10- acaba sen bana ==> 1,0000
11- acaba seryojka mı ==> 1,0000
12- acaba yakov ne ==> 1,0000
13- acaba yine seryojka ==> 1,0000
14- acele bir işleri ==> 1,0000
15- acele etmeden dudaklarını ==> 1,0000
16- acele ettin onlar ==> 1,0000
17- acelen ne diye ==> 1,0000
18- acemi acemi yüzüp ==> 1,0000
19- acemi bir kürekçiydi ==> 1,0000
20- acemi yüzüp gelen ==> 1,0000

The ngram probability calculation of the BOZKIRDA.txt took 567 millisecond.

*******************************************************
*******************************************************
```

1.4.3 DEĞİŞİM.txt

1.4.3.1 DEĞİŞİM.txt Unigram Count and Probabilities

```
**********************************************************
top 20 unigram count

1- bir ==> 537
2- ve ==> 256
3- bu ==> 160
4- ama ==> 145
5- gregor ==> 143
6- de ==> 123
7- da ==> 102
8- gibi ==> 95
9- gregorun ==> 95
10- için ==> 93
11- kızkardeşi ==> 93
12- daha ==> 80
13- diye ==> 77
14- ne ==> 70
15- babası ==> 66
16- çünkü ==> 61
17- üzerine ==> 57
18- ancak ==> 56
19- sonra ==> 55
20- her ==> 54
```

```
**********************************************************
top 20 unigram probabilities

1- bir ==> 0,034313
2- ve ==> 0,016358
3- bu ==> 0,010224
4- ama ==> 0,0092652
5- gregor ==> 0,0091374
6- de ==> 0,0078594
7- da ==> 0,0065176
8- gibi ==> 0,0060703
9- gregorun ==> 0,0060703
10- için ==> 0,0059425
11- kızkardeşi ==> 0,0059425
12- daha ==> 0,0051118
13- diye ==> 0,0049201
14- ne ==> 0,0044728
15- babası ==> 0,0042173
16- çünkü ==> 0,0038978
17- üzerine ==> 0,0036422
18- ancak ==> 0,0035783
19- sonra ==> 0,0035144
20- her ==> 0,0034505
```

1.4.3.2 DEĞİŞİM.txt Bigram Count and Probabilities

```
**********************************************************
top 20 bigram count

1- anne ve ==> 29
2- bir şey ==> 28
3- müdür bey ==> 22
4- müdür beyin ==> 18
5- bundan böyle ==> 17
6- bay samsa ==> 15
7- bir an ==> 15
8- hiç de ==> 15
9- ya da ==> 15
10- böyle bir ==> 14
11- büyük bir ==> 14
12- sağa sola ==> 14
13- kendi kendine ==> 13
14- bir türlü ==> 12
15- bunun üzerine ==> 12
16- yavaş yavaş ==> 11
17- baylardan orta ==> 10
18- gibi bir ==> 10
19- o anda ==> 10
20- beri yandan ==> 9
```

```
**********************************************************
top 20 bigram probabilities

1- acıkmış hissediyordu ==> 1,0000
2- acımaklı bir ==> 1,0000
3- adamakıllı bir ==> 1,0000
4- adamdı beş ==> 1,0000
5- adamları maalesef ==> 1,0000
6- adamın kızkardeşine ==> 1,0000
7- adı geçen ==> 1,0000
8- adım atarak ==> 1,0000
9- adıma karşılık ==> 1,0000
10- adımlar atarak ==> 1,0000
11- adımlarla yürüyüp ==> 1,0000
12- adımlarını atıyordu ==> 1,0000
13- adına konuşuyor ==> 1,0000
14- adını ağzıma ==> 1,0000
15- afallamış ve ==> 1,0000
16- ahlaya poflaya ==> 1,0000
17- ahlayıp oflayarak ==> 1,0000
18- ahşap olduğu ==> 1,0000
19- ailede kimin ==> 1,0000
20- aileden izin ==> 1,0000
```

1.4.3.3 DEĞİŞİM.txt Trigram Count and Probabilities

```
**************************************************
top 20 trigram count

1- anne ve babası ==> 9
2- baylardan orta boylusu ==> 9
3- ne var ki ==> 8
4- en ufak bir ==> 6
5- anne ve babasına ==> 5
6- diye geçirdi içinden ==> 5
7- ama yine de ==> 4
8- anne ve babasının ==> 4
9- böyle bir şeyin ==> 4
10- anne ve babasıyla ==> 3
11- bir an bile ==> 3
12- bir an önce ==> 3
13- bir çeyrek saat ==> 3
14- diye söylendi kendi ==> 3
15- ne de olsa ==> 3
16- ne yazık ki ==> 3
17- parmak uçlarına basarak ==> 3
18- söylendi kendi kendine ==> 3
19- tam bir sessizlik ==> 3
20- zahmetli bir iş ==> 3
```

```
**************************************************
top 20 trigram probabilities

1- acaba bugünkü kuruntusu ==> 1,0000
2- acaba kendisi yataktan ==> 1,0000
3- acaba kızkardeşi süte ==> 1,0000
4- acayip kaprislere kaptırmış ==> 1,0000
5- acayip oturuş öyle ==> 1,0000
6- acele bir kovanın ==> 1,0000
7- acele davranan kızkardeşiydi ==> 1,0000
8- acele etmeksizin gerekli ==> 1,0000
9- acele ve suskun ==> 1,0000
10- aceleci bir kadın ==> 1,0000
11- aceleci ellerle pencereyi ==> 1,0000
12- aceleci hizmetçi tarafından ==> 1,0000
13- acı çekiyorlardı o ==> 1,0000
14- acı çektirdiğini düşündü ==> 1,0000
15- acıkmış hissediyordu bütün ==> 1,0000
16- acımaklı bir ses ==> 1,0000
17- acınacak bir ses ==> 1,0000
18- acınacak kadar cılız ==> 1,0000
19- acıyla kafasını çevirip ==> 1,0000
20- acıyla o anda ==> 1,0000

The ngram probability calculation of the DEĞİŞİM.txt took 374 millisecond.

**************************************************
**************************************************
```

1.4.4 DENEMELER.txt

1.4.4.1 DENEMELER.txt Unigram Count and Probabilities

```
**************************************************
top 20 unigram count

1- bir ==> 1510
2- ve ==> 895
3- bu ==> 568
4- de ==> 477
5- daha ==> 440
6- da ==> 432
7- ki ==> 391
8- için ==> 379
9- her ==> 318
10- ne ==> 317
11- kadar ==> 308
12- o ==> 301
13- en ==> 285
14- kendi ==> 255
15- gibi ==> 248
16- çok ==> 231
17- ama ==> 228
18- insan ==> 205
19- kitap ==> 192
20- bütün ==> 181
```

```
**************************************************
top 20 unigram probabilities

1- bir ==> 0,029992
2- ve ==> 0,017777
3- bu ==> 0,011282
4- de ==> 0,0094742
5- daha ==> 0,0087393
6- da ==> 0,0085805
7- ki ==> 0,0077661
8- için ==> 0,0075278
9- her ==> 0,0063162
10- ne ==> 0,0062963
11- kadar ==> 0,0061175
12- o ==> 0,0059785
13- en ==> 0,0056607
14- kendi ==> 0,0050648
15- gibi ==> 0,0049258
16- çok ==> 0,0045882
17- ama ==> 0,0045286
18- insan ==> 0,0040717
19- kitap ==> 0,0038135
20- bütün ==> 0,0035951
```

1.4.4.2 DENEMELER.txt Bigram Count and Probabilities

```
*****************************************************
top 20 bigram count

1- kitap bölüm ==> 177
2- o kadar ==> 75
3- ne kadar ==> 56
4- bir şey ==> 51
5- ya da ==> 49
6- her şeyi ==> 42
7- başka bir ==> 40
8- bu kadar ==> 37
9- büyük bir ==> 37
10- hiç de ==> 34
11- daha fazla ==> 32
12- her şey ==> 32
13- daha çok ==> 31
14- hem de ==> 28
15- her zaman ==> 27
16- böyle bir ==> 26
17- ki bu ==> 26
18- onun için ==> 25
19- çok daha ==> 25
20- der ki ==> 24
```

```
*****************************************************
top 20 bigram probabilities

1- abc öğreniyor ==> 1,0000
2- abderiadan gelen ==> 1,0000
3- abenea turis ==> 1,0000
4- abest meus ==> 1,0000
5- abique habitat ==> 1,0000
6- ablatum medüs ==> 1,0000
7- absit mente ==> 1,0000
8- absolvitur juvenalis ==> 1,0000
9- acayipliğini görüp ==> 1,0000
10- accipimus enim ==> 1,0000
11- acervus et ==> 1,0000
12- acquiritur plus ==> 1,0000
13- acre revenum ==> 1,0000
14- acres sollicitum ==> 1,0000
15- acri fingendus ==> 1,0000
16- acta vetutas ==> 1,0000
17- acuant mentem ==> 1,0000
18- acıdan ve ==> 1,0000
19- acıdığı vahlandığı ==> 1,0000
20- acıdığımız şeye ==> 1,0000
```

1.4.4.3 DENEMELER.txt Trigram Count and Probabilities

```
*****************************************************
top 20 trigram count

1- başka bir şey ==> 10
2- o kadar ki ==> 10
3- ben kendi hesabıma ==> 8
4- kitap bölüm insan ==> 7
5- pek o kadar ==> 6
6- ruh ve beden ==> 6
7- öyle sanıyorum ki ==> 6
8- bana öyle geliyor ==> 5
9- bir bu yana ==> 5
10- bir o yana ==> 5
11- bir şey değildir ==> 5
12- kitap bölüm insanın ==> 5
13- o yana bir ==> 5
14- ve daha başka ==> 5
15- yana bir bu ==> 5
16- öyle geliyor ki ==> 5
17- bu kadar büyük ==> 4
18- daha büyük bir ==> 4
19- değildir kitap bölüm ==> 4
20- hiçbir şey yoktur ==> 4
```

```
*****************************************************
top 20 trigram probabilities

1- a collo trahitur ==> 1,0000
2- a far opre ==> 1,0000
3- a karşı açtıkları ==> 1,0000
4- a limine moverat ==> 1,0000
5- a malis ennfus ==> 1,0000
6- a onları kendilerine ==> 1,0000
7- a quo ceu ==> 1,0000
8- ab aemonio num ==> 1,0000
9- ab curo lucretius ==> 1,0000
10- ab imo ejiciuntur ==> 1,0000
11- ab ipso ducis ==> 1,0000
12- ab numine distant ==> 1,0000
13- ab origine pendet ==> 1,0000
14- abanır oralarda rahat ==> 1,0000
15- abanır yüreksiz ederler ==> 1,0000
16- abc öğreniyor henüz ==> 1,0000
17- abderiadan gelen bir ==> 1,0000
18- abenea turis non ==> 1,0000
19- abest meus particeps ==> 1,0000
20- abique habitat maxime ==> 1,0000

The ngram probability calculation of the DENEMELER.txt took 2665 millisecond.

*****************************************************
*****************************************************
```

1.4.5 grimms-fairy-tales_P1.txt

1.4.5.1 grimms-fairy-tales_P1.txt Unigram Count and Probabilities

```
**********************************************************
top 20 unigram count

1- the ==> 744
2- and ==> 541
3- to ==> 271
4- a ==> 244
5- he ==> 232
6- his ==> 142
7- was ==> 139
8- in ==> 133
9- of ==> 127
10- it ==> 108
11- that ==> 105
12- you ==> 105
13- said ==> 95
14- as ==> 93
15- but ==> 89
16- him ==> 89
17- they ==> 88
18- so ==> 86
19- had ==> 85
20- for ==> 82
```

```
**********************************************************
top 20 unigram probabilities

1- the ==> 0,072352
2- and ==> 0,052611
3- to ==> 0,026354
4- a ==> 0,023728
5- he ==> 0,022562
6- his ==> 0,013809
7- was ==> 0,013517
8- in ==> 0,012934
9- of ==> 0,012350
10- it ==> 0,010503
11- that ==> 0,010211
12- you ==> 0,010211
13- said ==> 0,0092385
14- as ==> 0,0090441
15- but ==> 0,0086551
16- him ==> 0,0086551
17- they ==> 0,0085578
18- so ==> 0,0083633
19- had ==> 0,0082661
20- for ==> 0,0079743
```

1.4.5.2 grimms-fairy-tales_P1.txt Unigram Count and Probabilities

```
**********************************************************
top 20 bigram count

1- in the ==> 60
2- and the ==> 53
3- to the ==> 47
4- of the ==> 35
5- said the ==> 32
6- into the ==> 27
7- on the ==> 24
8- the fox ==> 22
9- he was ==> 21
10- then the ==> 21
11- and said ==> 20
12- as he ==> 20
13- at last ==> 20
14- but the ==> 20
15- the king ==> 20
16- so he ==> 19
17- came to ==> 18
18- the golden ==> 17
19- then he ==> 17
20- to be ==> 17
```

```
**********************************************************
top 20 bigram probabilities

1- able to ==> 1,0000
2- abode there ==> 1,0000
3- about’ ‘why’ ==> 1,0000
4- above its ==> 1,0000
5- according to ==> 1,0000
6- accordingly so ==> 1,0000
7- account of ==> 1,0000
8- across and ==> 1,0000
9- added the ==> 1,0000
10- advice’ he ==> 1,0000
11- alas there ==> 1,0000
12- alas’ said ==> 1,0000
13- alight upon ==> 1,0000
14- all’ thought ==> 1,0000
15- aloud ‘what ==> 1,0000
16- already half ==> 1,0000
17- altogether from ==> 1,0000
18- amiss let ==> 1,0000
19- amongst the ==> 1,0000
20- answered the ==> 1,0000
```

1.4.5.3 grimms-fairy-tales_P1.txt Unigram Count and Probabilities

```
**************************************************
top 20 trigram count

1- came to the ==> 11
2- the king and ==> 10
3- and in the ==> 9
4- at last he ==> 9
5- he came to ==> 8
6- said the ass ==> 8
7- in the morning ==> 7
8- out of the ==> 7
9- the golden bird ==> 7
10- the young man ==> 7
11- and when he ==> 6
12- king and queen ==> 6
13- the side of ==> 6
14- till at last ==> 6
15- to the king ==> 6
16- he could not ==> 5
17- he went to ==> 5
18- in his hand ==> 5
19- that he could ==> 5
20- the fox said ==> 5
```

```
**************************************************
top 20 trigram probabilities

1- a bad job ==> 1,0000
2- a beam on ==> 1,0000
3- a better fate ==> 1,0000
4- a bit of ==> 1,0000
5- a bone or ==> 1,0000
6- a box on ==> 1,0000
7- a bridge’ the ==> 1,0000
8- a broomstick in ==> 1,0000
9- a burning coal ==> 1,0000
10- a bush and ==> 1,0000
11- a butcher soon ==> 1,0000
12- a cage and ==> 1,0000
13- a candle and ==> 1,0000
14- a certain king ==> 1,0000
15- a chamber in ==> 1,0000
16- a christening ‘feel’ ==> 1,0000
17- a club and ==> 1,0000
18- a coach to ==> 1,0000
19- a cock perched ==> 1,0000
20- a common rough ==> 1,0000

The ngram probability calculation of the grimms-fairy-tales_P1.txt took 146 millisecond

**************************************************
**************************************************
```