Assignment 01

By Shivam Bansal, Roll No. 170101063

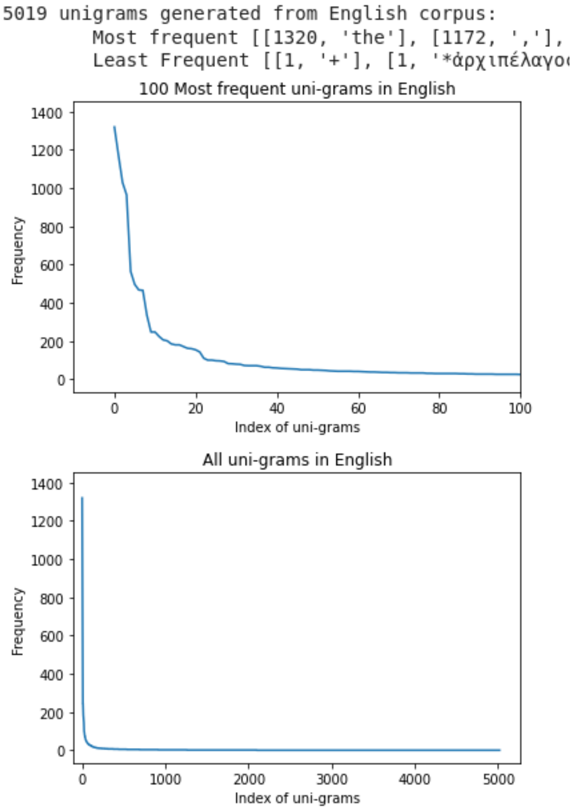
Link to Colab Notebook

1.3.1 Analysis using existing NLP tools

Tools explored

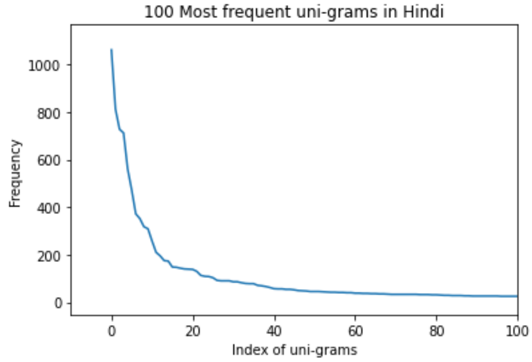
- I used NLTK and spaCy libraries for sentence segmentation and word tokenization of English corpus
- I used Stanza and Indic NLP libraries for sentence segmentation and word tokenization of Hindi corpus
- Since Stanza is very slow I used only first 1000 lines of input with Stanza library
- For all subsequent questions NLTK library was used for English and Indic NLP for Hindi

Uni-grams

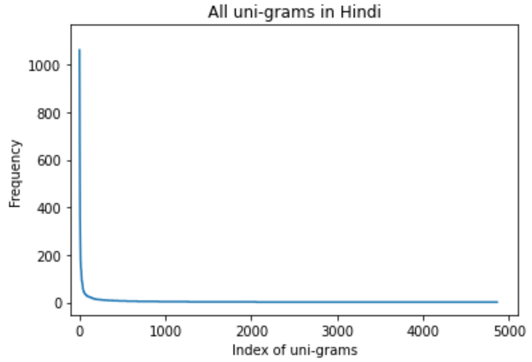


```
5019 unigrams generated from English corpus:
        Most frequent [[1320, 'the'], [1172, ','],
        Least Frequent [[1, '+'], [1, '*ἀρχιπέλαγο
```

```
4863 unigrams generated from Hindi corpus:
        Most frequent [[1062, 'के'], [813, '।'], [72
        Least Frequent [[1, '1218'], [1, '1206516']
```



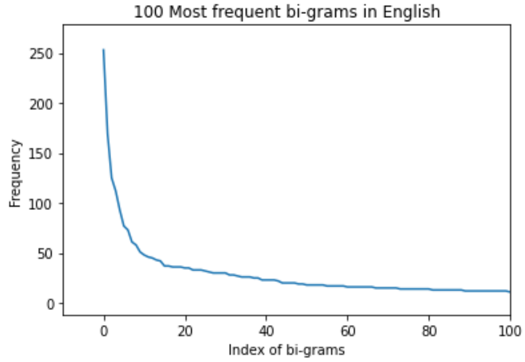
100 Most frequent uni-grams in Hindi



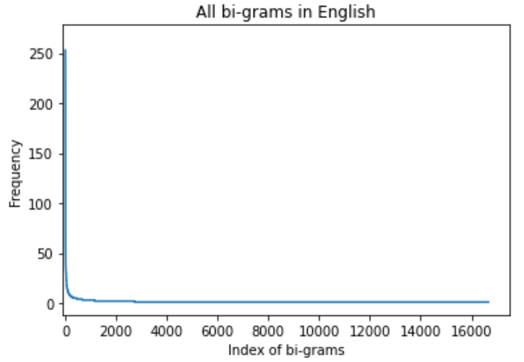
All uni-grams in Hindi

Bi-grams

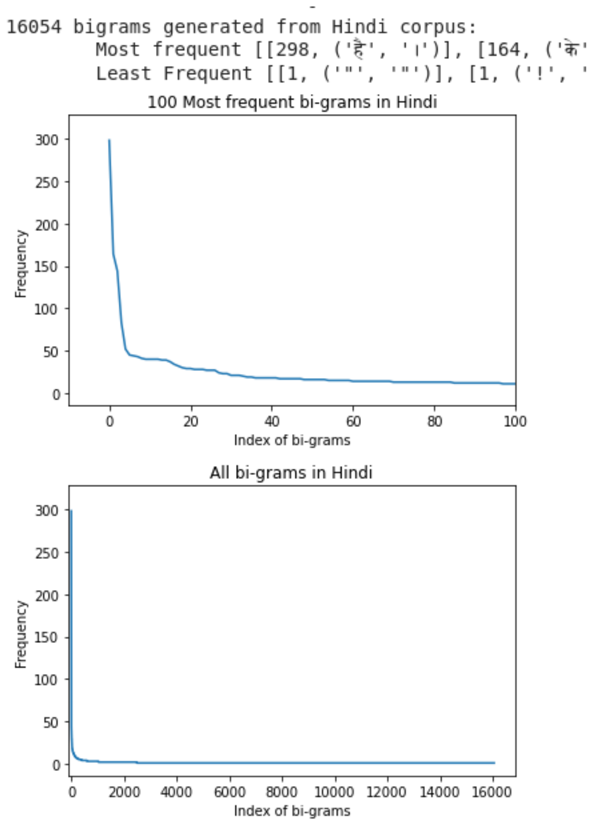
```
16665 bigrams generated from English corpus:
        Most frequent [[253, ('of', 'the')], [169, ('
        Least Frequent [[1, ('%', 'every')], [1, ('%'
```



100 Most frequent bi-grams in English



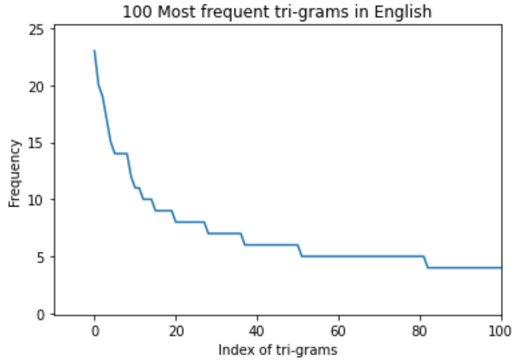
All bi-grams in English



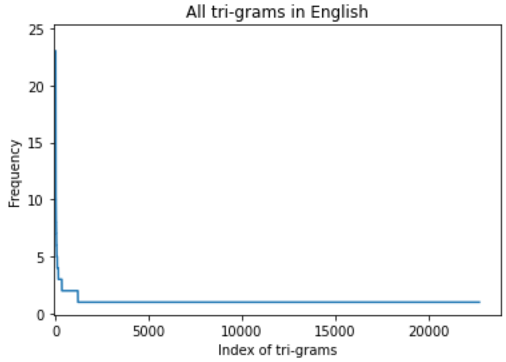
```
16054 bigrams generated from Hindi corpus:
        Most frequent [[298, ('है', '।')], [164, ('के'
        Least Frequent [[1, ('"', '"')], [1, ('!', '
```

Tri-grams

```
22748 trigrams generated from English corpus:
        Most frequent [[23, ('.', 'However', ',')], [
        Least Frequent [[1, ('%', ')', 'to')], [1, ('
```

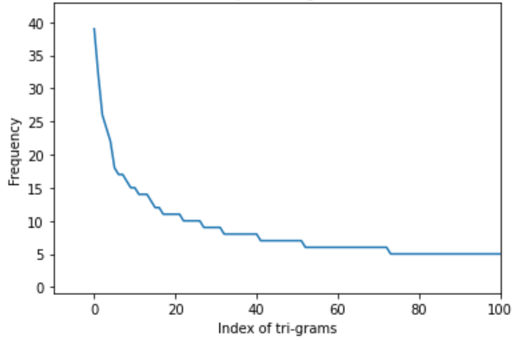


100 Most frequent tri-grams in English

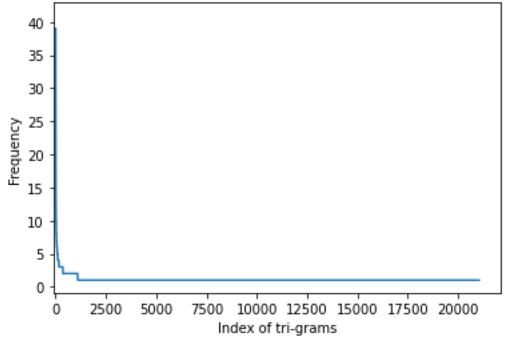


All tri-grams in English

```
21043 trigrams generated from Hindi corpus:
        Most frequent [[39, ('के', 'रूप', 'में')], [32, (
        Least Frequent [[1, ('!', 'शंकर', 'तो')], [1, (
```



100 Most frequent tri-grams in Hindi



All tri-grams in Hindi

1.3.2 Few Basic Questions

Coverage using n-grams

Unigrams

```
ENGLISH
=======
Size of corpora = 24709 words, 5019 unigrams
90% of corpora = 22238 words
No. of unigrams to cover 90% corpora = 2549 unigrams

HINDI
=====
Size of corpora = 23135 words, 4863 unigrams
90% of corpora = 20821 words
No. of unigrams to cover 90% corpora = 2550 unigrams
```

Bi-grams

```
ENGLISH
=======
Size of corpora = 24709 words, 16665 bigrams
80% of corpora = 19767 words
No. of bigrams to cover 80% corpora = 2284 bigrams

HINDI
=====
Size of corpora = 23135 words, 16054 bigrams
80% of corpora = 18508 words
No. of bigrams to cover 80% corpora = 2333 bigrams
```

Tri-grams

```
ENGLISH
=======
Size of corpora = 24709 words, 22748 trigrams
70% of corpora = 17296 words
No. of trigrams to cover 70% corpora = 6690 trigrams

HINDI
=====
Size of corpora = 23135 words, 21043 trigrams
70% of corpora = 16194 words
No. of trigrams to cover 70% corpora = 6008 trigrams
```

Stemming

Uni-grams

```
ENGLISH
=======
3942 stemmed unigrams generated:
  Most frequent [[1320, 'the'], [1172, ','], [1030, 'of'], [965, '.'], [565, 'and'], [497, 'in'], [468, 'to'], [466, 'a'], [336, 'is
  Least Frequent [[1, '1.007825'], [1, '1,839'], [1, '1,836'], [1, '1,386'], [1, '0.25'], [1, '0.012'], [1, '+'], [1, '*ἀρχιπέλαγος'
Size of corpora = 24709 words, 3942 unigrams
90% of corpora = 22238 words
No. of unigrams to cover 90% corpora = 1684 unigrams
HINDI
=====
3975 stemmed unigrams generated:
  Most frequent [[2324, 'क'], [813, '।'], [786, 'म'], [712, ','], [561, 'है'], [372, 'और'], [355, 'स'], [303, 'कर'], [259, 'ज'], [258
  Least Frequent [[1, '1218'], [1, '1206516'], [1, '119'], [1, '1048'], [1, '104'], [1, '10,15'], [1, '1.6'], [1, '06'], [1, '04'],
Size of corpora = 23135 words, 3975 unigrams
90% of corpora = 20821 words
No. of unigrams to cover 90% corpora = 1718 unigrams
```

Bi-grams

```
ENGLISH
=======
16009 stemmed bigrams generated:
  Most frequent [[253, ('of', 'the')], [169, ('.', 'The')], [125, ('in', 'the')], [112, (',', 'and')], [93, (',', 'the')], [77, ('.'
  Least Frequent [[1, ('%', 'everi')], [1, ('%', 'compar')], [1, ('$', '35')], [1, ('$', '300')], [1, ('$', '3.74')], [1, ('$', '25–
Size of corpora = 24709 words, 16009 bigrams
80% of corpora = 19767 words
No. of bigrams to cover 80% corpora = 2044 bigrams
HINDI
=====
14588 stemmed bigrams generated:
  Most frequent [[298, ('है', '।')], [164, ('क', 'ल')], [144, ('हैं', '।')], [105, ('थ', '।')], [83, ('है', ',')], [75, ('कर', 'क')], [60,
  Least Frequent [[1, ('"', '"')], [1, ('!', 'शम्भुक')], [1, ('!', 'शंकर')], [1, ('!', 'यद')], [1, ('!', 'मयद्वार')], [1, ('!', 'मय')], [1,
Size of corpora = 23135 words, 14588 bigrams
80% of corpora = 18508 words
No. of bigrams to cover 80% corpora = 1677 bigrams
```

Tri-grams

```
ENGLISH
=======
22587 stemmed trigrams generated:
  Most frequent [[23, ('.', 'Howev', ',')], [20, ('.', 'It', 'is')], [19, (',', 'and', 'the')], [17, (',', 'such', 'as')], [15, (')'
  Least Frequent [[1, ('%', ')', 'to')], [1, ('%', ')', 'have')], [1, ('$', '35', 'per')], [1, ('$', '300', 'in')], [1, ('$', '3.74'
Size of corpora = 24709 words, 22587 trigrams
70% of corpora = 17296 words
No. of trigrams to cover 70% corpora = 6529 trigrams
HINDI
=====
20559 stemmed trigrams generated:
  Most frequent [[39, ('क', 'रूप', 'म')], [37, ('ज', 'है', '।')], [32, ('कर', 'क', 'ल')], [26, ('ज', 'सक', 'है')], [25, ('सक', 'है', '।'
  Least Frequent [[1, ('!', 'शंकर', 'त')], [1, ('!', 'यद', 'आप')], [1, ('!', 'मैं', 'सद')], [1, ('!', 'मैं', 'तुझपर')], [1, ('!', 'मयद्वार', '
Size of corpora = 23135 words, 20559 trigrams
70% of corpora = 16194 words
No. of trigrams to cover 70% corpora = 5524 trigrams
```

Comparing no-stemming VS. stemming

```
           Without stemming    |   With stemming
           ----------------    |   -------------
=======
ENGLISH
=======
Size of corpora        24709       |   24709
No. of unigrams        5019        |   3942
Unigrams for 90% coverage    2549        |   1684
No. of bigrams         16665       |   16009
Bigrams for 80% coverage   2284        |   2044
No. of trigrams        22748       |   22587
Trigrams for 70% coverage    6690        |   6529
=======
HINDI
=======
Size of corpora        23135       |   23135
No. of unigrams        4863        |   3975
Unigrams for 90% coverage    2550        |   1718
No. of bigrams         16054       |   14588
Bigrams for 80% coverage   2333        |   1677
No. of trigrams        21043       |   20559
Trigrams for 70% coverage    6008        |   5524
```

1.3.3 Writing basic codes

Heuristic based sentence segmentation and word tokenization

I followed the procedure described in class.

Sentence segmentation

- Mark temporary sentence end if punctuation like period, question mark or exclamation mark is encountered
- Check if there is following quotation mark
- Check if there is no space or small case alphabet after period
- Trim whitespace

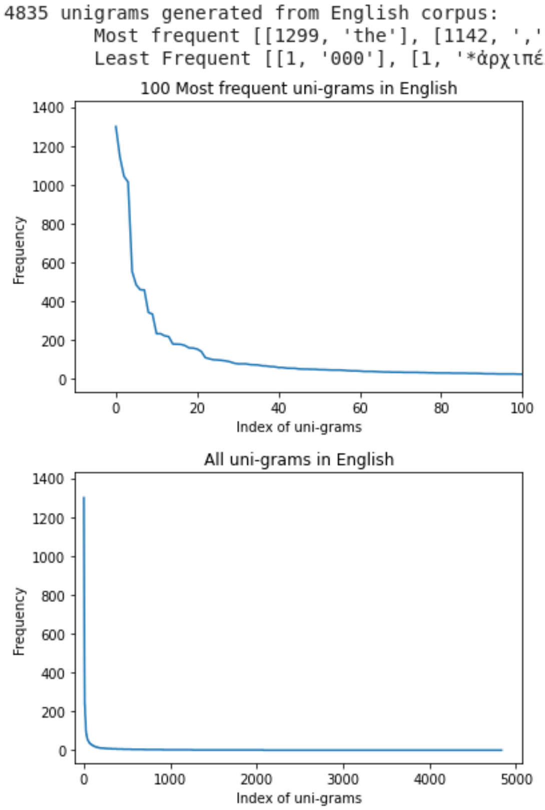
Word tokenization

- Try to split on whitespace
- Separate special characters like : ; ' " , . / ? \ |] [{ } ( )
- Do not separate each digit in numbers

```
32811 sentences in en_heur_sents. First 5 sentences:
  ['The word "atom" was coined by ancient Greek philosophers.', 'However, these ideas were founded in philosophical and theological
24756 words in en_heur_words. First 50 words:
  ['The', 'word', '"', 'atom', '"', 'was', 'coined', 'by', 'ancient', 'Greek', 'philosophers', '.', 'However', ',', 'these', 'ideas'
```

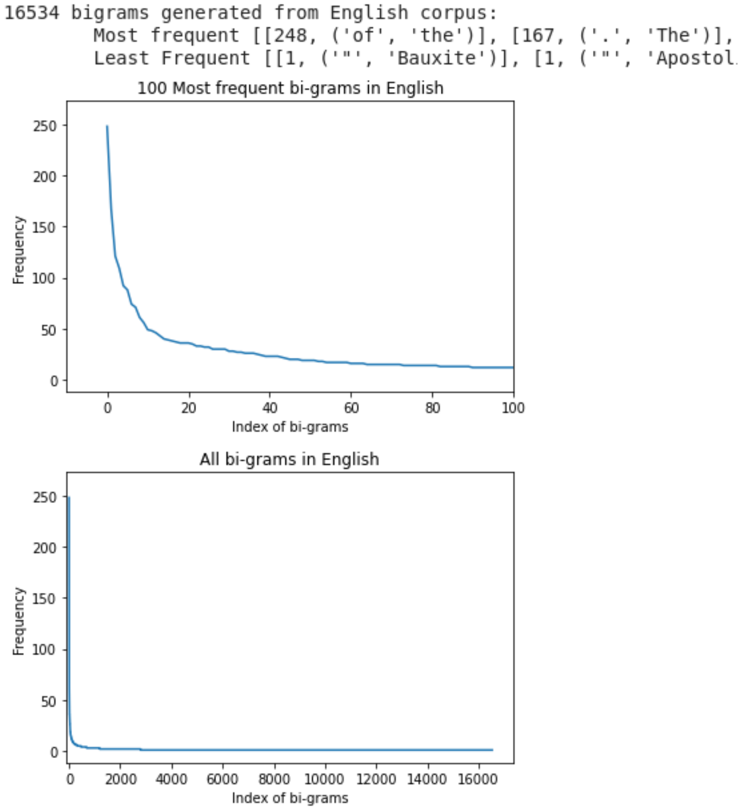
N-grams

Uni-grams



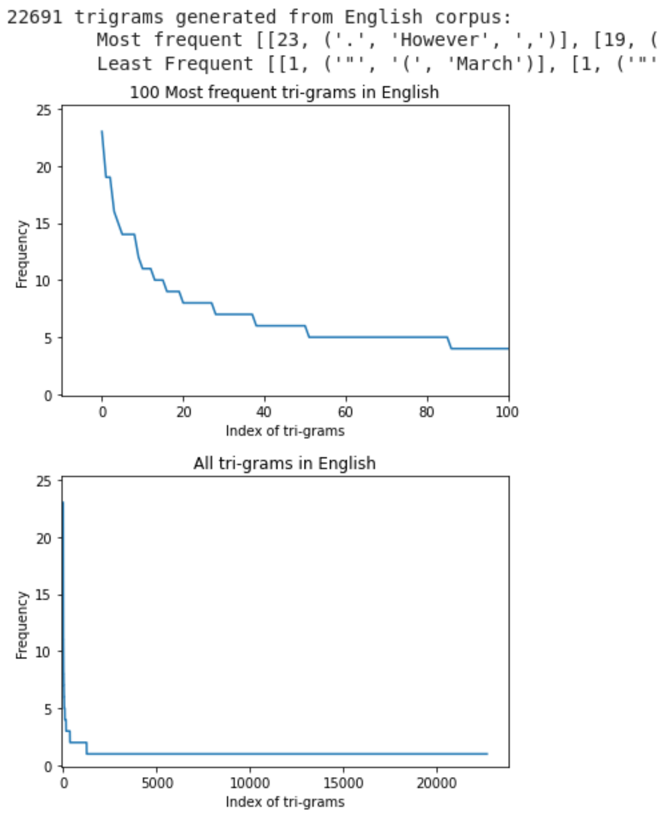
```
4835 unigrams generated from English corpus:
        Most frequent [[1299, 'the'], [1142, ','], [1043,
        Least Frequent [[1, '000'], [1, '*ἀρχιπέλαγος'], [
```

Bi-grams



```
16534 bigrams generated from English corpus:
        Most frequent [[248, ('of', 'the')], [167, ('.', 'The')],
        Least Frequent [[1, ('"', 'Bauxite')], [1, ('"', 'Apostol
```

Tri-grams



```
22691 trigrams generated from English corpus:
        Most frequent [[23, ('.', 'However', ',')], [19, (
        Least Frequent [[1, ('"', '(', 'March')], [1, ('"'
```

Coverage of n-grams

```
Size of corpora = 24756 words, 4835 unigrams
90% of corpora = 22280 words
No. of unigrams to cover 90% corpora = 2360 unigrams

Size of corpora = 24756 words, 16534 bigrams
80% of corpora = 19804 words
No. of bigrams to cover 80% corpora = 2236 bigrams

Size of corpora = 24756 words, 22691 trigrams
70% of corpora = 17329 words
No. of trigrams to cover 70% corpora = 6602 trigrams
```

Stemming

```
3763 stemmed unigrams generated:
  Most frequent [[1299, 'the'], [1142, ','], [1043, '.'], [1014, 'of'], [554, 'and'], [485, 'in'], [460, 'a'], [459, 'to'], [343, '"
  Least Frequent [[1, '000'], [1, '*ἀρχιπέλαγος'], [1, '&'], [1, '$300'], [1, '$3'], [1, '$25–$35'], [1, '$2000'], [1, '$20'], [1, '
Size of corpora = 24756 words, 3763 unigrams
90% of corpora = 22280 words
No. of unigrams to cover 90% corpora = 1584 unigrams
15879 stemmed bigrams generated:
  Most frequent [[248, ('of', 'the')], [167, ('.', 'The')], [121, ('in', 'the')], [109, (',', 'and')], [92, (',', 'the')], [88, ("'"
  Least Frequent [[1, ('"', 'Brownian')], [1, ('"', 'Belknap')], [1, ('"', 'Bauxit')], [1, ('"', 'Apostolica')], [1, ('"', 'Als')],
Size of corpora = 24756 words, 15879 bigrams
80% of corpora = 19804 words
No. of bigrams to cover 80% corpora = 1999 bigrams
22524 stemmed trigrams generated:
  Most frequent [[23, ('.', 'Howev', ',')], [19, ('.', 'It', 'is')], [19, (',', 'and', 'the')], [16, (',', 'such', 'as')], [15, (')'
  Least Frequent [[1, ('"', '(', 'March')], [1, ('"', '(', 'Latin')], [1, ('"', '(', 'Foreign')], [1, ('"', '(', '2011')], [1, ('"',
Size of corpora = 24756 words, 22524 trigrams
70% of corpora = 17329 words
No. of trigrams to cover 70% corpora = 6435 trigrams
```

Comparing no-stemming VS. stemming

```
           Without stemming    |   With stemming
           ----------------    |   -------------
=======
ENGLISH
=======
Size of corpora        24756       |   24756
No. of unigrams        4835        |   3763
No. of unigrams for 90% coverage   2360        |   1584
No. of bigrams         16534       |   15879
No. of bigrams for 80% coverage    2236        |   1999
No. of trigrams        22691       |   22524
No. of trigrams for 70% coverage   6602        |   6435
```