CS 565
Intelligent Systems and Interfaces

Assignment I
Basic Preprocessing: Segmentation, N-Gram Analysis, Collocation

Group: Agent007
Dheeraj Khatri, 120101021
Dhruv Kohli, 120123054

Python-NLTK

Corpus: **austen-emma (from nltk)**

1.1

Number of sentences: **7493**

1.2

Number of words in dictionary: **8466**

1.3

Number of unigrams: **8466**
Top 10 unigrams in monotonically decreasing order of frequencies:

Unigram	**Frequency**
,	12016
.	6357
to	5124
the	4842
and	4652
of	4272
I	3164
--	3100
a	3001
''	2452
was	2383

***Graphs at the end.**

1.4

Total Number of Cont. bigrams: **65313**
Top 10 bigrams in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, and	1880
. ''	1158
'' ``	959
; and	867
to be	592
, ''	584
. I	570
, I	569
of the	556
in the	434
; but	427

Top 10 Cont. bigrams collocations using Mutual Information measure:

26th	ult.
Abominable	scoundrel
Agricultural	Reports
Austen	1816
Baronne	d'Almane
Candles	everywhere.
Clayton	Park
Comtesse	d'Ostalis
DEAR	MADAM
Farmer	Mitchell
Former	provocations

Total Number of Non Cont. bigrams (**Window size = 25**): **1022485**
Top 10 Non Cont. bigrams (**Window size = 25**) in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, ,	17660
. ,	9002
, .	8543
, and	8242
, to	7847
the ,	7555
to ,	7447
, the	7422
and ,	7020
of ,	6640
, of	6406

Top 10 Non Cont. bigrams collocations (**Window size = 25**) in monotonically decreasing order of frequencies:

could.	Hughes
26th	ult.
8th	23rd
8th	birthday
Abbots	peeped
Abdy	clerk
Abominable	scoundrel
Abominable	steadier
According	pressingly
Acquit	acquittal
Adopt	educate

1.5

Number of Cont. trigrams: **138793**
Top 10 Cont. trigrams in monotonically decreasing order of frequencies:

Trigram	Frequency
. " ``	758
, '' said	225
? " ``	147
" `` I	136
I do not	135
. It was	117
I am sure	105
, and the	89
, however ,	89
, my dear	87
Miss Woodhouse ,	86

Top 10 Cont. trigrams collocations using Mutual Information measure:

MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
The	_Rev._	_Philip_
_be_	_a_	_source_
repentance	_and_	_misery_
Austen	1816	]
C.	WESTON	CHURCHILL
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
Rev.	_Philip_	_Elton_

Number of Non Cont. trigrams (**Window size = 10**): **4674727**
Top 10 Non Cont. trigrams (**Window size = 10**) in monotonically decreasing order of frequencies:

Trigram	**Frequency**
, , ,	1096
, and ,	1024
, , and	1013
. " ``	928
, the of	788
. `` ,	776
. " ,	740
" `` ,	727
, the ,	697
the of ,	680
. , ,	665

Top 10 Non Cont. trigrams collocations (**Window size = 10**) in monotonically decreasing order of frequencies:

MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
The	_Rev._	_Philip_
_be_	_a_	_source_
repentance	_and_	_misery_
Austen	1816	]
C.	WESTON	CHURCHILL
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
Rev.	_Philip_	_Elton_

1.7

Sentence Tokenizer

- **Punkt Sentence Tokenizer**:This tokenizer divides a text into a list of sentences, by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences.
- **LineTokenizer**:Tokenize a string into its lines, optionally discarding blank lines.
- **Stanford Tokenizer**:Use stanford's PTBTokenizer to tokenize multiple sentences.

Word Tokenizer

- NLTK Tokenizer Package: Tokenizers divide strings into lists of substrings.
- **TreebankWordTokenizer**:The Treebank tokenizer uses regular expressions to tokenize text as in Penn Treebank.This is the method that is invoked by word_tokenize().
- **MWETokenizer:Multi-Word Expression Tokenizer** : takes a string which has already been divided into tokens and re-tokenizes it, merging multi-word expressions into single tokens, using a lexicon of MWEs
- **RegexpTokenizer**: splits a string into substrings using a regular expression.
- **WhitespaceTokenizer**: Tokenize a string on whitespace (space, tab, newline).
- **WordPunctTokenizer**: Tokenize a text into a sequence of alphabetic and non-alphabetic characters, using the regexp \w+|[^\w\s]+.
- **S-Expression Tokenizer**: is used to find parenthesized expressions in a string.
- **Simple Tokenizers**:These tokenizers divide strings into substrings using the string split() method. When tokenizing using a particular delimiter string, use the string split() method directly, as this is more efficient.
- **SpaceTokenizer**: Tokenize a string using the space character as a delimiter.
- **TabTokenizer**: Tokenize a string use the tab character as a delimiter.

ML Based:

- **Memory Networks paper by Facebook AI Research presents a straight-forward way of using Embeddings and Threshold based Linear Regression to segment lines with a streaming words as input (represented as BOW)**
- **Logistic regression with BOW representation of input (streaming words) are also popular to classify an accumulation of words as sentence.**

2.1

Number of most frequent words required for **90.0 %** coverage: **1204** out of **8466**

2.2

Number of most frequent Cont. bigrams required for **80.0 %** coverage: **26979** out of **65313**
Number of most frequent Non. Cont. bigrams (**Window size = 25**) required for **80.0 %** coverage: **225079** out of **1022485**

2.3

Number of most frequent Cont. trigrams required for **70.0 %** coverage: **81292** out of **138793**
Number of most frequent Non. Cont. trigrams (**Window size = 10**) required for **70.0 %** coverage: **2604710** out of **4674727**

2.4. Lemmatization + NLTK

Number of unigrams: **7811**
Top 10 unigrams in monotonically decreasing order of frequencies:

Unigram	**Frequency**
,	12016
.	6357
to	5124
the	4842
and	4652
a	4388
of	4272
I	3164
--	3100
''	2452

Number of Cont. bigrams: **63732**
Top 10 Cont. bigrams in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, and	1880
. "	1158
" ``	959
; and	867
to be	592
, "	584
. I	570
, I	569
of the	556
in the	434
; but	427

Top 10 Cont. bigrams collocations using Mutual Information measure:

26th	ult.
Abominable	scoundrel
Agricultural	Reports
Austen	1816
Baronne	d'Almane
Candles	everywhere.
Clayton	Park
Comtesse	d'Ostalis
DEAR	MADAM
Farmer	Mitchell
Italian	singing.

Number of Non Cont. bigrams (**Window size = 25**): **977512**
Top 10 Non Cont. bigrams (**Window size = 25**) in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, ,	17660
. ,	9002
, .	8543
, and	8242
, to	7847
the ,	7555
to ,	7447
, the	7422
and ,	7020
, a	6845
a ,	6699

Top 10 Non Cont. bigrams collocations (**Window size = 25**) in monotonically decreasing order of frequencies:

could.	Hughes
26th	ult.
8th	23rd
8th	birthday
Abbots	peeped
Abominable	scoundrel
Abominable	steadier
According	pressingly
Acquit	acquittal
Adopt	educate
Agreed	Low

Number of Cont. trigrams: **138091**
Top 10 Cont. trigrams in monotonically decreasing order of frequencies:

Trigram	**Frequency**
. " ``	758
, " said	225
? " ``	147
" `` I	136
I do not	135
. It wa	117
I am sure	105
, and the	89
, however ,	89
, my dear	88
Miss Woodhouse ,	86

Top 10 Cont. trigrams collocations using Mutual Information measure:

MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
The	_Rev._	_Philip_
_be_	_a_	_source_
repentance	_and_	_misery_
Austen	1816	]
C.	WESTON	CHURCHILL
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
Rev.	_Philip_	_Elton_

Number of Non Cont. trigrams (**Window size = 25**): **4610090**
Top 10 Non Cont. trigrams (**Window size = 10**) in monotonically decreasing order of frequencies:

Trigram	**Frequency**
, , ,	1096
, and ,	1024
, , and	1013
. " ``	928
, the of	788
. `` ,	776
. " ,	740
" `` ,	727
, a ,	715
, the ,	697
the of ,	680

Top 10 Non Cont. trigrams collocations (Window size = 10) in monotonically decreasing order of frequencies:

MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
The	_Rev._	_Philip_
_be_	_a_	_source_
repentance	_and_	_misery_
Austen	1816	]
C.	WESTON	CHURCHILL
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
Rev.	_Philip_	_Elton_

Number of most frequent words required for **90.0 %** coverage: **1081** out of **7811**

Number of most frequent Cont. bigrams required for **80.0 %** coverage: **25398** out of **63732**
Number of most frequent Non. Cont. bigrams (**Window size = 25**) required for **80.0 %** coverage: **200615** out of **977512**

Number of most frequent Cont. trigrams required for **70.0 %** coverage: **80590** out of **138091**
Number of most frequent Non. Cont. trigrams (**Window size = 10**) required for **70.0 %** coverage: **2540073** out of **4610090**

2.5

- As expected, the total number of unigrams, bigrams and trigrams decreased after lemmatization.

3.1 Section 1 and 2 after Heuristic based segmentation

Using putative delimiters: **['.', '?', '!', '-', ';', ':']**
Number of sentences after putting putative sentence boundaries: **19868**
Using sentence mergers: **['dr', 'mr', 'ms', 'mrs', 'vs'] and** lower case after ? and name after !
Number of sentences after merging: **18055**
Number of words in dictionary: **9169**
Number of unigrams: **9169**
Top 20 unigrams in monotonically decreasing order of frequencies:

Unigram	**Frequency**
,	11461
.	8844
to	5173
the	4837
and	4643
of	4268
a	3001
I	2833
-	2776
was	2376
her	2355

Number of Cont. bigrams: **67058**
Top 20 Cont. bigrams in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, and	1879
Mr .	1124
; and	907
Mrs .	687
. I	646
to be	592
, I	569
of the	556
. She	496
; but	448
in the	434

Top 10 Cont. bigrams collocations using Mutual Information measure:

"Christmas	weather,
"Four	o'clock!
"Happy	couple!
"Highbury	gossips!
"Lord	bless
"Men's	Beavers
"Success	supposes
"York	Tan
"_His_	sufferings,
"ready	wit"
"soft	eyes"

Number of Non Cont. bigrams (**Window size = 25**): **1076147**
Top 10 Non Cont. bigrams (**Window size = 25**) in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, ,	16385
. ,	12324
, .	11632
. .	9421
, and	8016
, to	7708
the ,	7295
to ,	7283
, the	7225
and ,	6829
of ,	6410

Top 10 Non Cont. bigrams collocations (**Window size = 25**) in monotonically decreasing order of frequencies:

"About	Oh!
"About	owning
"Agreed	"Low
"Agreed	reckon?
"Be	outcry
"Better	doat
"Both	lady?
"Cautious	believes
"Cautious	cautious,
"Charming	interpret
"Christmas	seasonable

Number of Cont. trigrams: **139314**
Top 10 Cont. trigrams in monotonically decreasing order of frequencies:

Trigrams	**Frequency**
Mr . Knightley	254
Mrs . Weston	222
. Mr .	174
Mr . Elton	173
. It was	138
Mr . Weston	138
Mrs . Elton	118
I do not	108
Mr . Woodhouse	107
. Weston ,	105
I am sure	104

Top 10 Cont. trigrams collocations using Mutual Information measure:

Hymen's	saffron	robe
MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
_be_	_a_	_source_
repentance	_and_	_misery_
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
_a_	_source_	_of_
felt	_the_	_engagement_
_of_	_repentance_	_and_

Number of Non Cont. trigrams (**Window size = 10**): **4665456**
Top 10 Non Cont. trigrams (**Window size = 10**) in monotonically decreasing order of frequencies:

Trigram	**Frequency**
, and ,	987
, , ,	958
, , and	942
. , ,	916
, the of	785
, and .	708
Mr . ,	679
, the ,	665
the of ,	659
. , and	652
, and the	640

Top 10 Non Cont. trigrams collocations (**Window size = 10**) in monotonically decreasing order of frequencies:

Hymen's	saffron	robe
MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
_be_	_a_	_source_
repentance	_and_	_misery_
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
_a_	_source_	_of_
felt	_the_	_engagement_
_of_	_repentance_	_and_

Number of most frequent words required for **90.0 %** coverage: **1400** out of **9169**

Number of most frequent Cont. bigrams required for **80.0 %** coverage: **29406** out of 67058
Number of most frequent Non. Cont. bigrams (**Window size = 25**) required for **80.0 %** coverage: **261989** out of **1076147**

Number of most frequent Cont. trigrams required for **70.0 %** coverage: **82836** out of **139314**
Number of most frequent Non. Cont. trigrams (**Window size = 10**) required for **70.0 %** coverage: **2632245** out of **4665456**

Lemmatization + Heuristic

Number of words in dictionary: **8506**
Number of unigrams: **8506**
Top 10 unigrams in monotonically decreasing order of frequencies:

Unigram	**Frequency**
,	11461
.	8844
to	5173
the	4837
and	4643
a	4376
of	4268
I	2833
-	2776
wa	2376
her	2355

Number of Cont. bigrams: **65470**
Top 10 Cont. bigrams in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, and	1879
Mr .	1124
; and	907
Mrs .	687
. I	646
to be	592
, I	569
of the	556
. She	496
; but	448

Top 10 Cont. bigrams collocations using Mutual Information measure:

"Christmas	weather,
"Four	o'clock!
"Happy	couple!
"Highbury	gossips!
"Lord	bless
"Men's	Beavers
"Success	supposes
"York	Tan
"_His_	sufferings,
"ready	wit"
"soft	eyes"

Number of Non Cont. bigrams (**Window size = 25**): **1031412**
Top 10 Non Cont. bigrams (**Window size = 25**) in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, ,	16385
. ,	12324
, .	11632
. .	9421
, and	8016
, to	7708
the ,	7295
to ,	7283
, the	7225
and ,	6829
, a	6643

Top 10 Non Cont. bigrams collocations (**Window size = 25**) in monotonically decreasing order of frequencies:

"About	Oh!
"About	owning
"Agreed	"Low
"Agreed	reckon?
"Be	outcry
"Better	doat
"Both	lady?
"Cautious	cautious,
"Charming	interpret
"Christmas	seasonable
"Christmas	weather,

Number of Cont. trigrams: **138630**
Top 10 Cont. trigrams in monotonically decreasing order of frequencies:

Trigram	**Frequency**
Mr . Knightley	254
Mrs . Weston	222
. Mr .	174
Mr . Elton	173
. It wa	138
Mr . Weston	138
Mrs . Elton	118
I do not	108
Mr . Woodhouse	107
. Weston ,	105
I am sure	104

Top 10 Cont. trigrams collocations using Mutual Information measure:

Hymen's	saffron	robe
MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
_be_	_a_	_source_
repentance	_and_	_misery_
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
_a_	_source_	_of_
felt	_the_	_engagement_
_of_	_repentance_	_and_

Number of Non Cont. trigrams (**Window size = 10**): **4601826**
Top 10 Non Cont. trigrams (Window size = 10) in monotonically decreasing order of frequencies:

Trigram	**Frequency**
, and ,	987
, , ,	958
, , and	942
. , ,	916
, the of	785
, and .	708
, a ,	681
Mr . ,	679
, the ,	665
the of ,	659

Top 10 Non Cont. trigrams collocations (Window size = 10) in monotonically decreasing order of frequencies:

Hymen's	saffron	robe
MY	DEAR	MADAM
Madame	de	Genlis
Most	_precious_	_treasures_
_be_	_a_	_source_
repentance	_and_	_misery_
La	Baronne	d'Almane
La	Comtesse	d'Ostalis
_a_	_source_	_of_
felt	_the_	_engagement_
_of_	_repentance_	_and_

Number of most frequent words required for **90.0 %** coverage: **1258** out of **8506**

Number of most frequent Cont. bigrams required for **80.0 %** coverage: **27818** out of **65470**
Number of most frequent Non. Cont. bigrams (**Window size = 25**) required for **80.0 %** coverage: **236471** out of **1031412**

Number of most frequent Cont. trigrams required for **70.0 %** coverage: **82152** out of **138630**
Number of most frequent Non. Cont. trigrams (**Window size = 10**) required for **70.0 %** coverage: **2568615** out of **4601826**

3.2 Chi-Square test Collocations

Java-Stanford CoreNLP

Corpus: **austen-emma (from nltk)**

1.1

Number of sentences: **8618**

1.2

Number of words in dictionary: **7755**

1.3

Number of unigrams: **7755**
Top 10 unigrams in monotonically decreasing order of frequencies:

Unigram	**Frequency**
,	11979
.	6934
to	5120
the	4821
and	4644
of	4258
I	3191
--	3077
a	2998
her	2395
was	2377

1.4

Total Number of Cont. bigrams: **64208**
Top 10 bigrams in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, and	1877
. "	1158
" ``	982
; and	864
. --	714
to be	594
. I	570
, I	569
of the	550
, "	438
in the	431

1.5

Number of Cont. trigrams: **138192**
Top 10 Cont. trigrams in monotonically decreasing order of frequencies:

Trigram	**Frequency**
. " ``	754
, " said	225
? " ``	146
" `` I	138
I do not	135
. It was	117
I am sure	109
`` Oh !	92
, and the	89
, however ,	88
, my dear	87

2.1

Number of most frequent words required for **90.0 %** coverage: **1114** out of **7755**

2.2

Number of most frequent Cont. bigrams required for **80.0 %** coverage: **25665** out of **64208**

2.3

Number of most frequent Cont. trigrams required for **70.0 %** coverage: **80378** out of **138192**

2.4. Lemmatization + Stanford coreNLP

Number of unigrams: **5678**
Top 10 unigrams in monotonically decreasing order of frequencies:

Unigram	**Frequency**
,	11979
be	8201
.	6934
to	5176
the	5175
and	4868
she	4844
of	4270
I	3774
he	3716

Number of Cont. bigrams: **54709**
Top 10 Cont. bigrams in monotonically decreasing order of frequencies:

Bigram	**Frequency**
, and	1879
. "	1158
" ``	982
; and	864
it be	777
. --	714
have be	610
to be	608
. I	570
, I	569
of the	553

Number of Cont. trigrams: **130090**
Top 10 Cont. trigrams in monotonically decreasing order of frequencies:

Trigram	**Frequency**
. " ``	754
, '' say	225
. it be	175
I do not	162
" `` I	147
? " ``	146
I be sure	111
it be a	105
. she be	104
, it be	99
it be not	96

Number of most frequent words required for **90.0 %** coverage: **700** out of **5678**

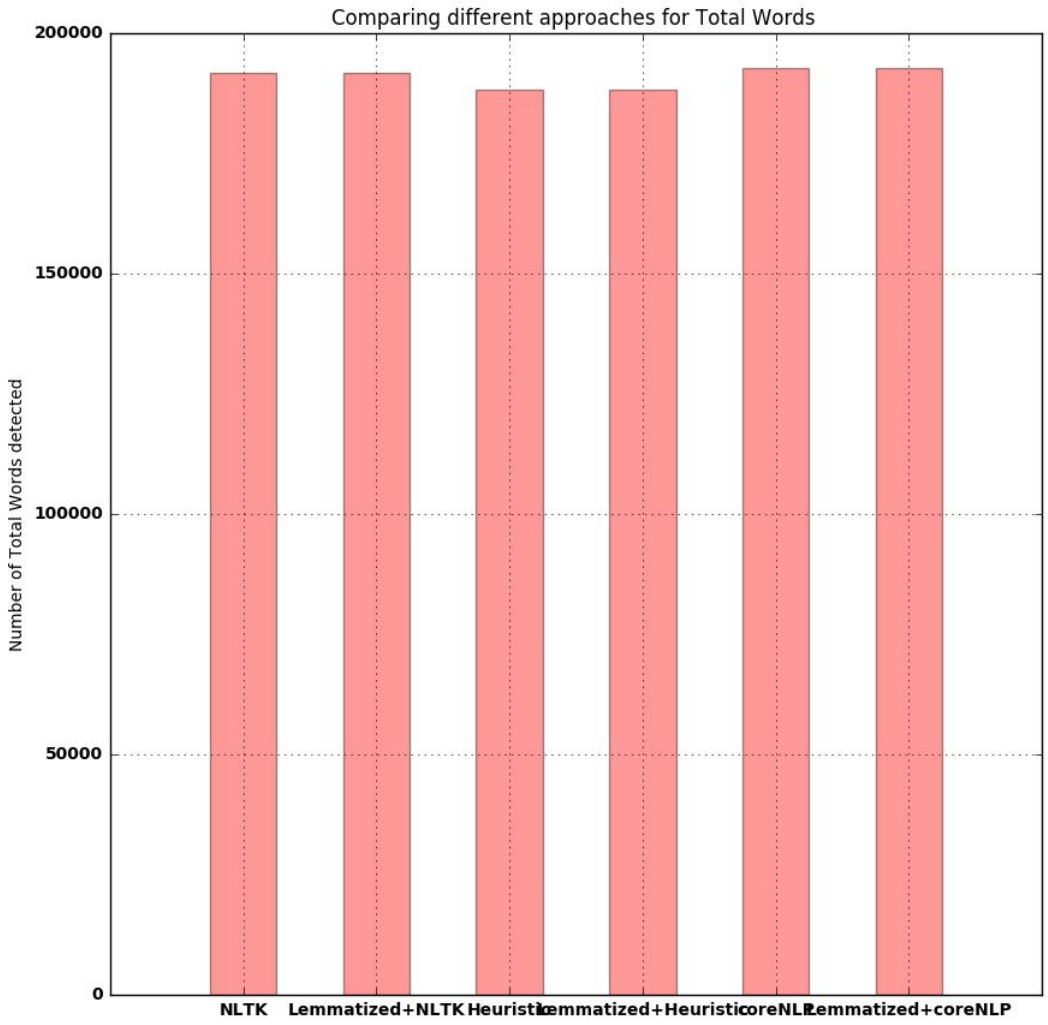
Number of most frequent Cont. bigrams required for **80.0 %** coverage: **17150** out of **54709**

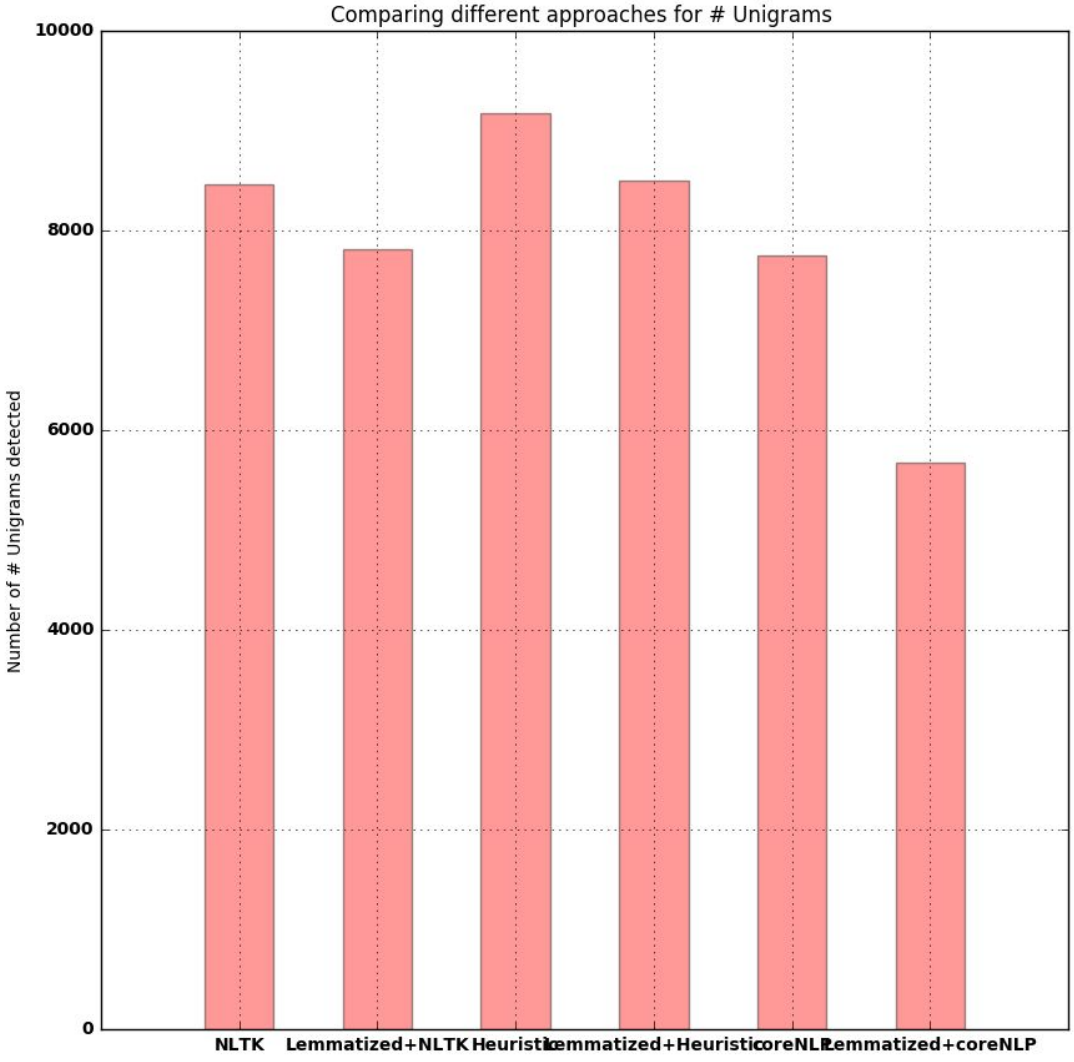
Number of most frequent Cont. trigrams required for **70.0 %** coverage: **72276** out of **130090**

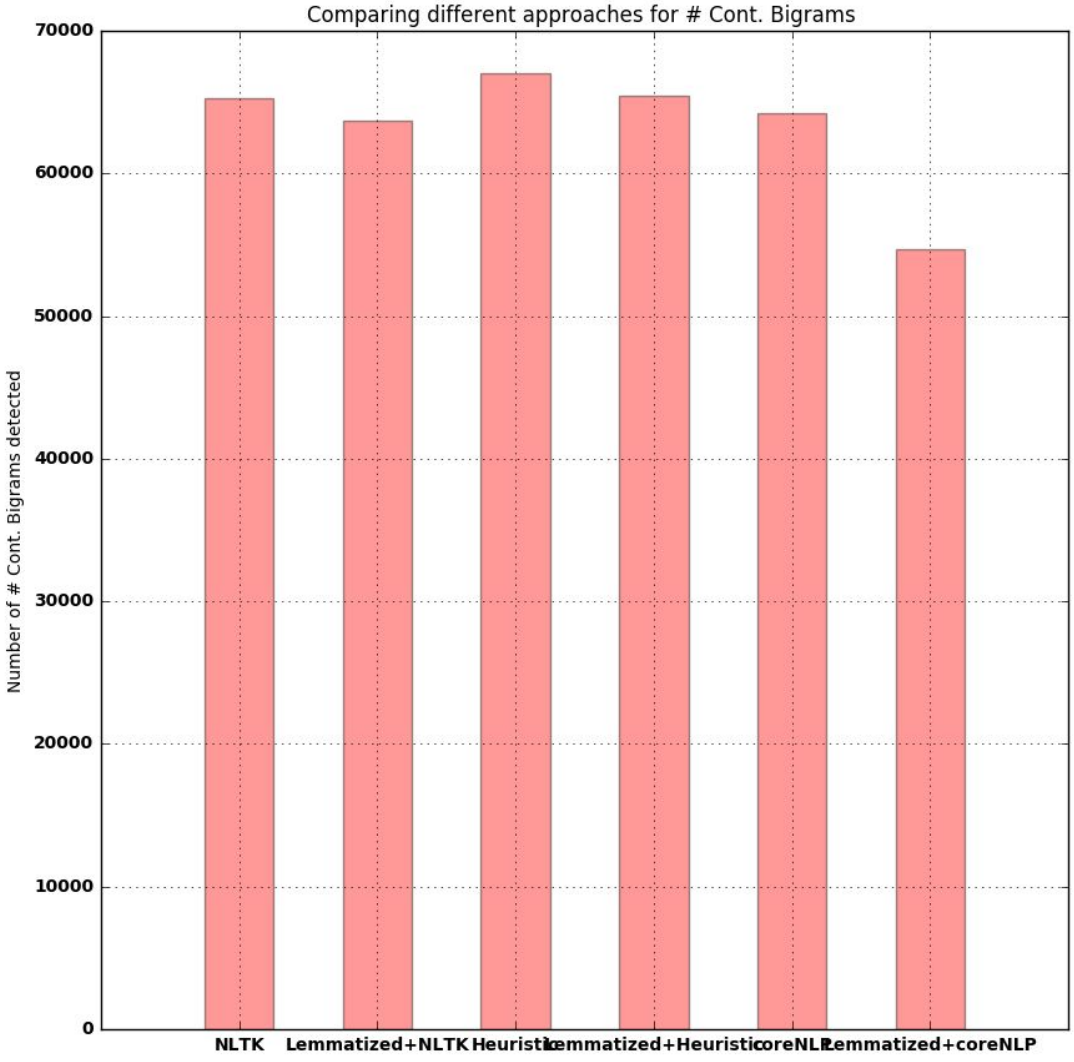
2.5

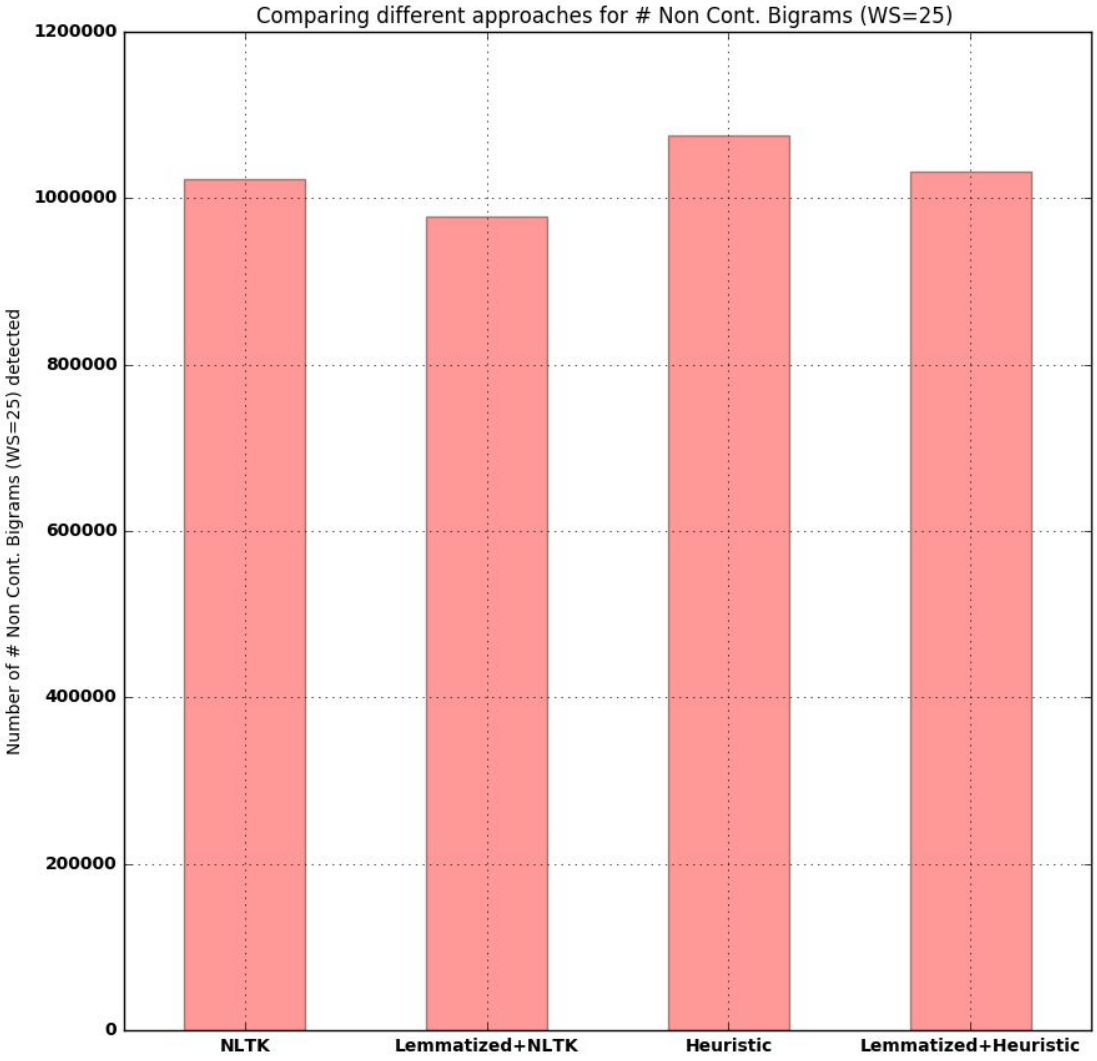
- Same as in NLTK.

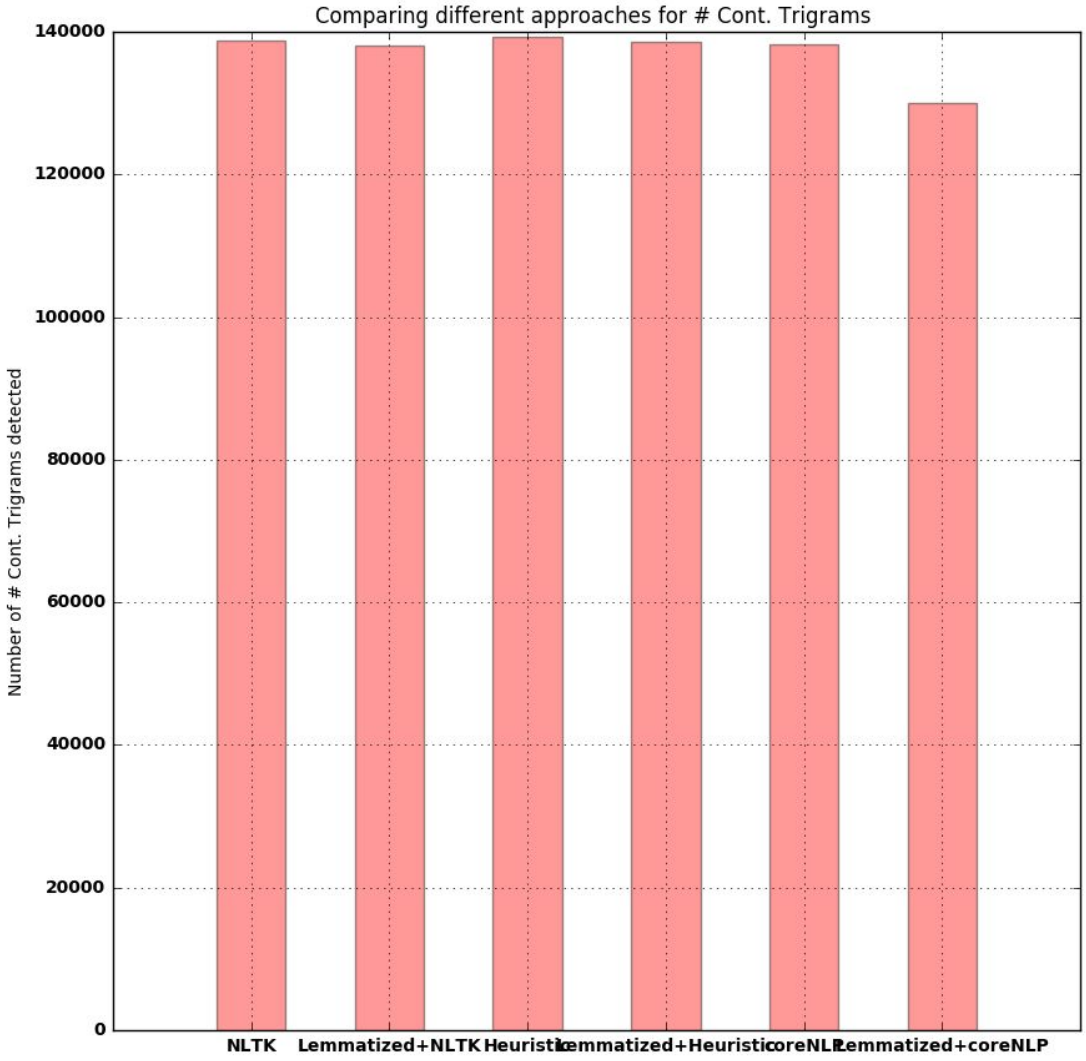
Graphs

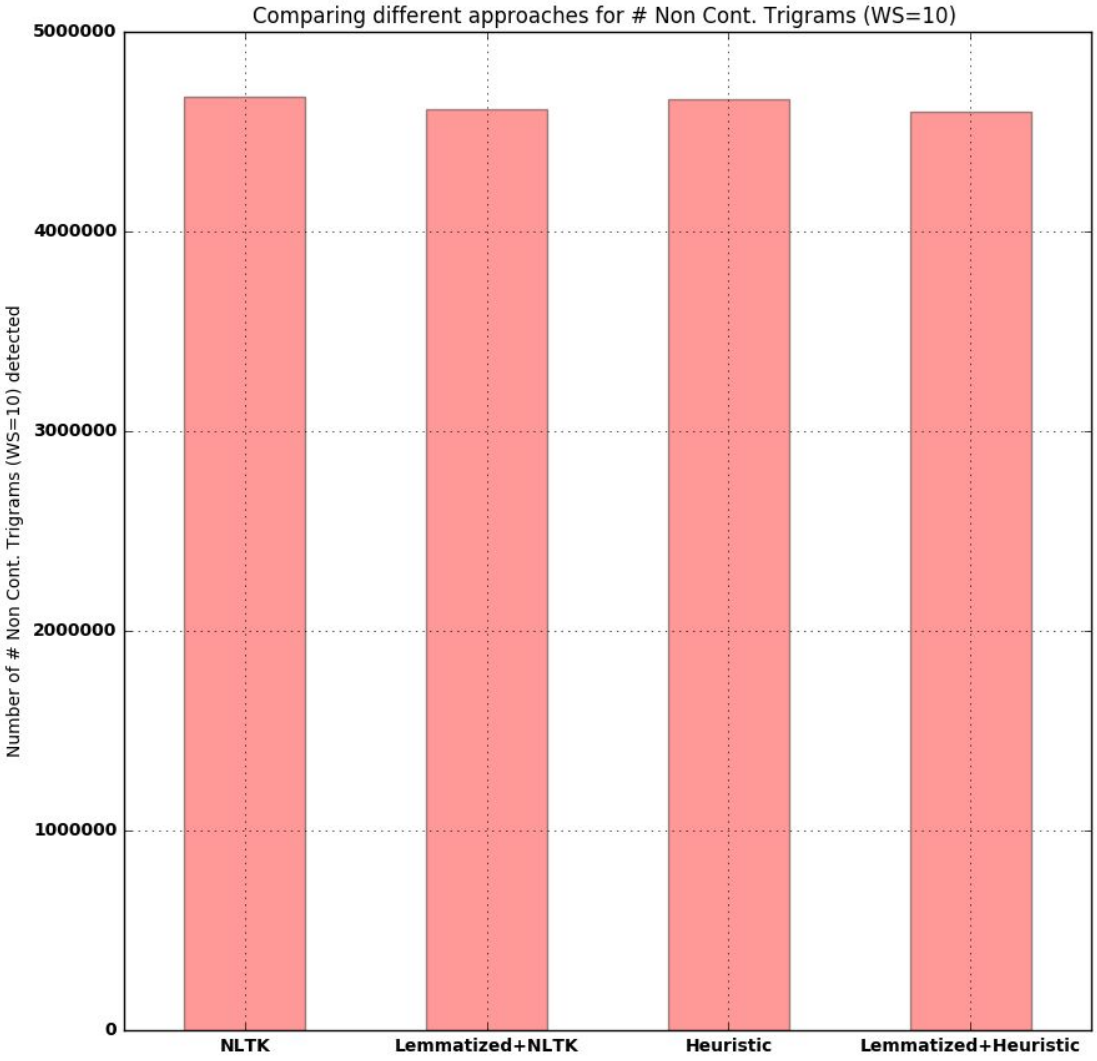

Comparing different approaches for Total Words
Number of Total Words detected
200000
150000
100000
50000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic
coreNLP
Lemmatized+coreNLP

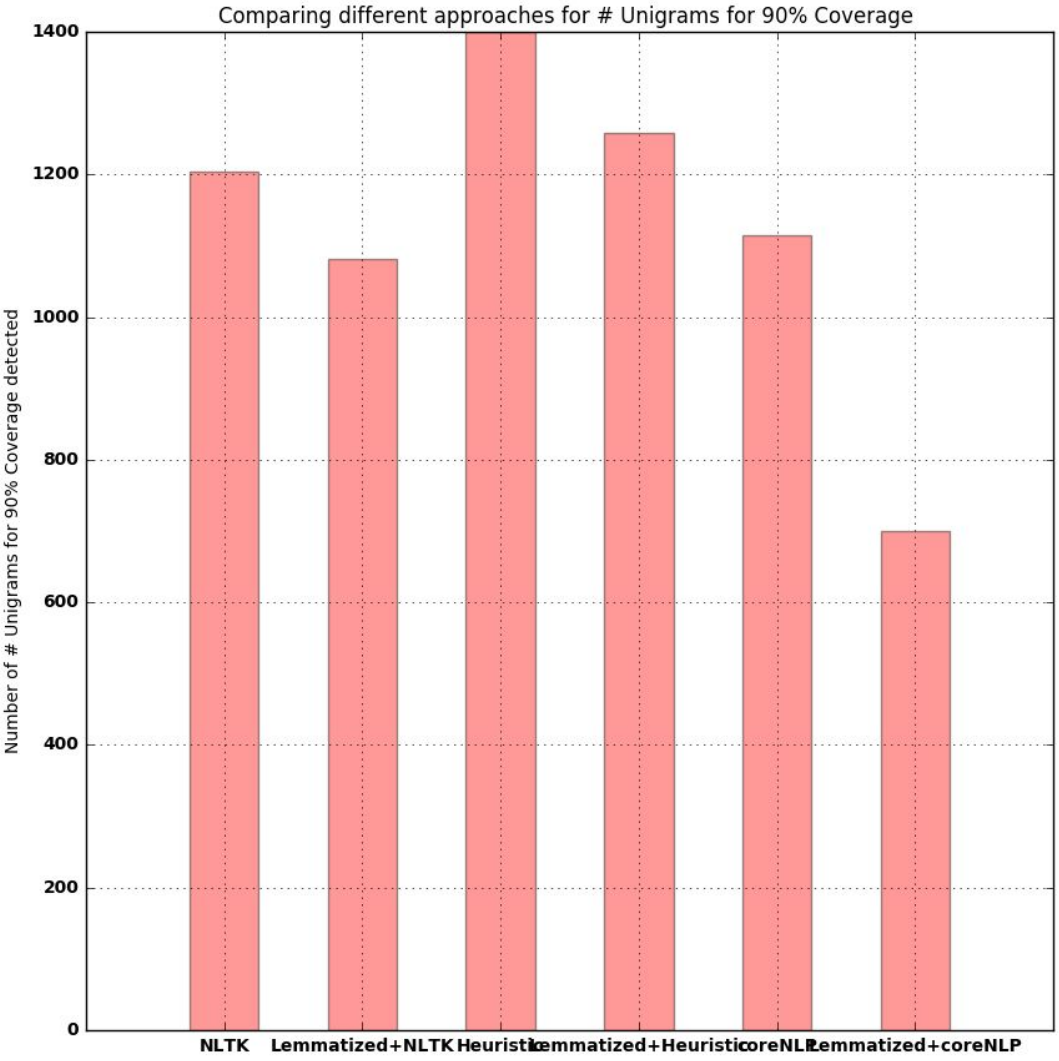

Comparing different approaches for # Unigrams
Number of # Unigrams detected
10000
8000
6000
4000
2000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic
coreNLP
Lemmatized+coreNLP

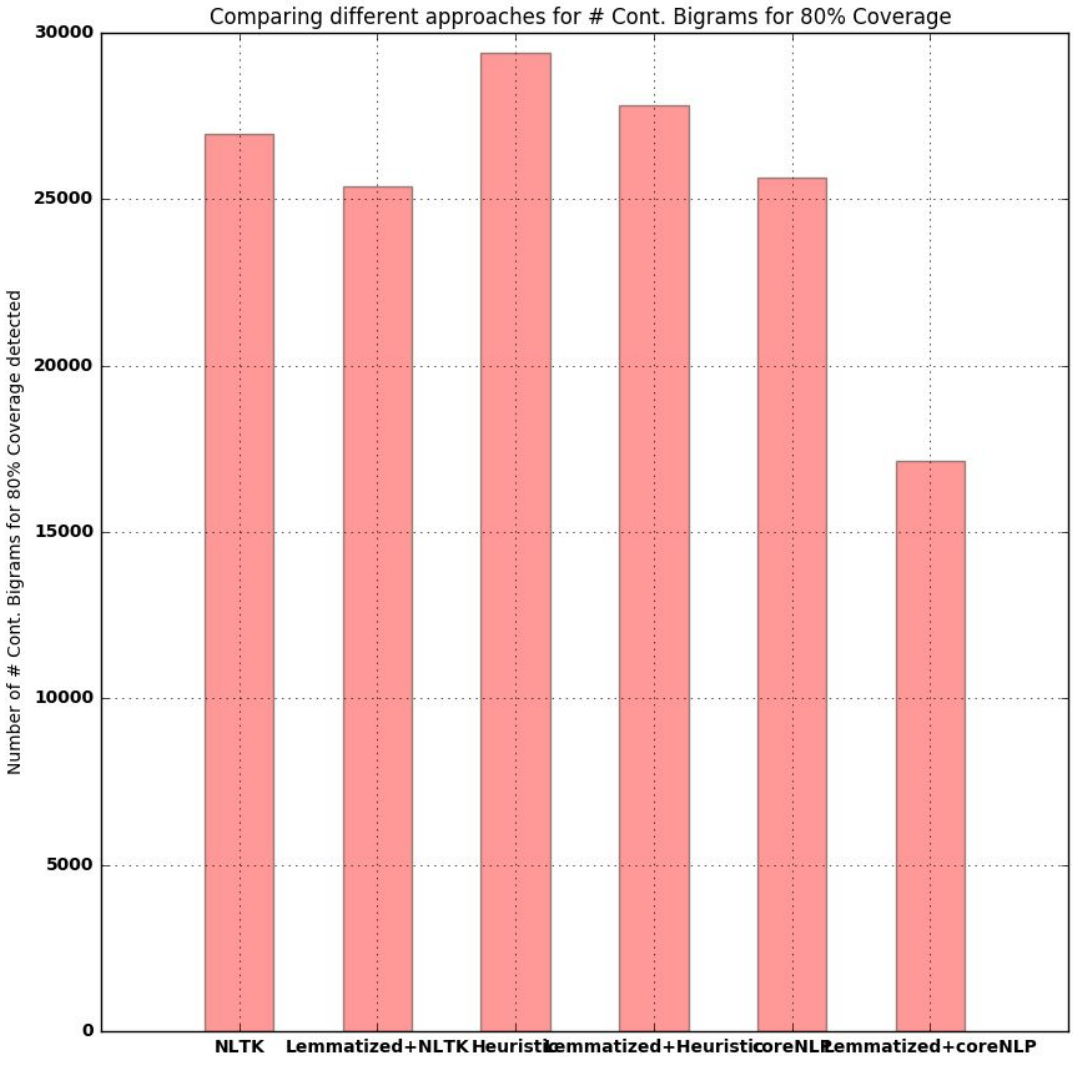

Comparing different approaches for # Cont. Bigrams
Number of # Cont. Bigrams detected
70000
60000
50000
40000
30000
20000
10000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic
coreNLP
Lemmatized+coreNLP

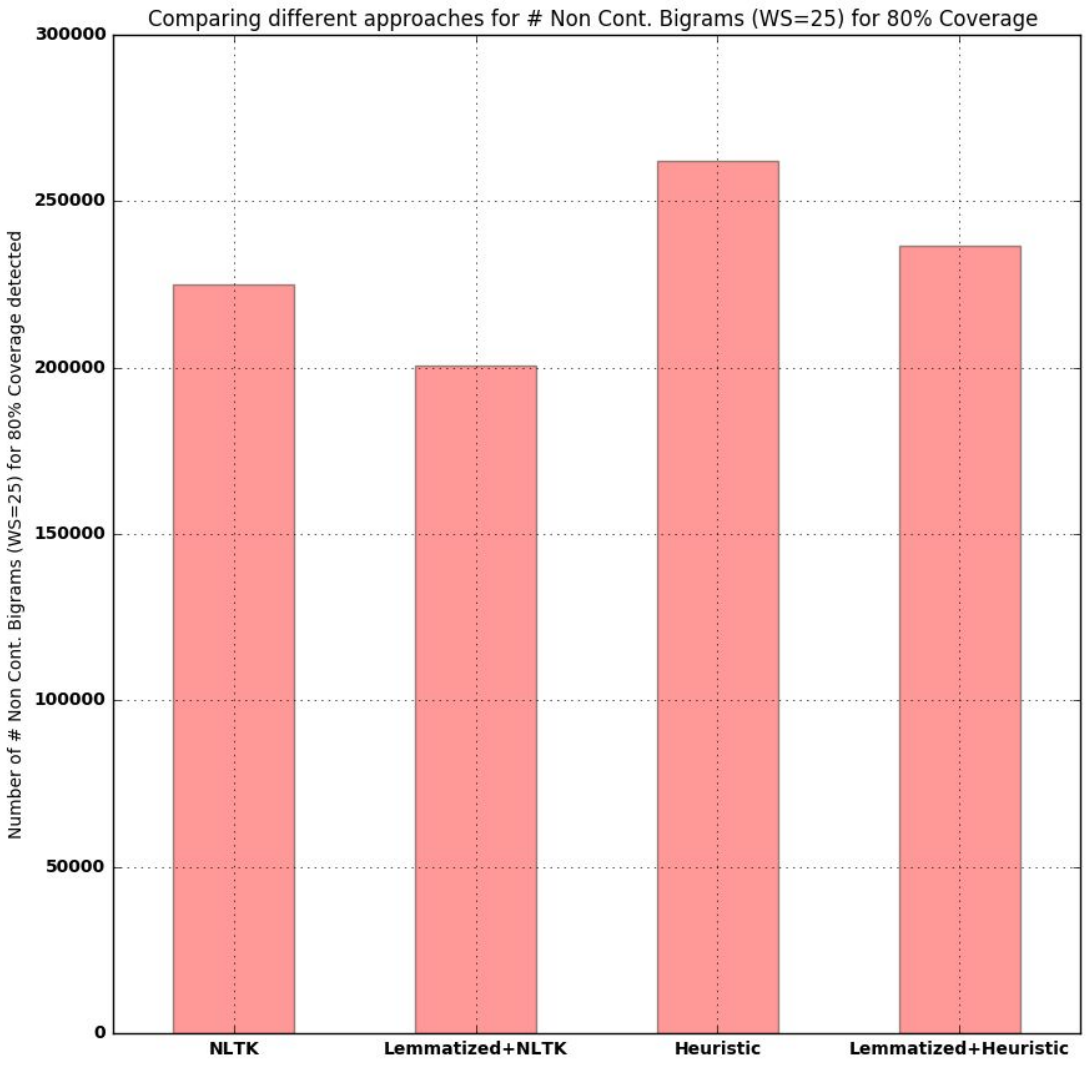

Comparing different approaches for # Non Cont. Bigrams (WS=25)
Number of # Non Cont. Bigrams (WS=25) detected
1200000
1000000
800000
600000
400000
200000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic

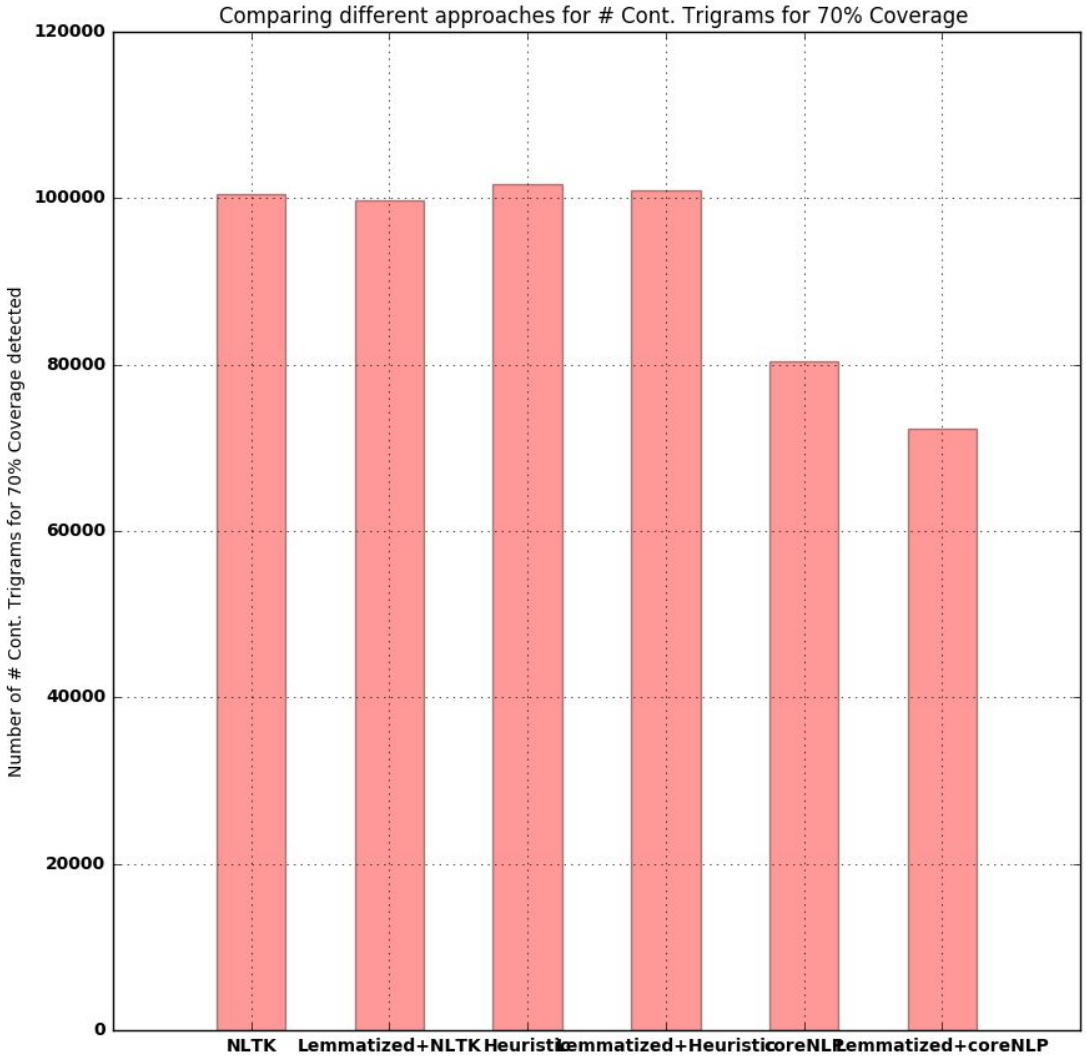

Comparing different approaches for # Cont. Trigrams
Number of # Cont. Trigrams detected
140000
120000
100000
80000
60000
40000
20000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic
coreNLP
Lemmatized+coreNLP

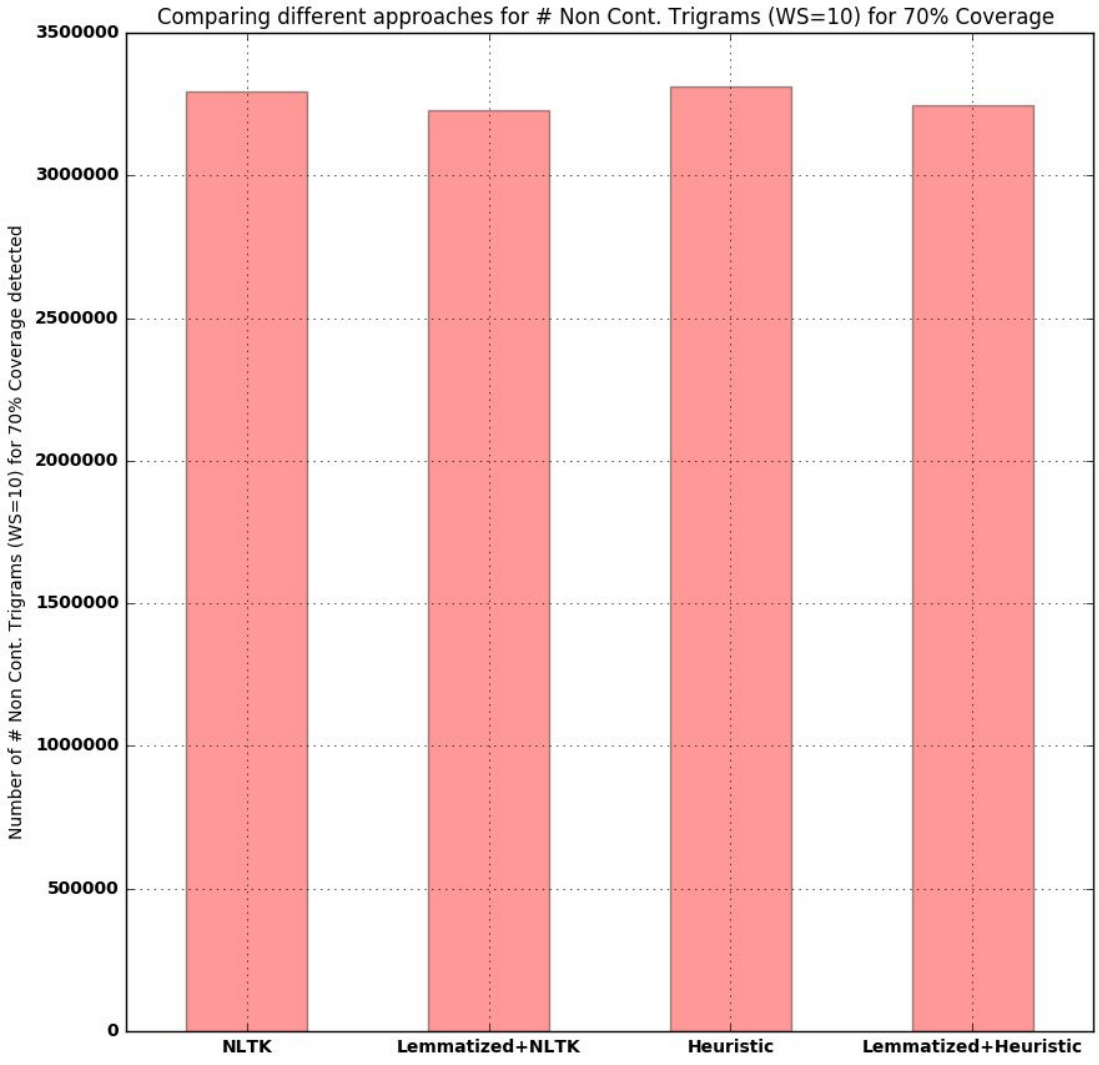

Comparing different approaches for # Non Cont. Trigrams (WS=10)
Number of # Non Cont. Trigrams (WS=10) detected
5000000
4000000
3000000
2000000
1000000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic


Comparing different approaches for # Unigrams for 90% Coverage
Number of # Unigrams for 90% Coverage detected
1400
1200
1000
800
600
400
200
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic
coreNLP
Lemmatized+coreNLP


Comparing different approaches for # Cont. Bigrams for 80% Coverage
Number of # Cont. Bigrams for 80% Coverage detected
30000
25000
20000
15000
10000
5000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic
coreNLP
Lemmatized+coreNLP

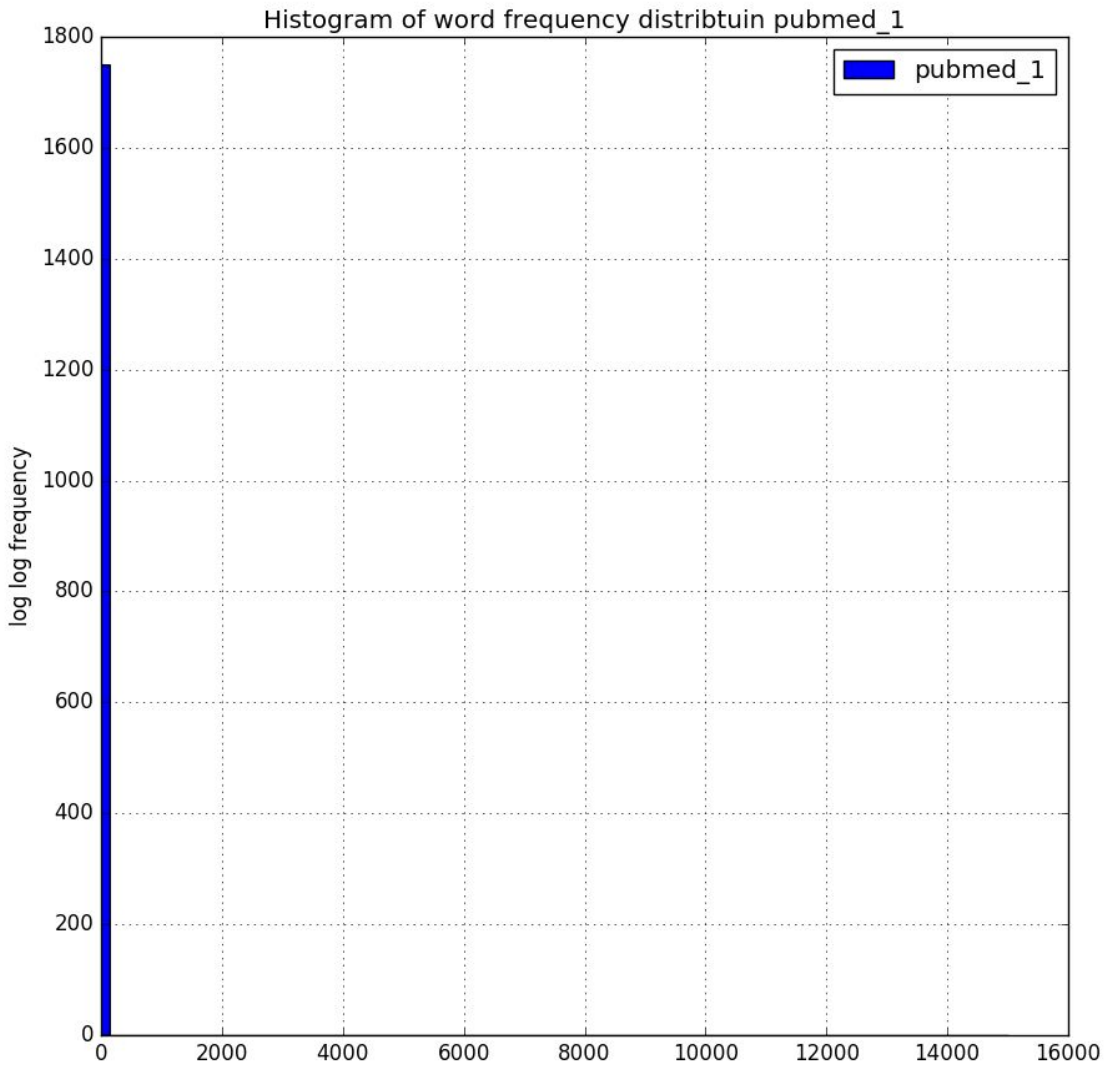

Comparing different approaches for # Non Cont. Bigrams (WS=25) for 80% Coverage
Number of # Non Cont. Bigrams (WS=25) for 80% Coverage detected
300000
250000
200000
150000
100000
50000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic


Comparing different approaches for # Cont. Trigrams for 70% Coverage
Number of # Cont. Trigrams for 70% Coverage detected
120000
100000
80000
60000
40000
20000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic
coreNLP
Lemmatized+coreNLP


Comparing different approaches for # Non Cont. Trigrams (WS=10) for 70% Coverage
Number of # Non Cont. Trigrams (WS=10) for 70% Coverage detected
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
NLTK
Lemmatized+NLTK
Heuristic
Lemmatized+Heuristic

Bonus Problem

Due to large data, only first 1e6 characters of text were used for this part.
Some floating point numbers and then this:



Top 20 collocations:

100x oil-immersion
19th century
2.28e-04 .444
250µm mesh
27-gauge 0.5-inch
3-ketoacyl-coa thiolases
3-β-hydroxyacyl coa
3.11e-04 .412
3utr ccggcctatacgtttctgtggagtactcgagtactccacagaaacgtataggtttttg
3β-hydroxysteroid dehydrogenase-isomerase
4.74e-06 .447
50mm k2hpo4
5µm 5-aza-2
6.73e-06 .495
8.63e-05 .286
8th abdominal
a-coated paramagnetics
aaatcggctcacaagggattc ctcccagcttaaagattttggaaa
abdominal segments
abl800 flex
adenine dinucleotide