Analysis of different approaches

The top 10 passages for each search method are listed in Appendix.B.

BM25 scoring function

The BM25 scoring function does not work well when the keyword searched for is contained in more than half of the corpus documents. In this case, the inverse document frequency function will return a negative value, and a document that does not contain the phrase queried for might be scored higher than a document that contains the query.

However, this behavior can be treated in a number of ways. We implemented one of these fixes, and we fixed it by giving each term in the function a floor of 0.005, to avoid negative contribution to scoring but still take into account some globally popular words.

N-gram

Using n-grams allows us to calculate how often words appear close to each other. Words that often appear close to each other are likely more related to each other than words that do not. However, the size of the collection of n-grams can easily get overwhelming when implementing this. What we did to fix this problem is what was mentioned in the article *Identifying Implicit Relationships*. We used stemming, i.e. reducing words to their word stem, and stop-word removal, i.e. removal of unnecessary words such as “the”, “for”. By doing this, we could reduce our number of n-grams. In implementation, we used 5-gram as proposed in the article.

Query 1 – “adams”

'adams' 'president' 'john' 'quincy' 'secretary'

Query 2 – “Lincoln”

'lincoln' 'president' 'lincoln's' 'abraham' 'in'

Query 3 – “president”

'president' 'vice' 'presidential' 'he' 'became'

Skip bi-gram

Skip bi-grams gives higher credits to passages where the candidate answer appears together with words from the clue. However, one of the major weaknesses is that skip bi-grams do not take into account words that are similar, but not identical. For example, we tried to use search queries like “good president” and “president who is good”. Ther results from those two queries are quite different.

Passage term matching

Passage term matching checks how often a candidate answer appears in the same passage as words from the clue. One of the two major strengths that come from this is that the order of the words and syntactic structure does not matter. The other one is that passage term matching will take into account passages that do not contain the correct answer, but will help extract information from that passage to support candidate answers extracted from other passages.

However, this second strength mentioned can also be a weakness. By using passages that do not contain the correct answer to support other candidate answers, passages that are closely related to the clue but are not actually correct will be assigned too much credit.

Interesting things

In order to refine the search result returned by BM25, we tried to combine n-gram and BM25 together. The proposed method is: for each search query, instead of just search this query with BM25, we first search each term from the query using n-gram to get the similar concepts. We get the intersect of the similar concept sets return for each query term to form a new search query. Then we use the new search query to search.

By searching for terms, using n-grams, that generated sets with intersections, we wanted to use the intersection when using BM25. E.g. searching for ‘civil’ with n-grams generated:

> ‘civil’ ‘war’ ‘service’ ‘president’ ‘right’

And searching for ‘war’ generated:

> ‘war’ ‘civil’ ‘toward’ ‘world’ ‘president’

Then what we will do is to search for {‘civil’, ‘war’, ‘president’} instead of just {‘civil’, ‘war’}. In order to compare the search results, we used BM25 to search for both {‘civil’, ‘war’} and {‘war’, ‘civil’, ‘president’}. We were expecting this would yield ‘lincoln’, or at least the latter query will yield better result. However, ‘president’ is a common word throughout all the documents, which meant that adding this term did not change the outcome of the search by a large margin. And we failed to find other good queries that can well illustrate the benefit of the proposed method. We believe if we can have larger corpus and better stemming and stop-words pruning techniques, we would be able to use this technique to refine the search.

Appendix. A

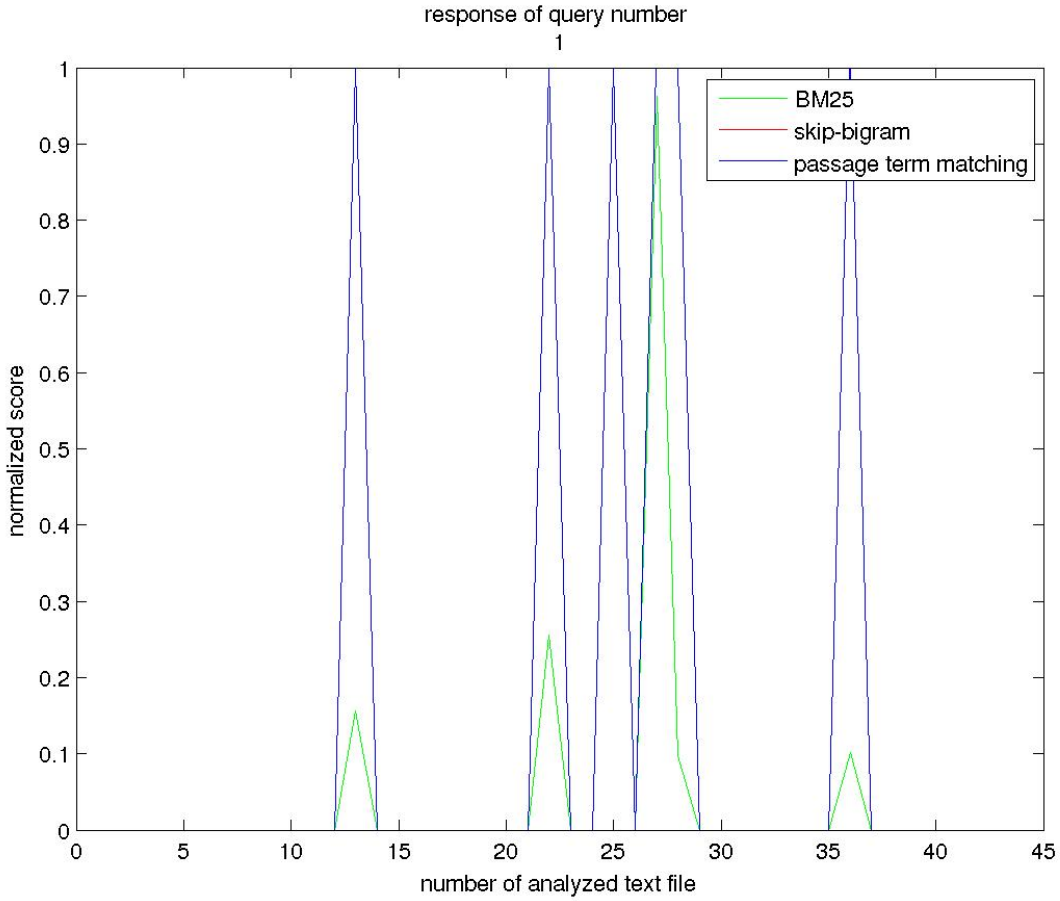
Andreas Palsson – palssoa@rose-hulman.edu
Zhihao Li – liz3@rose-hulman.edu

Comparison of the scores of the different methods

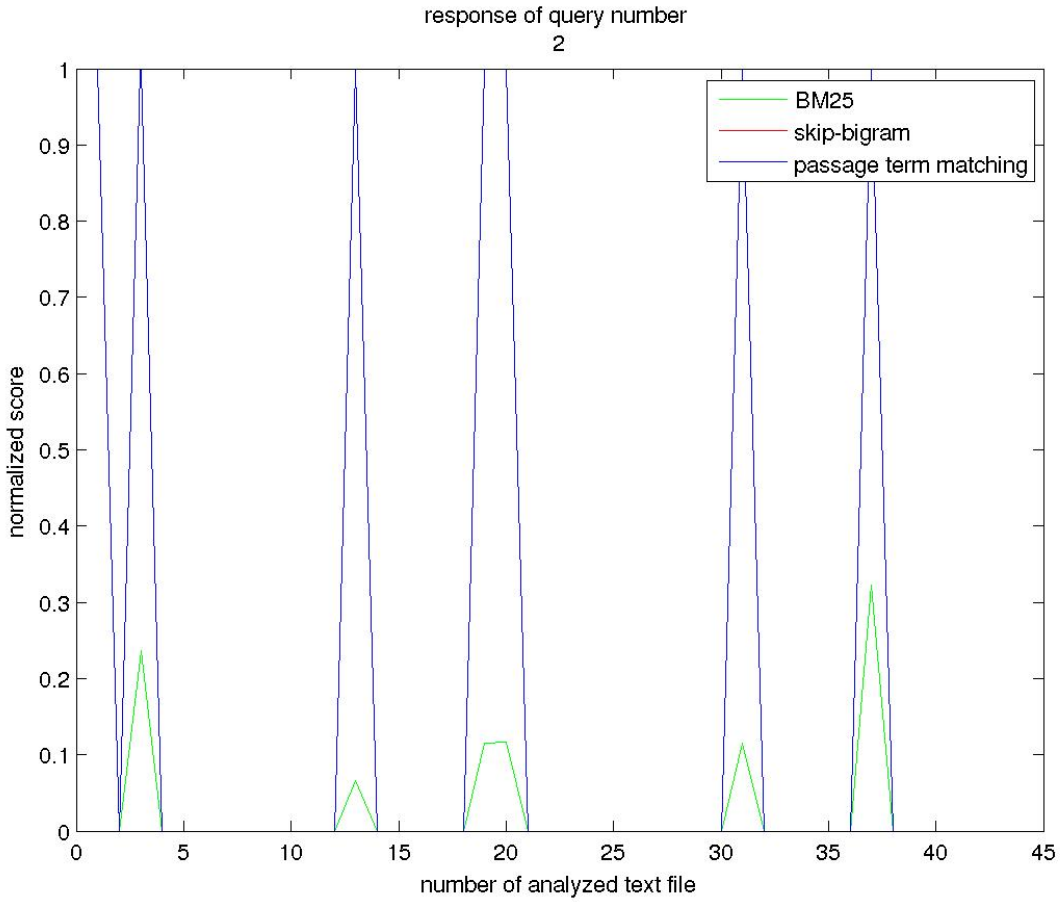
As we can see in the graphs below, the score from BM25 and passage term matching are looking quite similar. This is due to the fact that they are both based on word frequency. However, the skip bigram graphs look a lot different. One of the differences between skip bigram and BM25 is that skip bigram is not based on term frequency, but it is based on how often the answer appears in passages with words from the clue.

You can find some similarities between passage term matching and the skip-bigrams in some of the graphs, since passage term matching is based both on term frequency and on how often the answer appears in passages with words from the clue.

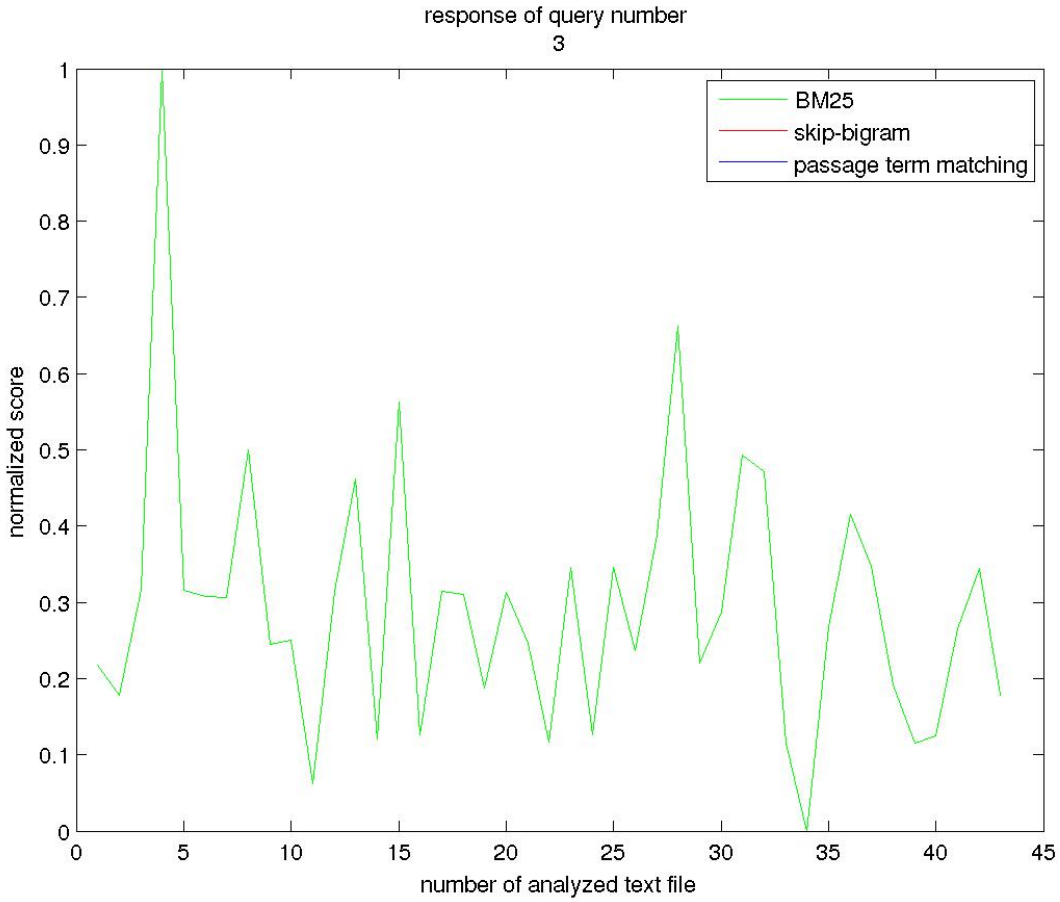
Query 1: “adams”



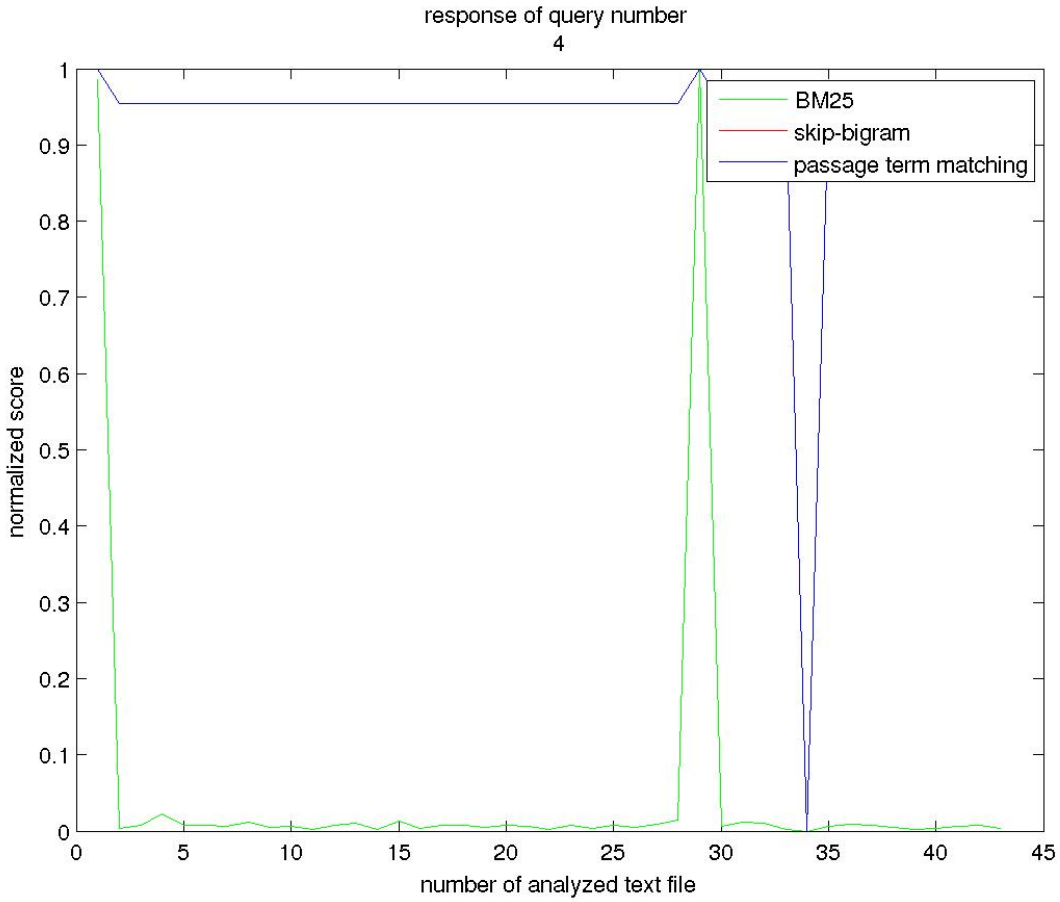
Query 2: “lincoln”



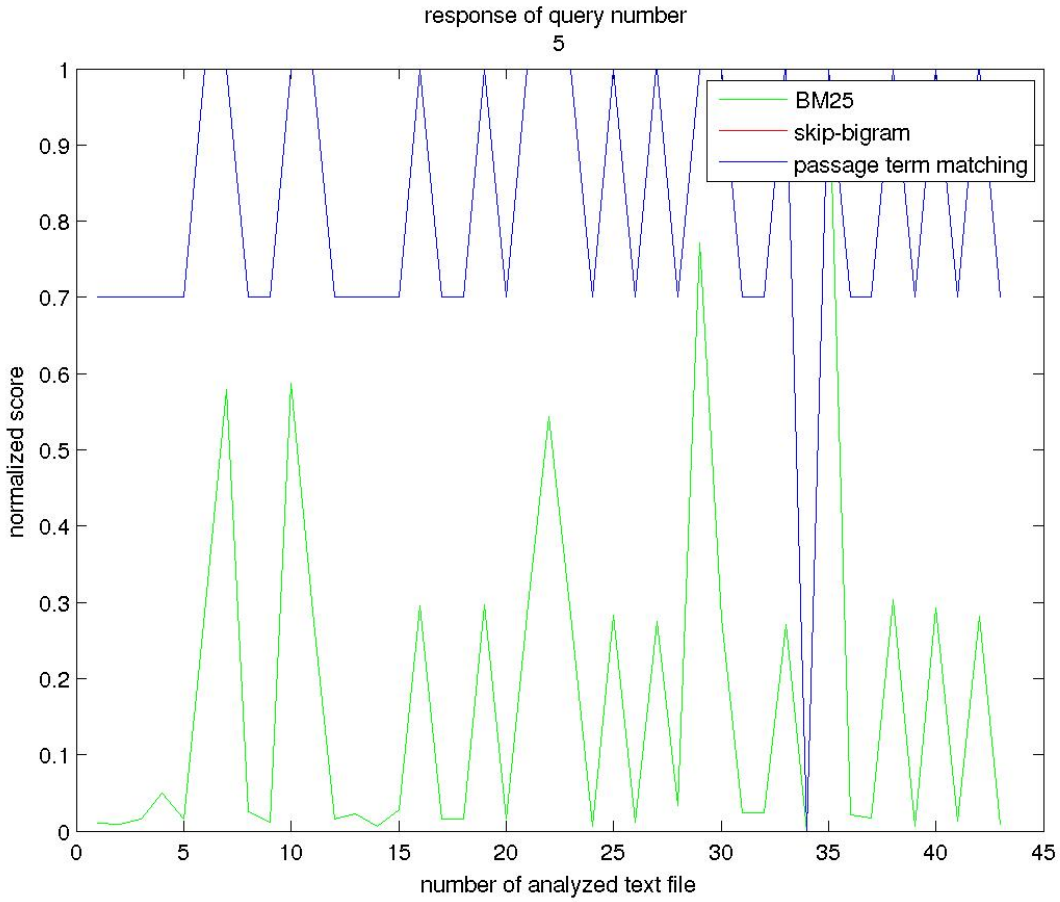
Query 3: “president”



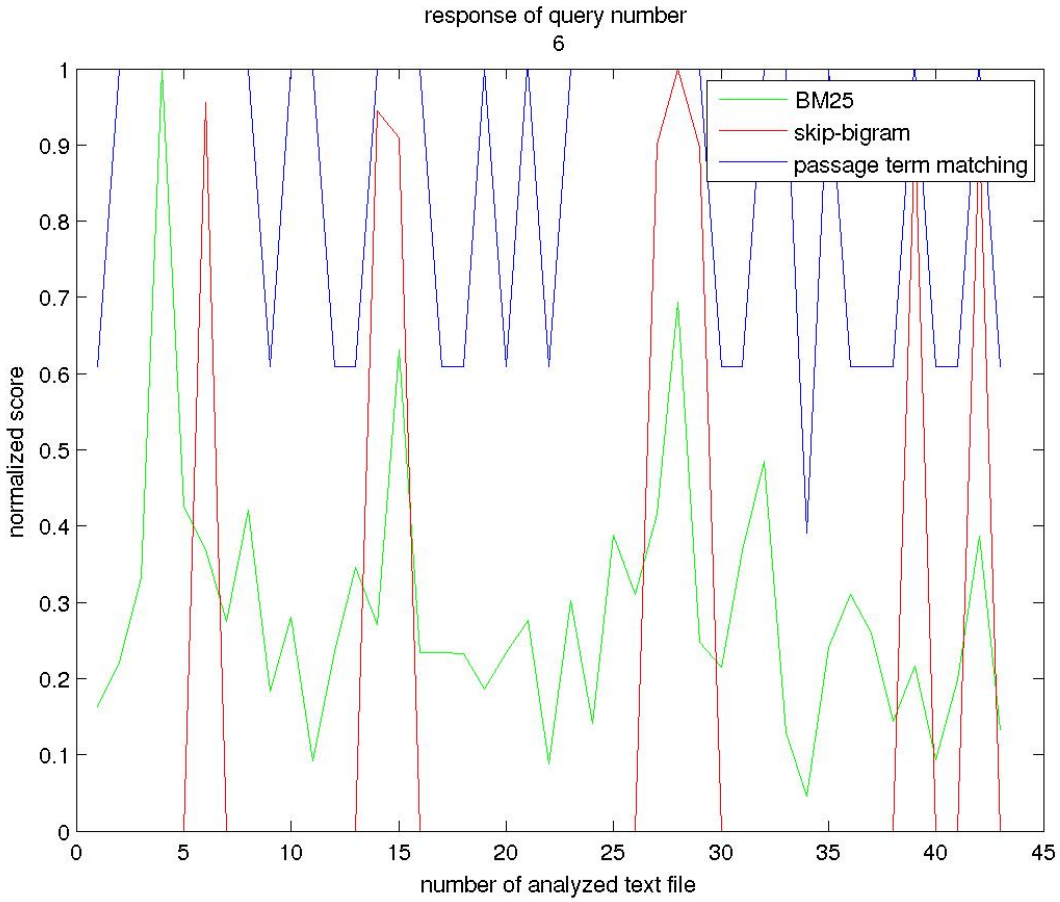
Query 4: “assassinated president”



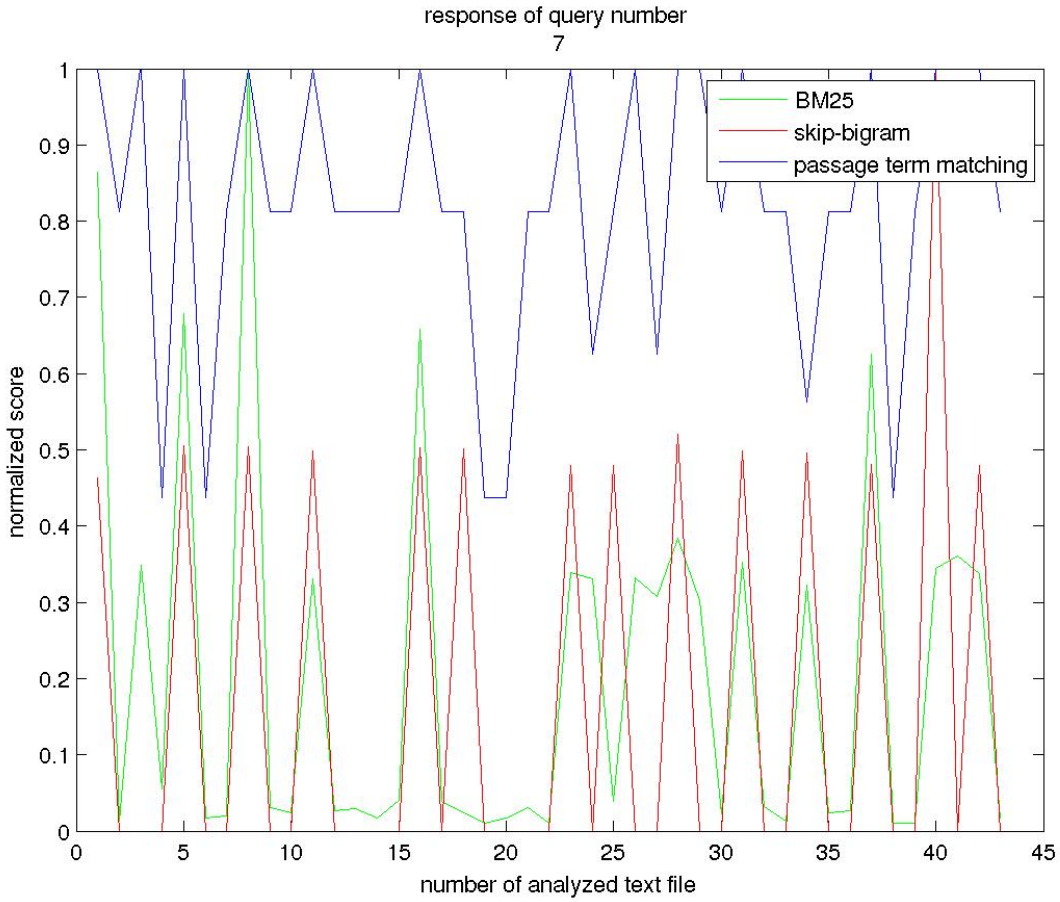
Query 5: “great president”



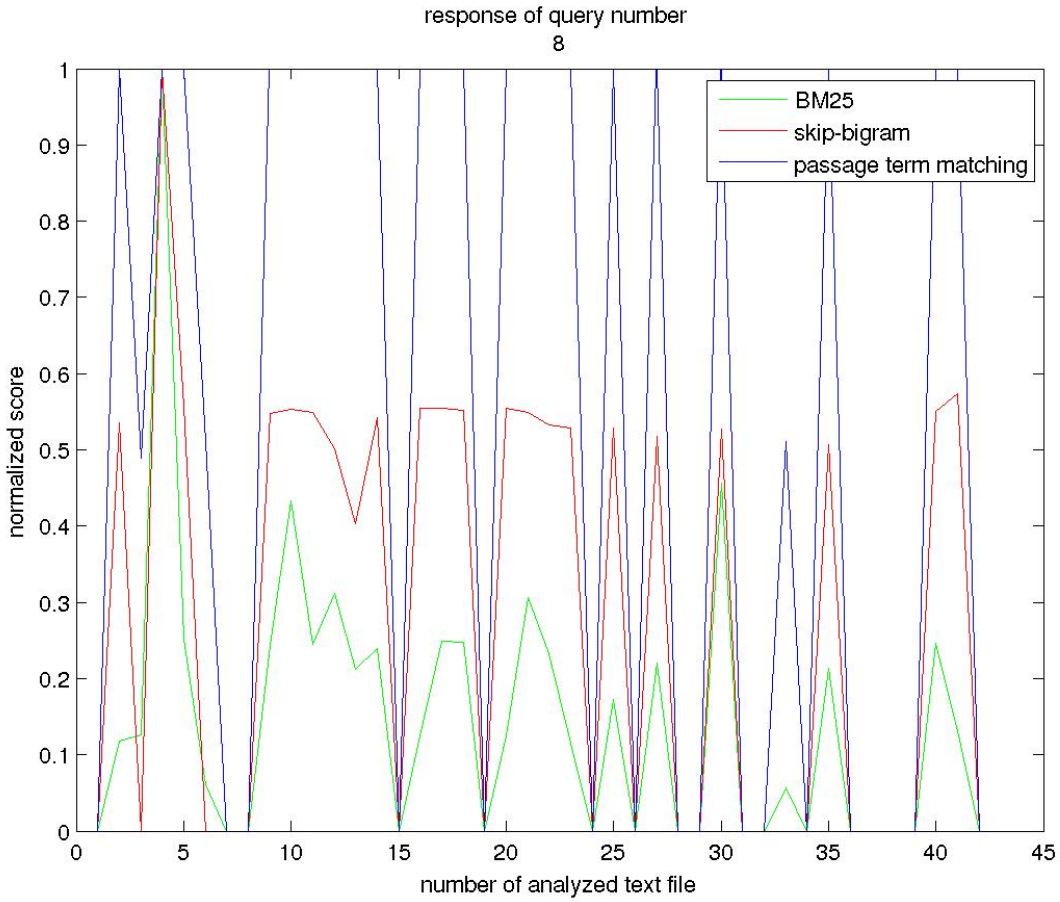
Query 6: “first president”



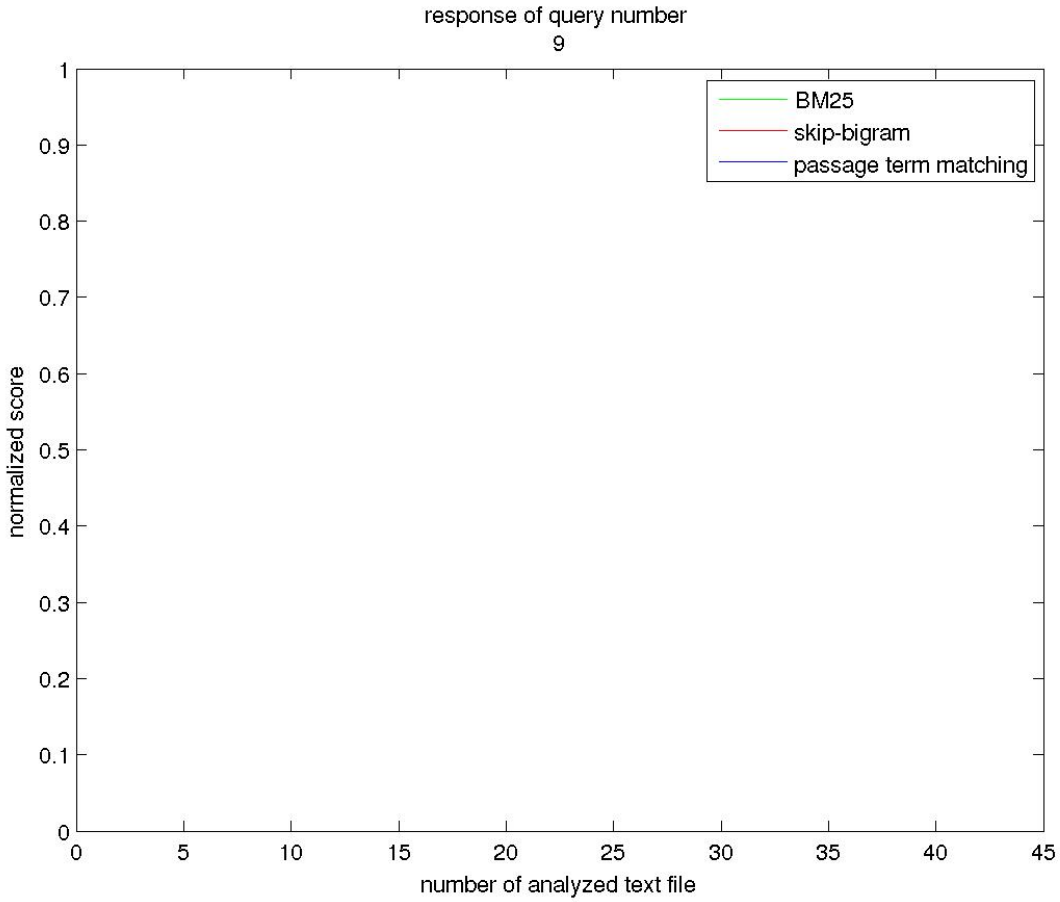
Query 7: “civil war president”



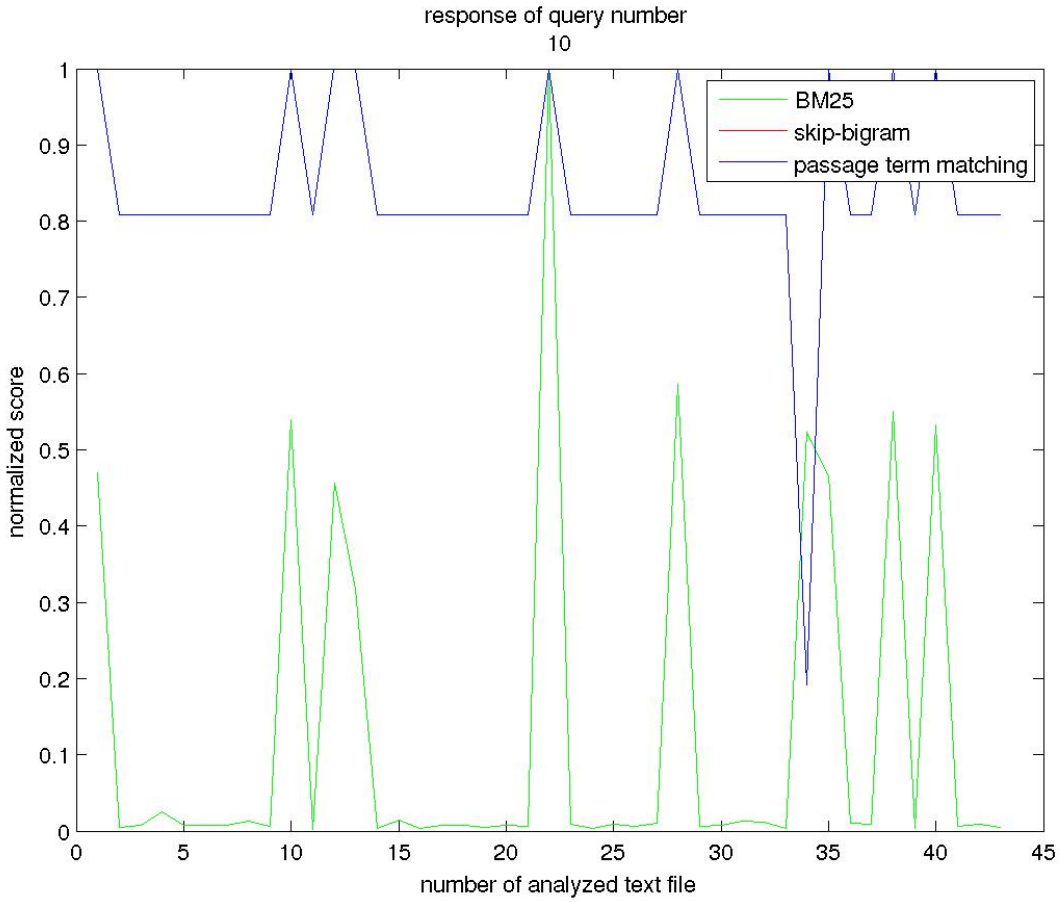
Query 8: “united states”



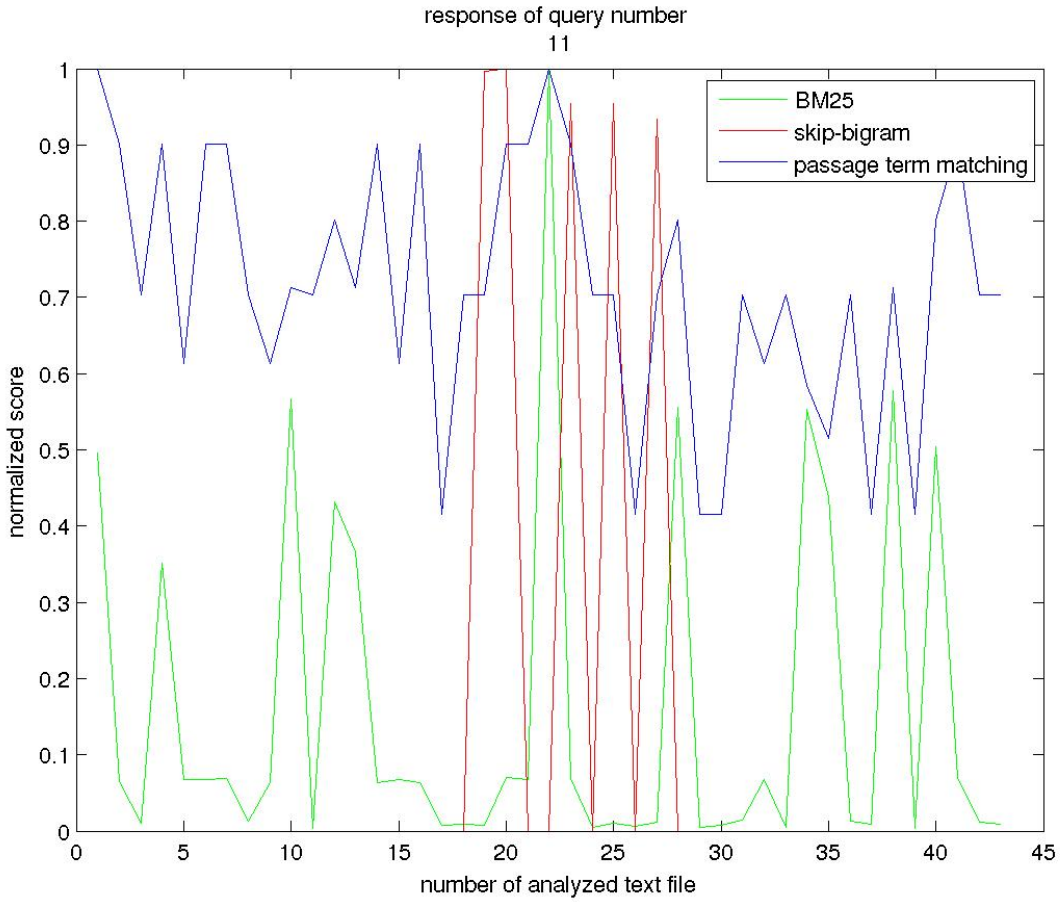
Query 9: “USA”


response of query number
9
BM25
skip-bigram
passage term matching
normalized score
number of analyzed text file
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
45

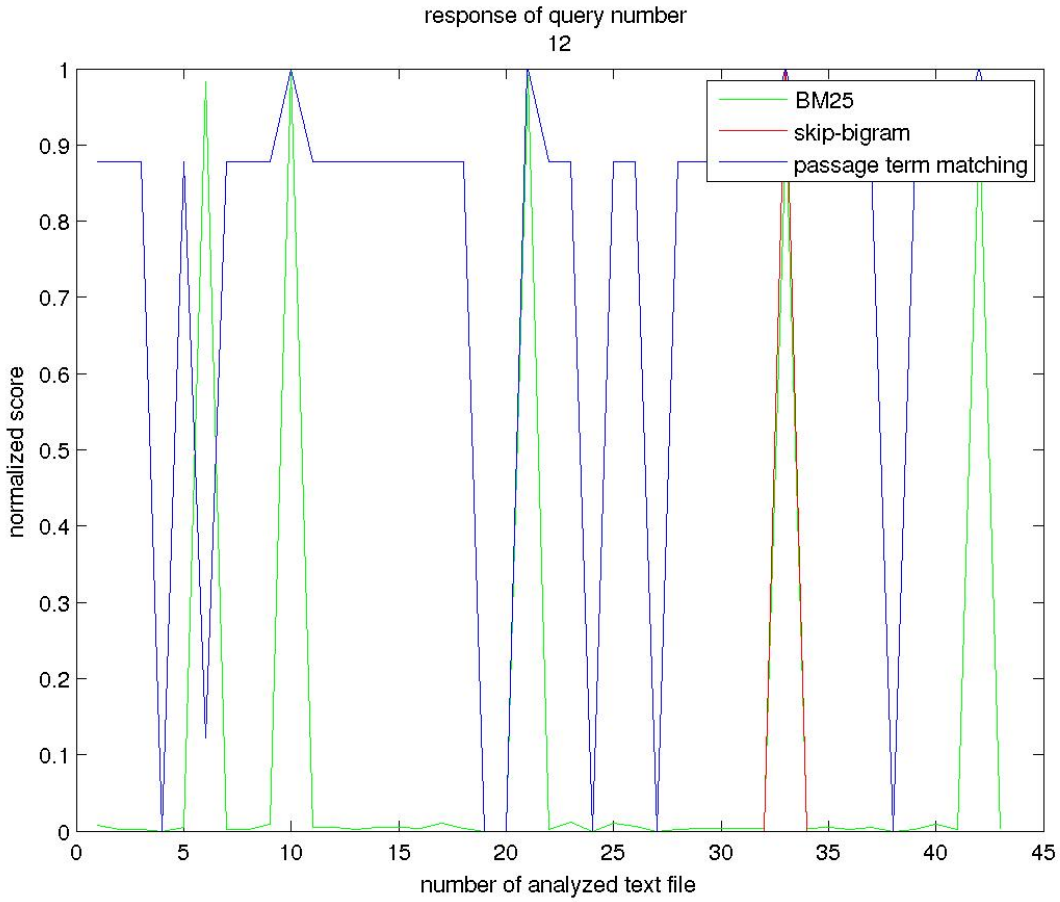
Query 10: “good president”



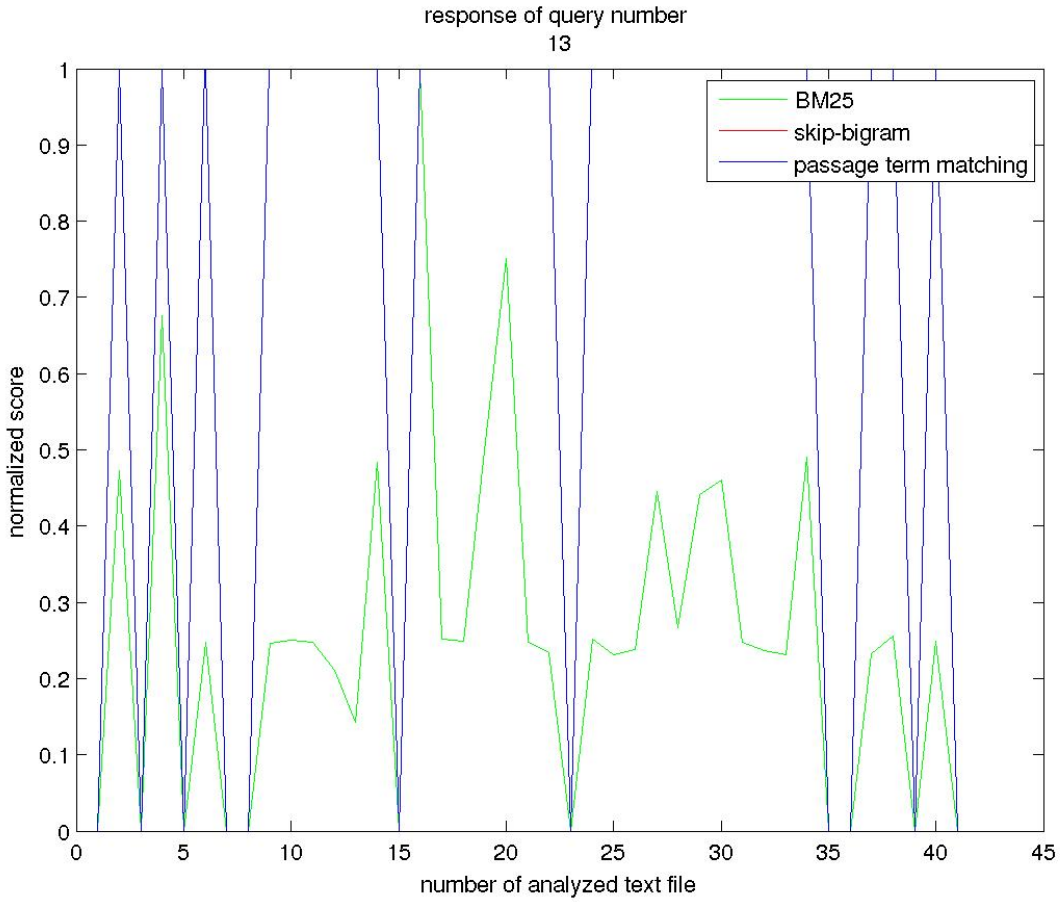
Query 11: “president who is good”



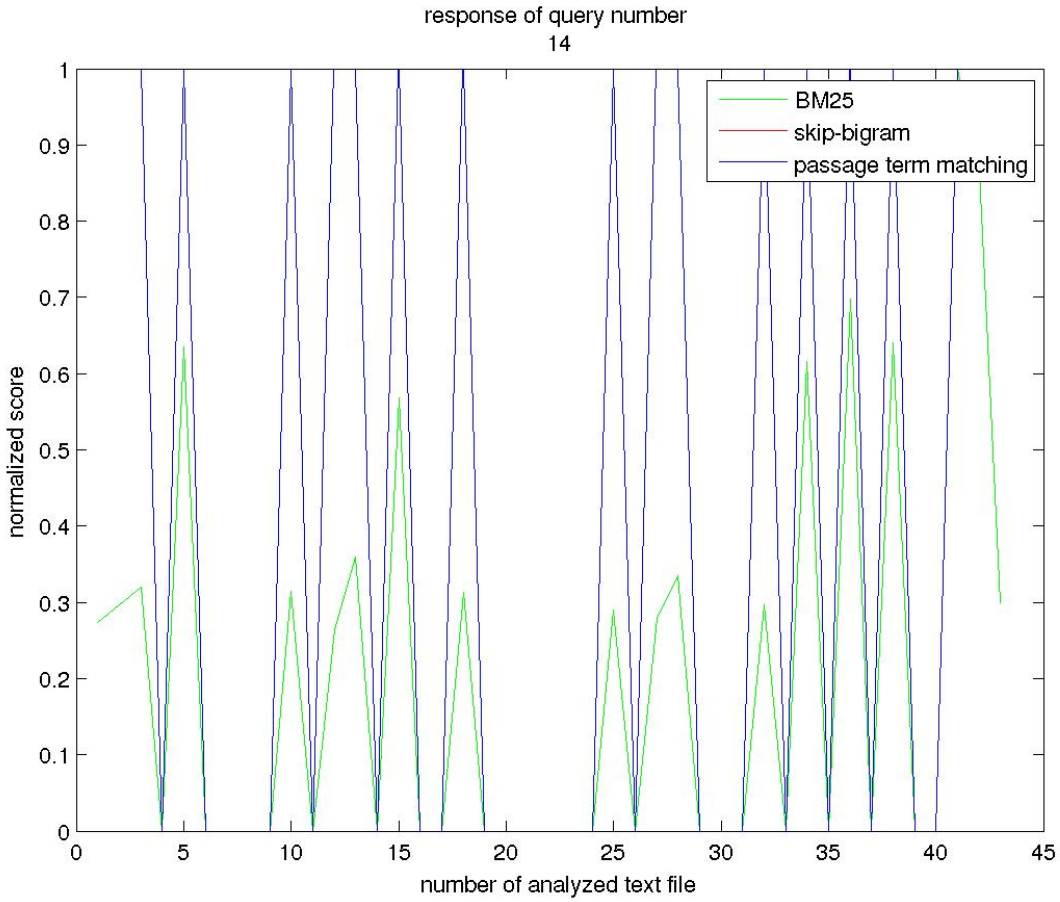
Query 12: “international war”



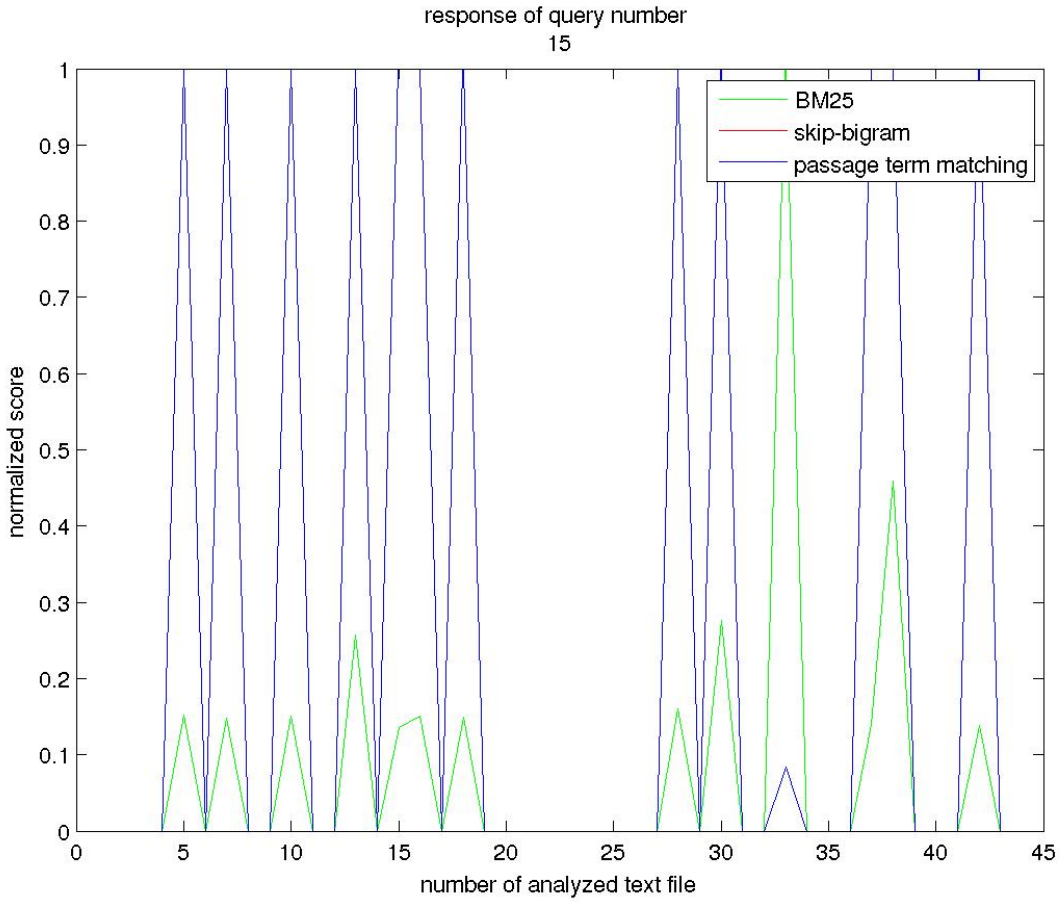
Query 13: “elected 1982”



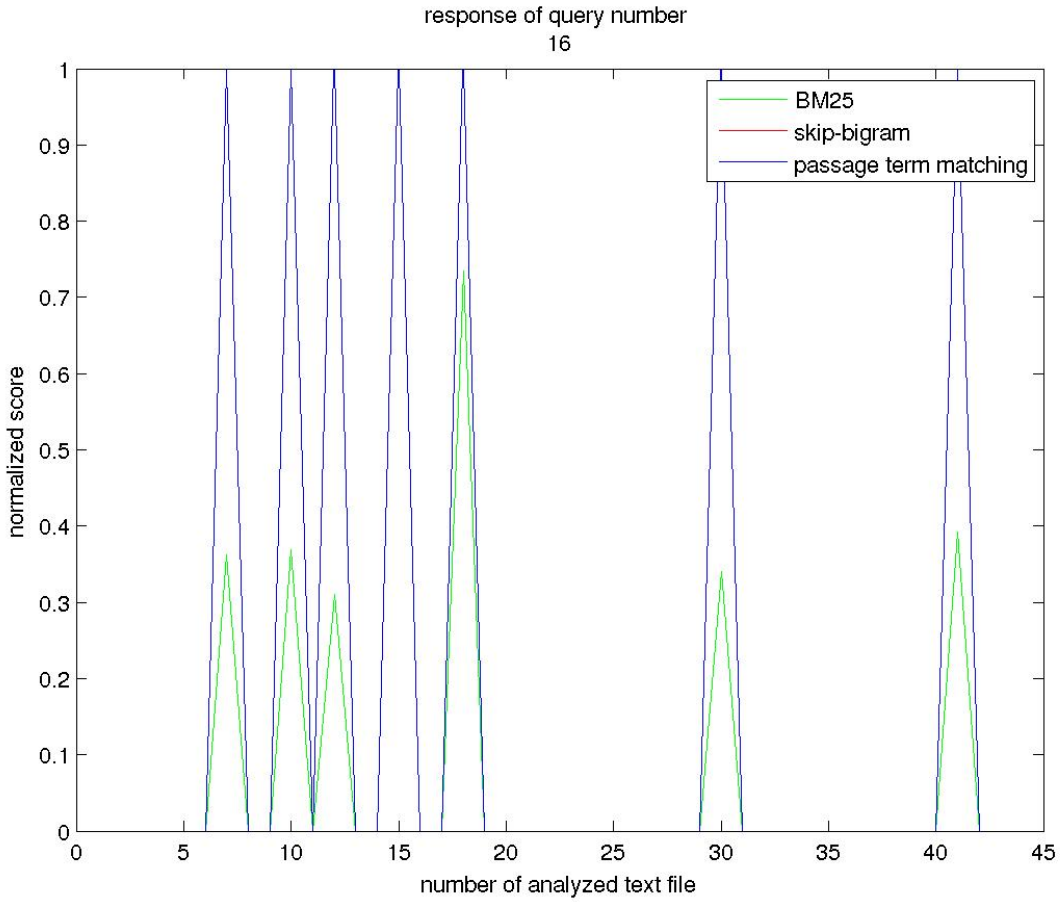
Query 14: “election 1982”



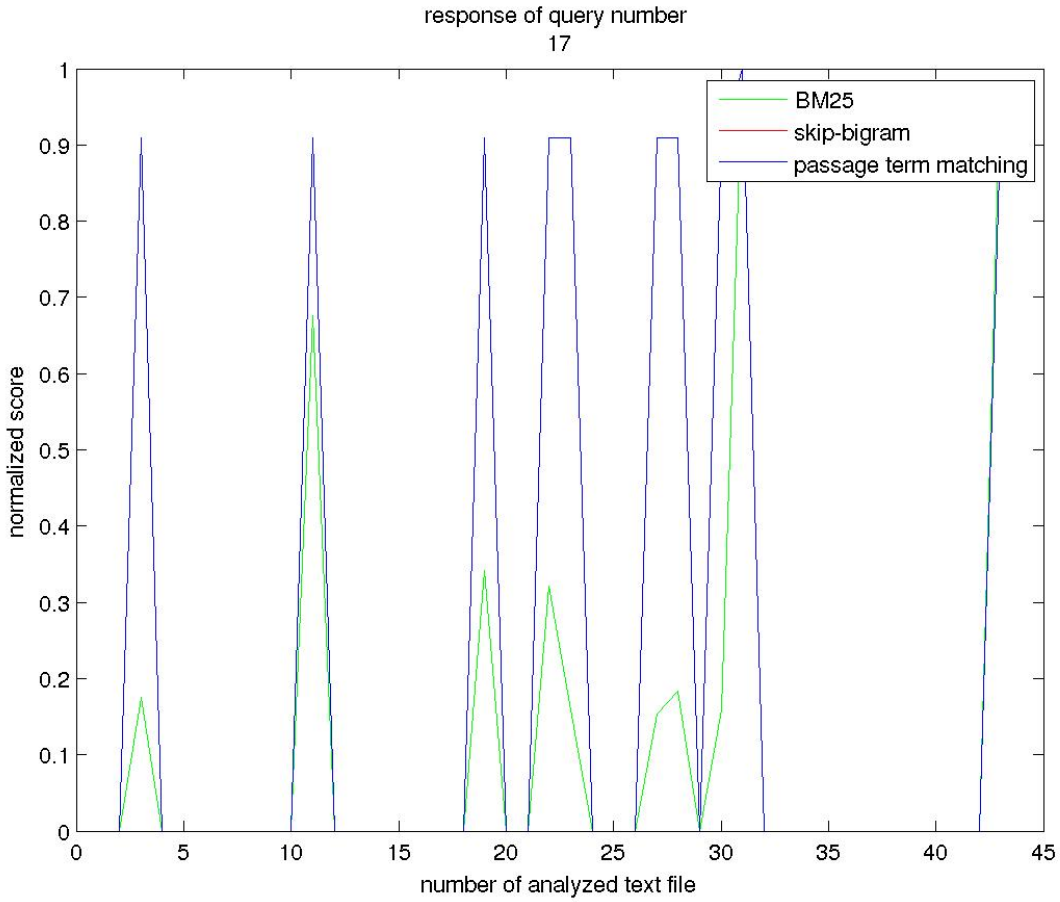
Query 15: “business growth”



Query 16: “economy”



Query 17 “abolish slavery”



Appendix. B

The queries we used in the tables below were
Query 1 – “adams”
Query 2 – “Lincoln”
Query 3 – “president”
Query 4 – “assassinated president”
Query 5 – “great president”
Query 6 – “first president”
Query 7 – “civil war president”
Query 8 – “united states”
Query 9 – “USA”
Query 10 – “good president”
Query 11 – “president who is good”
Query 12 – “international war”
Query 13 – “elected 1982”

Query 14 – “election 1982”
Query 15 – “business growth”
Query 16 – “economy”
Query 17 – “abolish slavery”

The tables shown in the appendix are the top 10 documents that the search approach found. The left column is the score, and the right column is the document number. The document numbers can be found in the table below.

'AbrahamLincoln.txt'	1
'AndrewJackson.txt'	2
'AndrewJohnson.txt'	3
'BarackObama.txt'	4
'BenjaminHarrison.txt'	5
'BillClinton.txt'	6
'CalvinCoolidge.txt'	7
'ChesterArthur.txt'	8
'DwightEisenhower.txt'	9
'FranklinDRoosevelt.txt'	10
'FranklinPierce.txt'	11
'GeorgeBush.txt'	12
'GeorgeWBush.txt'	13
'GeorgeWashington.txt'	14
'GeraldFord.txt'	15
'GroverCleveland.txt'	16
'HarryTruman.txt'	17
'HerbertHoover.txt'	18
'JamesBuchanan.txt'	19
'JamesGarfield.txt'	20
'JamesMadison.txt'	21
'JamesMonroe.txt'	22
'JamesPolk.txt'	23
'JimmyCarter.txt'	24
'JohnAdams.txt'	25
'JohnKennedy.txt'	26
'JohnQuincyAdams.txt'	27
'JohnTyler.txt'	28
'LyndonJohnson.txt'	29
'MartinVanBuren.txt'	30
'MillardFillmore.txt'	31
'RichardNixon.txt'	32
'RonaldReagan.txt'	33
'RutherfordHayes.txt'	34
'TheodoreRoosevelt.txt'	35

'ThomasJefferson.txt'	36
'UlyssesGrant.txt'	37
'WarrenHarding.txt'	38
'WilliamHenryHarrison.txt'	39
'WilliamMcKinley.txt'	40
'WilliamTaft.txt'	41
'WoodrowWilson.txt'	42
'ZacharyTaylor.txt'	43

BM25 SCORING FUNCTION

Query 1

0.0824992124008572	25
0.0794538704333526	27
0.0210766456939933	22
0.0128955579173646	13
0.00837981554156189	36
0.00803673233826394	28
0	43
0	42
0	41
0	40

Query 2

0.0585564324557654	1
0.0188623212369058	37
0.0138279146965538	3
0.00681608987369571	20
0.00677427698620173	19
0.00673285018208420	31
0.00388652209760521	13
0	43
0	42
0	41

Query 3

0.000237474038231101	4
0.000157132466136599	28
0.000133502390484170	15
0.000118799497161705	8
0.000117002194757061	31
0.000111849253852882	32
0.000109760650916130	13

9.87743716225431e-05	36
9.21383156070192e-05	27
8.23509935180430e-05	37

Query 4

0.0106973229229128	29
0.0105445921358247	1
0.000237474038231101	4
0.000157132466136599	28
0.000133502390484170	15
0.000118799497161705	8
0.000117002194757061	31
0.000111849253852882	32
0.000109760650916130	13
9.87743716225431e-05	36

Query 5

0.00478902472828646	35
0.00369251791574913	29
0.00281210535517798	10
0.00277544654369421	7
0.00260357976228817	22
0.00145371433904064	38
0.00143439052640571	6
0.00141982695826946	21
0.00141820743064821	19
0.00141210312915129	16

Query 6

0.000317647470101474	4
0.000220463115041372	28
0.000200611957229650	15
0.000154028610526309	32
0.000135027312772853	5
0.000133773961424647	8
0.000131793192974886	27
0.000123298870364250	25
0.000122936778132214	42
0.000117160292952778	6

Query 7

0.00428348599578618	8
0.00370436175532328	1
0.00290700964924122	5
0.00282090856505984	16

0.00267662212072603	37
0.00164463317579907	28
0.00154590989633296	41
0.00151186081770390	31
0.00149443028537228	3
0.00147274795254648	40

Query 8

0.000239751427820550	4
0.000109396419835634	30
0.000103859290046850	10
7.48882202275899e-05	12
7.33669417737936e-05	21
6.01967583359599e-05	5
5.97904817154892e-05	17
5.92767548421126e-05	18
5.90953020204664e-05	40
5.87349023115662e-05	11

Query 9

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 10

0.0093	22
0.0055	28
0.0051	38
0.0050	10
0.0050	40
0.0049	34
0.0044	1
0.0043	35
0.0043	12

0.0030	13

Query 11

0.0099	22
0.0057	38
0.0056	10
0.0055	28
0.0055	34
0.0050	40
0.0049	1
0.0043	35
0.0043	12
0.0036	13

Query 12

0.0084	10
0.0083	21
0.0083	6
0.0078	42
0.0077	33
0.0001	23
0.0001	17
0.0001	25
0.0001	40
0.0001	9

Query 13

0.0001	16
0	20
0	4
0	19
0	34
0	14
0	2
0	30

0	27
0	29

Query 14

0.0031	41
0.0027	42
0.0022	36
0.0020	38
0.0020	5
0.0019	34
0.0018	15
0.0011	13
0.0010	28
0.0010	3

Query 15

0.0264	33
0.0121	38
0.0073	30
0.0068	13
0.0042	28
0.0040	5
0.0040	16
0.0040	10
0.0040	18
0.0039	7

Query 16

0.0184	15
0.0135	18
0.0072	41
0.0068	10
0.0067	7
0.0063	30
0.0057	12
0	43
0	42
0	40

Query 17

0.0290	31
0.0283	43
0.0196	11
0.0099	19
0.0093	22
0.0053	28
0.0051	3
0.0046	23
0.0046	30
0.0044	27

PASSAGE TERM MATCHING

Query 1

3.30525707202109	36
3.30525707202109	28
3.30525707202109	27
3.30525707202109	25
3.30525707202109	22
3.30525707202109	13
0	43
0	42
0	41
0	40

Query 2

4.16230968473294	37
4.16230968473294	31
4.16230968473294	20
4.16230968473294	19
4.16230968473294	13
4.16230968473294	3
4.16230968473294	1
0	43
0	42
0	41

Query 3

0	34
-1806006467323.23	43

-1806006467323.23	42
-1806006467323.23	41
-1806006467323.23	40
-1806006467323.23	39
-1806006467323.23	38
-1806006467323.23	37
-1806006467323.23	36
-1806006467323.23	35

Query 4

16.5252764828192	29
16.5252764828192	1
15.7741275517990	43
15.7741275517990	42
15.7741275517990	41
15.7741275517990	40
15.7741275517990	39
15.7741275517990	38
15.7741275517990	37
15.7741275517990	36

Query 5

73.4608899951204	42
73.4608899951204	40
73.4608899951204	38
73.4608899951204	35
73.4608899951204	33
73.4608899951204	30
73.4608899951204	29
73.4608899951204	27
73.4608899951204	25
73.4608899951204	23

Query 6

160.840808430583	42
160.840808430583	39
160.840808430583	35
160.840808430583	33
160.840808430583	32
160.840808430583	29
160.840808430583	28
160.840808430583	27
160.840808430583	26
160.840808430583	25

Query 7

99.2716624731154	42
99.2716624731154	41
99.2716624731154	40
99.2716624731154	37
99.2716624731154	31
99.2716624731154	29
99.2716624731154	28
99.2716624731154	26
99.2716624731154	23
99.2716624731154	16

Query 8

41.7603098606995	41
41.7603098606995	40
41.7603098606995	35
41.7603098606995	30
41.7603098606995	27
41.7603098606995	25
41.7603098606995	23
41.7603098606995	22
41.7603098606995	21
41.7603098606995	20

Query 9, USA

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 10, good president

38.1426	40
38.1426	38
38.1426	35
38.1426	28

38.1426	22
38.1426	13
38.1426	12
38.1426	10
38.1426	1
30.8075	43

Query 11, president who is good

41.3937	22
41.3937	1
37.2953	41
37.2953	23
37.2953	21
37.2953	20
37.2953	16
37.2953	14
37.2953	7
37.2953	6

Query 12, international war

19.3422	42
19.3422	33
19.3422	21
19.3422	10
16.9834	43
16.9834	41
16.9834	40
16.9834	39
16.9834	37
16.9834	36

Query 13, elected 1982

7.3378	40
7.3378	38
7.3378	37
7.3378	34
7.3378	33
7.3378	32
7.3378	31
7.3378	30

7.3378	29
7.3378	28

Query 14, election 1982

4.1973	43
4.1973	42
4.1973	41
4.1973	38
4.1973	36
4.1973	34
4.1973	32
4.1973	28
4.1973	27
4.1973	25

Query 15, business growth

2.8141	42
2.8141	38
2.8141	37
2.8141	30
2.8141	28
2.8141	18
2.8141	16
2.8141	15
2.8141	13
2.8141	10

Query 16, economy

4.1623	41
4.1623	30

4.1623	18
4.1623	15
4.1623	12
4.1623	10
4.1623	7
0	43
0	42
0	40

Query 17, abolish slavery

2.4823	31
2.2566	43
2.2566	30
2.2566	28
2.2566	27
2.2566	23
2.2566	22
2.2566	19
2.2566	11
2.2566	3

SKIP BI-GRAM

Query 1

0	43
0	42
0	41

0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 2

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 3

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 4

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 5

0	43

0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 6

0.00183486238532110	28
0.00175438596491228	6
0.00173310225303293	14
0.00168918918918919	39
0.00168634064080944	42
0.00166944908180300	15
0.00165289256198347	27
0.00164473684210526	29
0	43
0	41

Query 7

0.00351493848857645	40
0.00183150183150183	28
0.00177619893428064	5
0.00176991150442478	8
0.00176678445229682	16
0.00176056338028169	18
0.00175438596491228	31
0.00175131348511384	11
0.00174520069808028	34
0.00168918918918919	37

Query 8

0.00319488817891374	4
0.00183150183150183	41
0.00177935943060498	5
0.00177304964539007	17
0.00176991150442478	20
0.00176991150442478	16
0.00176678445229682	10
0.00176366843033510	18
0.00176056338028169	40
0.00175438596491228	21

Query 9, USA

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 10, good president

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 11, president who is good

0.0018	20
0.0018	19
0.0017	25
0.0017	23
0.0016	27
0	43
0	42
0	41
0	40
0	39

Query 12, international war

0.0017	33
0	43
0	42
0	41

0	40
0	39
0	38
0	37
0	36
0	35

Query 13, elected 1982

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 14, election 1982

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 15, business growth

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35

0	34

Query 16, economy

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34

Query 17, abolish slavery

0	43
0	42
0	41
0	40
0	39
0	38
0	37
0	36
0	35
0	34