The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.3.373
27(3), 373–386

English Bible Text Visualization Using Word Clouds and Dynamic Graphics Technology

Dae-Heung Jang[a,1]

[a]Department of Statistics, Pukyong National University

(Received October 30, 2013; Revised March 14, 2014; Accepted April 07, 2014)

Abstract

A word cloud is a visualization of word frequency in a given text. The importance of each word is shown in font size or color. This plot is useful for quickly perceiving the most prominent words and for locating a word alphabetically to determine its relative prominence. With dynamic graphics, we can find the changing pattern of prominent words and their frequencies according to the changing selection of chapters in a given text. We can define the word frequency matrix. In this matrix, rows are chapters in text and columns are ranks corresponding to word frequency about the words in the text. We can draw the word frequency matrix plot with this matrix. Dynamic graphic can indicate the changing pattern of the word frequency matrix according to the changing selection of the range of ranks of words. We execute an English Bible text visualization using word clouds and dynamic graphics technology.

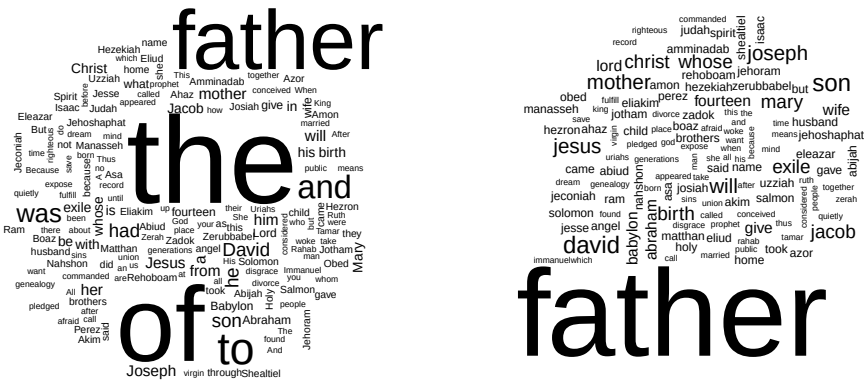
Keywords: Word clouds, word frequency matrix, dynamic graphics, Bible.

1. 서론

단어 구름은 문자 텍스트 상의 복수개의 단어들을 대상으로 그 단어들의 출현 빈도에 비례하는 글자의 크기나 글자의 색깔로 중요도를 나타내는 텍스트 시각화 방법이다 (Wikipedia, 2013). 즉 문자 텍스트 상의 특정 단어가 다른 단어에 비하여 상대적으로 출현 빈도가 높으면 글자의 크기를 크게 하거나 강조할 수 있는 색깔을 선택하여 이 단어를 표시한다. 이러한 작업을 통하여 문자 텍스트에서 어떤 단어들이 핵심단어들인 지를 한 눈에 파악할 수 있게 된다. 이 기법은 텍스트 시각화 기법의 하나로 등장하였다. 이 그림은 텍스트 상의 핵심단어를 재빨리 인지하고 단어들의 상대적 출현빈도수에 맞추어 배열하는 데 유용하다. 단어 구름은 최근까지도 학자들에 의하여 꾸준히 연구되고 있다 (Gottron, 2009; Kaptein과 Kamps, 2011; Seifert 등, 2011; Huh, 2013; Riggs와 Hu, 2013).

성경은 텍스트 시각화를 훈련시킬 수 있는 좋은 도구로 쓰일 수 있다. 본 논문은 신약 성서 중 마태복음을 대상으로 텍스트 시각화 기법으로서 단어 구름과 동적 그래픽스를 어떻게 사용할 수 있는 지를 보이고자 한다. 마태복음은 세무공무원이며 예수의 제자인 마태가 쓴 복음서로서 신약 성서에 나타나는 4개의 복음서 중 첫 번째 복음서이다. 이 복음서에는 예수의 가르침이 상세히 나와 있고 구약 성서의 예언서에서 약속했던 메시아가 다윗의 자손인 예수임을 유태인들에게 확신시키기 위하여 저술되

[1]Department of Statistics, Pukyong National University, 45 Yongso-Ro, Nam-Gu, Pusan 608-707, Korea.
E-mail: dhjang@pknu.ac.kr



(a) Word cloud with focusing all words (b) Word cloud with focusing noun and verb

Figure 2.1. Word clouds for first chapter of The gospel according to Matthew

었다. 총 28개의 장(chapters), 1,071개의 절(verses), 22,610개의 단어(words, 중복 허락, 중복 배제 시에는 2,462개)로 구성되어 있다. 예수 그리스도의 탄생으로부터 예수의 사역, 십자가 수난과 부활까지 드라마틱한 예수의 생애를 통한 예수의 가르침이 주된 내용이다. 성경은 영어성경 중 NIV판(New International Version; NIV)을 사용하였고 R-package로는 wordcloud와 KoNLP를 사용하였다. 영어 단어를 구분할 때 대소문자를 구분하였다. 예로 영어성경에서 'father'와 'son'의 의미는 아버지와 아들이나 'Father'와 'Son'의 의미는 하나님과 예수님으로 의미가 달라진다.

단어구름 기법이 텍스트마이닝에서 중요한 시각화 방법으로 쓰일 수 있기 때문에 본 논문을 통하여 이러한 단어구름을 동적으로 확인할 수 있는 동적 단어구름을 제안하고 시너지 효과를 얻기 위하여 동적 단어구름과 병행하여 그릴 수 있는 동적 점차트를 제시하고자 한다.

2절에서 단어 구름 기법이 마태복음에 대한 텍스트 시각화에 어떻게 쓰일 수 있는 지를 살펴보고 동적 그래픽스를 이용하여 동적 점차트와 동적 단어구름을 제안하였다. 3절에서는 단어출현빈도행렬을 제안하고 이 단어출현빈도행렬를 기반으로 하는 동적 그래픽스 기법이 마태복음에 대한 텍스트 시각화에 어떻게 쓰일 수 있는 지를 보였다. 4절에서는 부차적인 텍스트분석에 대하여 언급하였고 5절에서 결론으로 마무리하였다.

2. 단어 구름과 동적 그래픽스

Figure 2.1(a)는 마태복음 1장에 대한 단어 구름을 나타낸다. 이 그림을 통하여 정관사 the(출현빈도수: 59)와 전치사 of(출현빈도수: 50)를 제외하면 가장 출현빈도가 높은 단어가 father(출현빈도수: 39)임을 알 수 있다. 이는 마태복음 1장 전반부(1절에서 17절까지)의 내용이 예수 그리스도의 족보를 나타내어 father 단어가 많이 나타나게 되었다. 반면 mother는 출현빈도수는 5이다. 예수께서 다윗(David, 출현빈도수: 6)의 후손임을 강조하고 있고 son도 출현 빈도수가 6이다. 1장 후반부(18절에서 25절까지)는 예수그리스도의 탄생에 대한 이야기로 구성되어 있다. 예수의 부친 요셉(Joseph, 출현빈도수: 5)과 모친 마리아(Mary, 출현빈도수: 4), 탄생(birth, 출현빈도수: 4), 예수(Jesus, 출현빈도수: 5), 그리스도(Christ, 출현빈도수: 4)가 눈에 띈다. 하나의 단어 구름을 통하여 예수 그리스도의 족보와 예수 그리스도의 탄생과 관련된 단어들을 우리는 한 번에 인지하게 된다. 가장 핵심적인 단어는 father가 된다. 이 단어 구름은 일종의 막대그래프의 변형이라 볼 수 있는 데 텍스트 마이닝에서 핵심단어에 대한

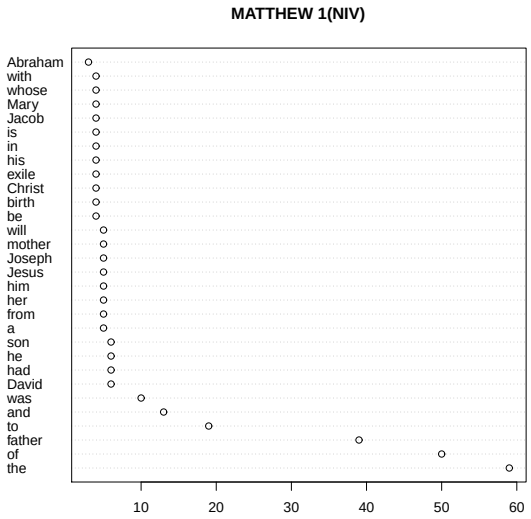


Figure 2.2. Dot chart(Rank: 1–30) for first chapter of The gospel according to Matthew

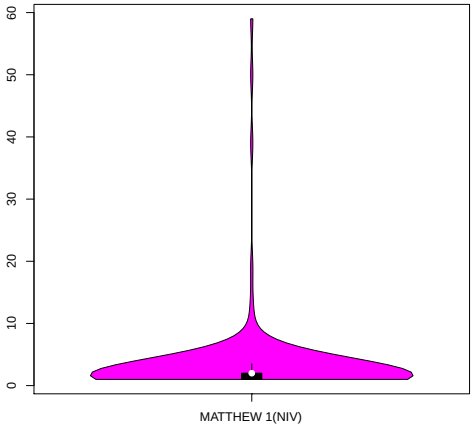


Figure 2.3. Violin plot for first chapter of The gospel according to Matthew

인지도를 높일 수 있는 장점이 있다. 만일 명사와 동사를 중심으로 하여 단어 구름(대소문자를 구별하지 않은)을 만들면 Figure 2.1(b)와 같고 Figure 2.1(a)와는 다른 그림이 그려진다. 이 단어 구름과 비교하기 위하여 출현빈도수 순위 1위에서 30위까지의 점챠트(dot chart)를 그려 보면 Figure 2.2와 같다. 마태복음 1장에 나타나는 단어 수는 172개이고 이 단어들의 출현 빈도수에 대한 분포를 바이올린그림으로 그려 보면 Figure 2.3과 같다. 전형적인 Zipf 분포를 이룸을 알 수 있다. 28장 각각에 대하여 출현 빈도수에 대한 분포들을 나타내어 보니 Figure 2.3처럼 Zipf 분포를 이룸을 알 수 있다. Zipf 분포는 거듭곱법칙 분포로서 다음과 같이 정의된다.

$$F = \frac{K}{R^{\beta}},$$

Table 2.1. Words of The gospel according to Matthew

Chapter	No. of Verses /No. of Words	Word(Rank: 1–15)
1	25 /171	the(59), of(50), **father(39)**, to(19), and(13), was(10), David(6), had(6), he(6), son(6), a(5), from(5), her(5), him(5), Jesus(5), Joseph(5), mother(5), will(5)
2	23 /219	the(45), and(30), to(21), of(19), in(17), **he(15)**, they(12), was(12), **Herod(9)**, a(8), child(8), for(8), had(8), him(7), his(7)
3	17 /200	the(21), and(14), to(13), of(12), **he(8)**, I(7), is(7), you(7), for(6), him(6), with(6), in(5), was(5), baptized(4), be(4), from(4), his(4), **John(4)**, will(4)
4	25 /239	the(39), and(26), of(16), **him(15)**, to(12), **Jesus(10)**, he(8), in(8), you(8), their(7), a(6), on(6), said(6), will(6), Galilee(5), God(5), is(5), was(5)
5	48 /349	the(60), **you(43)**, to(32), and(31), your(29), of(24), for(23), be(18), who(18), it(17), that(16), are(15), not(15), will(15), is(14), **heaven(10)**
6	34 /264	**you(35)**, the(30), your(30), not(23), and(20), will(19), do(18), in(17), be(15), is(15), to(14), **Father(12)**, of(12), they(11), or(8)
7	29 /253	the(28), and(26), **you(23)**, to(19), will(17), who(12), your(12), a(10), in(10), is(9), of(9), not(8), **fruit(7)**, good(7), it(7), them(7)
8	34 /304	the(48), and(37), to(24), **him(19)**, he(14), of(14), **Jesus(13)**, you(13), I(10), said(10), a(8), came(8), was(8), will(8), with(8)
9	38 /313	the(43), and(38), to(19), **he(16)**, **Jesus(14)**, said(13), him(12), of(11), they(11), a(10), this(10), will(10), that(9), his(8), is(8), on(8), was(8), your(8)
10	42 /341	the(42), of(25), **you(25)**, and(24), to(23), will(23), not(21), his(19), a(16), is(15), be(14), who(14), or(13), me(12), in(11), your(11), **Father(4)**, **peace(4)**, **town(4)**
11	30 /278	the(36), **you(30)**, and(29), to(24), of(14), in(13), is(11), I(9), who(9), for(8), will(8), **John(7)**, a(6), are(6), did(6), it(6), not(6)
12	50 /375	the(66), and(46), of(31), to(27), **will(27)**, a(18), it(18), he(17), is(17), out(15), you(15), be(14), not(14), in(13), him(12), I(12), **Jesus(7)**
13	58 /413	the(99), and(54), of(40), **it(23)**, in(21), is(21), he(20), a(19), to(19), them(17), will(17), The(16), his(15), they(15), who(14), **seed(13)**
14	36 /284	the(48), and(30), to(27), **he(19)**, of(16), him(14), **Jesus(11)**, on(11), they(10), a(9), them(9), said(8), was(8), disciples(7), had(7), his(7)

여기서 R은 단어들의 내림차순 순위, F는 단어들의 출현빈도수, K는 $N = 172$(출현단어개수(중복 배제))와 $M = 8{,}509$(출현단어개수 $N = 172$에 대응되는 총출현단어수(중복 허락), 출현빈도수들의 총합)가 주어졌을 때의 상수이다.

각 장에 대응되는 단어 구름들을 그려 보면 다음 Table 2.1과 Table 2.2와 같이 출현빈도수 순위가 1위에서 15위까지의 단어들을 확인할 수 있고 각 장을 대표하는 핵심단어들을 추출할 수 있게 된다. Table 2.1과 Table 2.2 내의 각 장에서 진하게 표시된 단어는 정관사, 전치사, 접속사를 제외한 단어들을 대상으로 할 때의 핵심단어이고, 밑줄이 있고 진하게 표시된 단어는 명사를 대상으로 할 때의 핵심단어이다. 예로, 1장에서 핵심단어가 'father'이었던 것처럼 예수의 산상설교 부분인 5–7장에서는 정관사, 전치사, 접속사를 제외한 단어들을 대상으로 할 때의 핵심단어가 'you'이고 명사를 대상으로 할 때의 핵심단어는 각각 'heaven', 'Father', 'fruit'이다. 예루살렘 멸망에 대한 예언이 나타나는 24장에서는 정관사, 전치사, 접속사를 제외한 단어들을 대상으로 할 때의 핵심단어가 'will'이고 명사를 대상으로 할 때의 핵심단어는 'Son' 등이다. 총 28개의 장 중 14개의 장에서 'Jesus'가 명사를 대상으로 할 때의 핵심단어가 된

Table 2.2. Words of The gospel according to Matthew(Continued)

Chapter	No. of Verses /No. of Words	Words(Rank: 1–15)
15	39 /321	the(45), and(29), to(23), of(18), a(12), **him(11)**, **Jesus(11)**, that(11), you(10), they(9), from(8), he(8), disciples(7), have(7), his(7), me(7), not(7), them(7)
16	28 /268	the(42), and(25), of(22), to(22), **you(21)**, will(13), he(12), his(12), be(9), for(9), in(9), **Jesus(9)**, not(8), it(7), on(7), that(7)
17	27 /269	the(36), and(24), to(21), **Jesus(14)**, you(14), he(11), him(11), them(10), of(9), they(9), I(8), came(7), disciples(7), will(7), a(6), be(6), for(6), from(6), it(6), said(6), with(6)
18	35 /290	the(36), to(31), **you(31)**, and(25), he(22), of(18), that(15), him(14), in(13), I(12), be(11), a(10), **heaven(10)**, his(9), on(9), will(9), your(9)
19	30 /270	the(26), to(24), and(23), **Jesus(12)**, of(12), you(12), a(11), for(11), man(11), his(9), them(9), will(9), is(8), not(8), this(8), who(8)
20	34 /267	the(42), to(35), and(33), **them(19)**, of(16), they(14), you(14), be(10), **Jesus(9)**, a(8), I(8), said(8), will(8), have(7), he(7)
21	46 /374	the(83), and(44), to(38), **you(34)**, he(27), of(27), a(20), will(20), them(18), they(15), him(14), **Jesus(14)**, it(12), on(12), was(11)
22	46 /316	the(52), to(32), and(29), of(15), is(14), **his(13)**, they(13), him(12), he(11), in(11), them(10), you(10), **God(9)**, said(9), The(9)
23	39 /300	the(64), **you(45)**, and(36), of(30), to(25), You(16), by(15), are(11), on(11), will(11), in(10), swears(10), not(9), they(9), who(9), **law(8)**, **teachers(8)**
24	51 /378	the(85), **will(50)**, and(41), of(40), be(26), to(23), in(20), you(18), not(15), that(15), is(14), he(12), his(10), one(10), all(8), him(8), it(8), **Son(7)**, **time(7)**
25	46 /297	the(53), and(46), **you(44)**, I(26), to(24), will(19), in(16), me(16), of(15), not(14), a(13), did(13), his(13), with(13), have(12), was(12), who(12), **talents(9)**
26	75 /488	the(102), and(54), to(54), of(45), **you(44)**, I(29), **Jesus(27)**, him(25), he(23), will(23), said(21), with(21), is(20), it(18), a(17), Then(17)
27	66 /467	the(103), and(51), to(42), **him(41)**, of(30), they(26), a(22), he(22), **Jesus(20)**, it(15), his(14), in(14), on(14), that(14), had(13), was(13)
28	20 /193	the(33), and(22), to(19), of(12), **him(8)**, them(7), you(7), they(6), **disciples(5)**, **Jesus(5)**, said(5), were(5), afraid(4), came(4), go(4), has(4), he(4), his(4), his(4), I(4), on(4), Then(4)

다. Table 2.1과 Table 2.2에서 단어 뒤의 괄호 안 숫자는 각 장에서의 출현빈도수를 나타낸다. 마태복음 28장 모두를 합친 전체 텍스트에서 출현빈도수 순위가 1에서 1,000인 단어들을 대상으로 단어 구름을 그리면 Figure 2.4와 같다. 출현빈도수 순위가 1위에서 4위인 단어들인 the(정관사), and(접속사), to(전치사), of(전치사)를 제외하면 핵심단어가 인칭대명사 'you', 인칭대명사 'he'(Jesus, I, his, him 포함, 이 단어들의 대부분은 예수 그리스도를 지칭하는 단어들임), 조동사 'will' 3개로 압축됨을 알 수 있다. 핵심단어 'you'는 예수의 가르침을 받는 상대자들을 지칭하는 단어이다. 마태복음 28장 모두를 합친 전체 텍스트에서 Figure 2.4의 단어 구름과 비교하기 위하여 출현빈도수 순위 1위에서 30위까지의 점챠트(dot chart)를 그려 보면 Figure 2.5와 같다.

만일 명사와 동사를 중심으로 하여 단어 구름을 만들면 Figure 2.6과 같고 앞에서 그렸던 단어 구름인

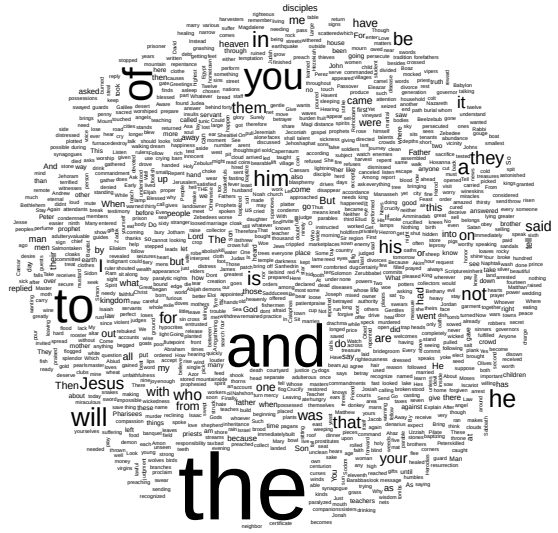


Figure 2.4. Word cloud for 28 chapters of The gospel according to Matthew with focusing all words

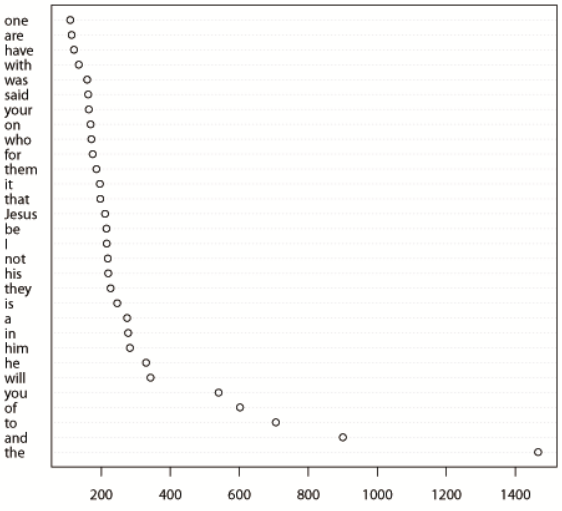


Figure 2.5. Dot chart(Rank: 1–30) for 28 chapters of The gospel according to Matthew

Figure 2.4와는 다른 그림이 그려진다. 핵심단어가 조동사 'will', 고유대명사 'Jesus', 동사 'said'로 나타난다. 이 세 단어가 마태복음의 성격(예수의 복음)을 잘 나타내는 단어들이라 볼 수 있다. 마태복음 26장을 모든 단어를 동등하게 보고 그린 단어 구름과 명사와 동사를 중심으로 하여 구한 단어 구름을 그려보면 다음 Figure 2.7과 같다. 모든 단어를 동등하게 보고 그린 단어 구름에서는 'you'가 핵심단어가 되나 명사와 동사를 중심으로 하여 구한 단어 구름에서는 'Jesus'가 핵심단어가 된다.

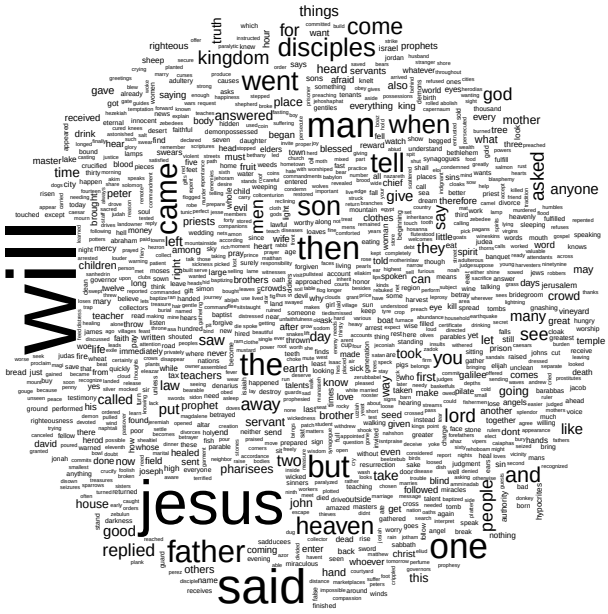
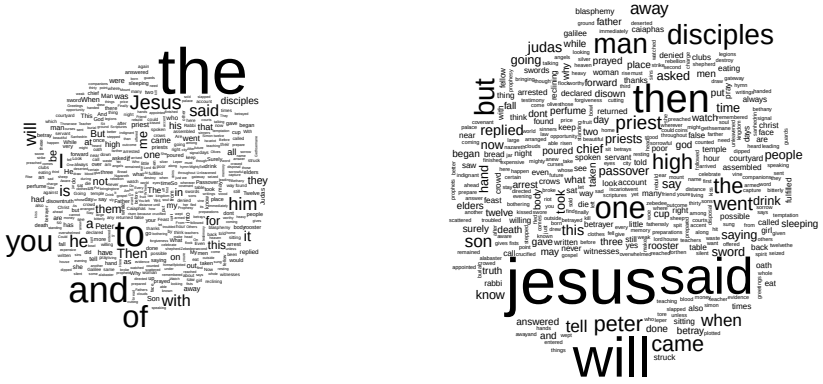


Figure 2.6. Word cloud for 28 chapters of The gospel according to Matthew with focusing noun and verb



(a) Word cloud with focusing all words (b) Word cloud with focusing noun and verb

Figure 2.7. Word clouds for 26th chapter of The gospel according to Matthew

우리는 동적 그래픽스를 이용하여 동적 점차트와 동적 단어구름을 그릴 수 있다. 이러한 동적 그래픽스를 이용하여 단어들의 출현빈도수의 패턴을 동적으로 확인할 수 있다. 마태복음 1장에 대하여 동적 점차트(출현빈도수 순위 1위에서 30위까지)와 동적 단어구름을 그리면 다음 Figure 2.8과 같다. 'father'가 핵심단어임을 동적으로 확인할 수 있다. 슬라이드바를 좌우로 이동시키면 마태복음 각 장을 바꿔가며 각 장의 출현빈도수의 패턴과 단어 구름 패턴의 변화를 연속적으로 확인할 수 있다. Figure 2.9는 마태복음 28장에 대하여 동적 점차트(출현빈도수 순위 1위에서 30위까지)와 동적 단어구름을 그린 그림이다. 명사를 중심으로 할 때 'disciples'와 'Jesus'가 핵심단어들임을 동적으로 확인할 수 있다.

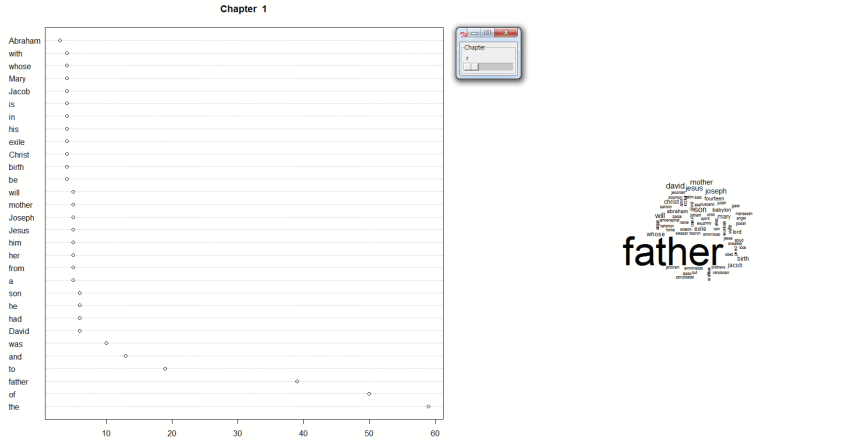


Figure 2.8. Dynamic dot chart(Rank: 1–30) and dynamic word cloud for first chapter of The gospel according to Matthew

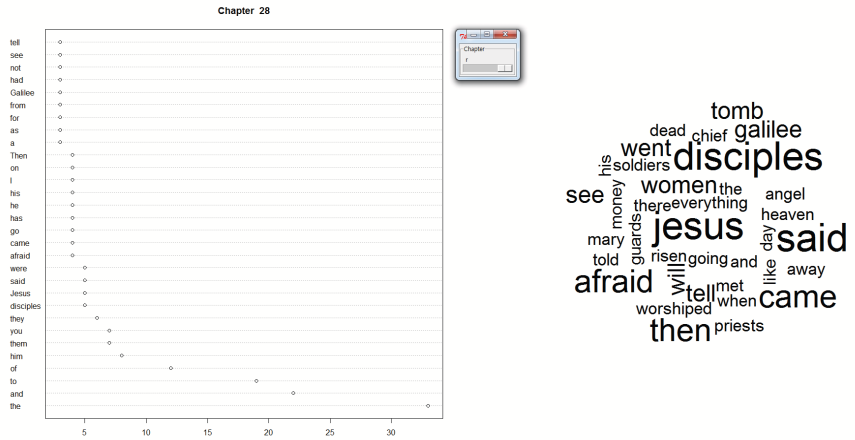


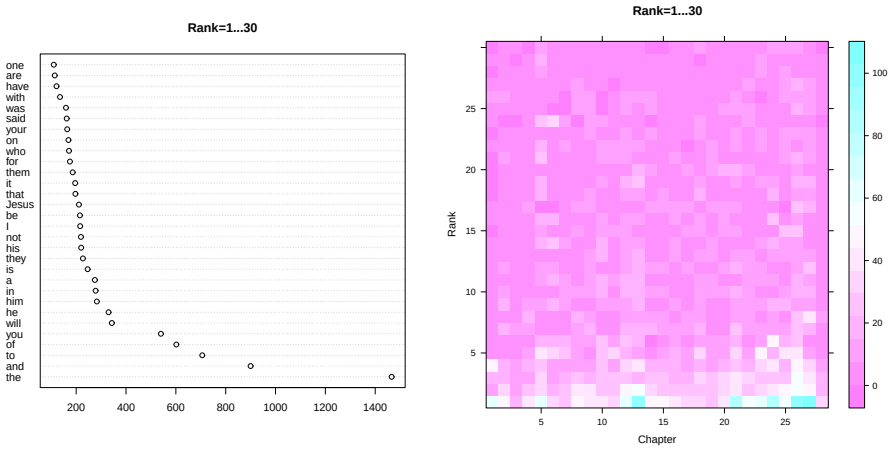
Figure 2.9. Dynamic dot chart(Rank: 1–30) and dynamic word cloud for 28th chapter of The gospel according to Matthew

3. 단어출현빈도행렬그림과 동적 그래픽스

주어진 텍스트(영어성경)을 대상으로 우리는 다음과 같은 단어출현빈도행렬을 정의할 수 있다.

$$O = \{o_{ij}\}, \quad i = 1, 2, \ldots, c; \; j = 1, 2, \ldots, r,$$

여기서 마태복음의 예에 적용하면 $c = 28, r = 2{,}462$이 되고 o_{ij}는 텍스트에 나타나는 2,462개의 각각 단어에 대하여 이 단어가 i-번째 장(chapter)에서 몇 번이나 나타나는 지를 알 수 있는 출현빈도수를 나타낸다. o_{ij} 값을 통하여 특정 단어가 마태복음 몇 장에 몇 번 나타나는지를 알 수 있고 이 단어의 출현빈도수 순위를 알 수 있다. 이러한 단어출현빈도행렬을 시각화한 그림을 단어출현빈도행렬그림이라 하자. 단어출현빈도행렬그림은 단어출현빈도행렬에서의 각 원소에 대응하여 명암을 주거나 색깔을 입혀 출현빈도수를 구별하는 그림이다. 마태복음의 예에서는 $c = 28$에 비하여 $r = 2{,}462$이어서 r이 c에



(a) Dot chart (b) Word frequency matrix plot

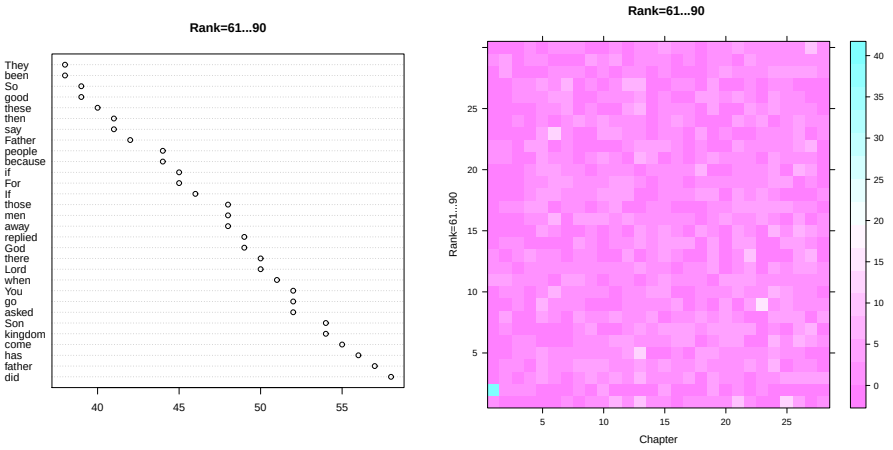
Figure 3.1. Dot chart and word frequency matrix plot(Rank: 1–30) for 28 chapters of The gospel according to Matthew

비하여 매우 크므로 출현빈도행렬그림을 그리면 출현빈도수 패턴을 보기가 어려움으로 출현빈도수 순위 개수를 적당히 주어 단어출현빈도행렬의 부분행렬에 대하여 연속적으로 그려 보면 출현빈도수 패턴의 변화하는 모습을 볼 수 있게 된다. 마태복음 텍스트를 이용하여 출현빈도수 순위 1위에서 30위까지의 점차트와 출현빈도행렬그림을 그려 보면 Figure 3.1과 같다. $c = 28$이므로 출현빈도수 순위 개수를 30개로 정하였다. Figure 3.1에서 우리는 몇 가지 특징들을 살펴 볼 수 있다.

1. 출연빈도수 1위인 정관사 'the'가 상대적으로 많이 나오는 장은 13, 21, 24, 26, 27장이다. 이는 이들 장들이 절수가 많은 장들이라는 연유에 기인한다.
2. 출현빈도수 순위가 5위인 'you'를 보면 출현빈도수가 40개 이상인 장들은 5, 23, 25, 26장이고 출현빈도수가 30개 이상인 장들은 6, 11, 18장이다.
3. 출현빈도수 순위가 6위인 'will'을 보면 출현빈도수가 50개인 24장(예루살렘 멸망에 대한 예언)이 압도적으로 많음을 알 수 있다.
4. 출현빈도수 순위가 7위인 'he'(예수를 지칭하는 단어)를 보면 출현 빈도수가 매우 두드러진 장이 없음에 비하여 출현빈도수 순위가 8위인 'him'(예수를 지칭하는 단어)를 보면 27장에 압도적으로 많은데(출현빈도수: 41)이 27장은 예수가 빌라도 법정에 서는 장면과 예수가 십자가상에서 돌아가심을 기술한 장면이 나오는 장이다.
5. 명사를 대상으로 할 때 출현빈도가 가장 많은 단어는 'Jesus'이다. 다른 장들에 비하여 26장과 27장에 언급이 많은 데 이 두 개의 장이 예수의 수난과 십자가 죽음을 나타내는 장들이다.

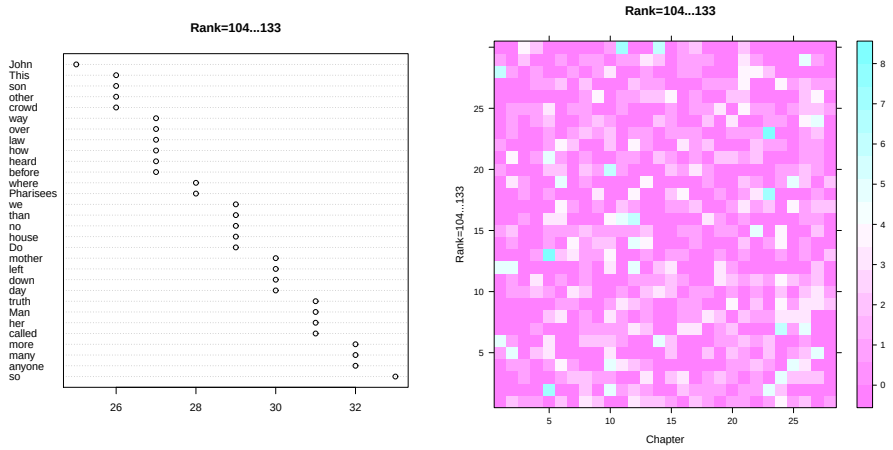
마태복음 텍스트를 이용하여 출현빈도수 순위 61위에서 90위까지의 점차트와 단어출현빈도행렬그림을 그려 보면 Figure 3.2와 같다. 재미있는 특징이 출현빈도수 순위가 62위인 'father'를 보면 1장에서 두드러지게 출현 빈도가 높다는 것이다. 28장 전체에 걸쳐 57번 출현하는데 1장에서 무려 39번 출현한다.

마태복음 텍스트를 이용하여 출현빈도수 순위 104위에서 133위까지의 점차트와 단어출현빈도행렬그림을 그려 보면 Figure 3.3과 같다. 28장에 걸친 출현빈도수에 따라 단어들이 그룹을 이룸을 알 수 있는데 같은 그룹에 속한 단어들이라 할지라도 각 장에 나타나는 출현 패턴이 다양함을 알 수 있다.



(a) Dot chart (b) Word frequency matrix plot

Figure 3.2. Dot chart and word frequency matrix plot(Rank: 61–90) for 28 chapters of The gospel according to Matthew

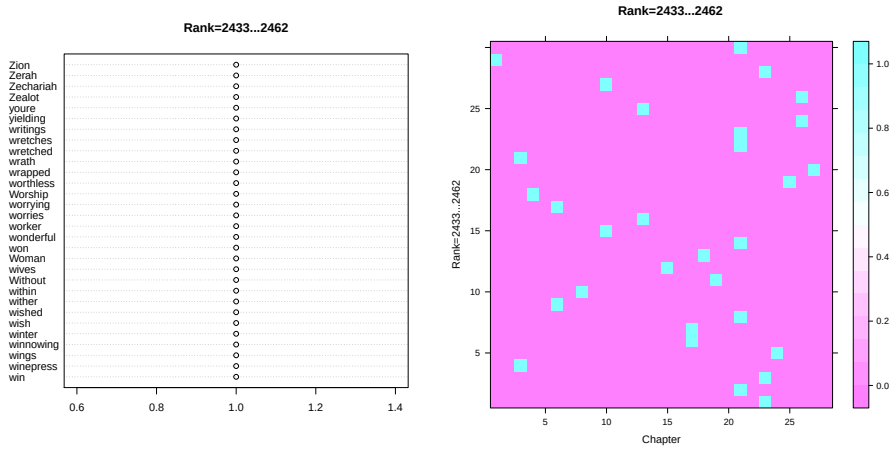


(a) Dot chart (b) Word frequency matrix plot

Figure 3.3. Dot chart and word frequency matrix plot(Rank: 104–133) for 28 chapters of The gospel according to Matthew

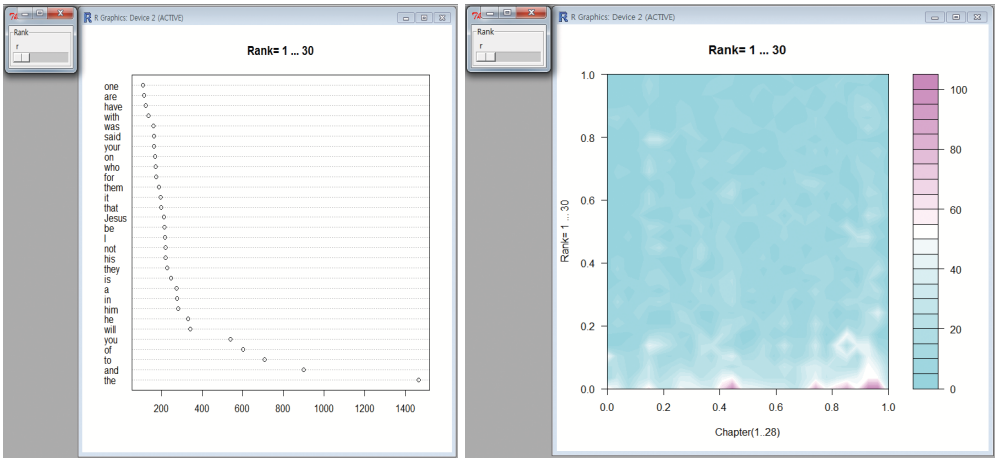
마태복음 텍스트를 이용하여 출현빈도수 순위 2,433위에서 2,462위(마지막 순위)까지의 점차트와 단어출현빈도행렬그림을 그려 보면 Figure 3.4와 같다. 모든 단어들이 28장에 걸쳐 1번씩만 나타남을 알 수 있다.

우리는 동적 그래픽스를 이용하여 동적 점차트와 동적 단어출현빈도행렬그림을 그릴 수 있다. 이러한 동적 그래픽스를 이용하여 단어들의 출현빈도수의 패턴을 동적으로 확인할 수 있다. 슬라이드바를 좌우로 이동시키면 출현빈도수 순위를 바꿔가며 출현빈도수의 패턴과 단어출현빈도행렬의 패턴의 변화를 연속적으로 확인할 수 있다. 예를 들어 슬라이드바를 조정하여 순위 31을 선택하면 그래픽 화면에 순위 31에서 60까지 대응되는 동적 점차트와 동적 단어출현빈도행렬그림을 우리는 얻을 수 있다. 슬라이드



(a) Dot chart (b) Word frequency matrix plot

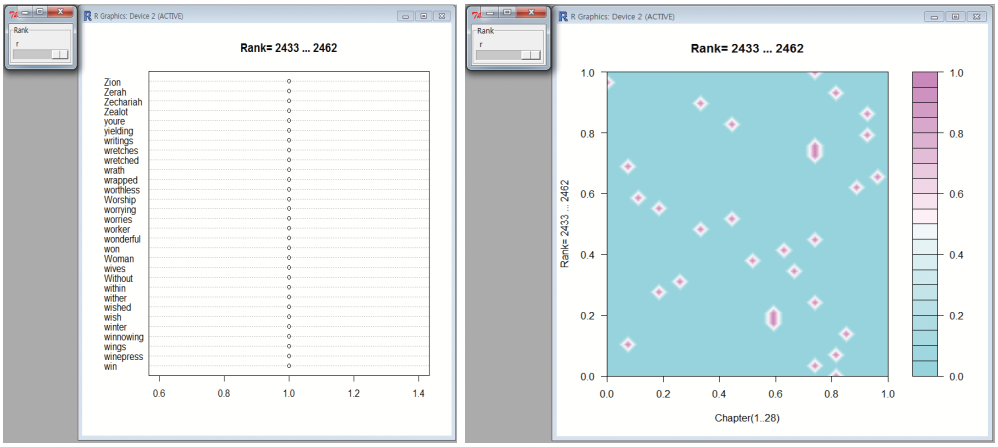
Figure 3.4. Dot chart and word frequency matrix plot(Rank: 2,433–2,462) for 28 chapters of The gospel according to Matthew



(a) Dynamic dot chart (b) Dynamic word frequency matrix plot

Figure 3.5. Dynamic dot chart and dynamic word frequency matrix plot(Rank: 1–30) for 28 chapters of The gospel according to Matthew

바를 좌에서 우로 이동시키면서 출현빈도수의 패턴과 단어출현빈도행렬의 패턴의 변화를 보면 출현빈도수 순위 1–30에서는 순위가 1에서 30으로 바뀌며 출현빈도수가 기하급수적으로 줄고 있음을 알 수 있고 출현빈도수 순위가 30을 넘어가면 출현빈도수가 선형적으로 주는 것을 알 수 있다. 또한 출현빈도수 순위가 100을 넘어가면 출현빈도수가 계단 모양을 나타내며 주는 것을 알 수 있다. 출현빈도수 순위 1위에서 30위까지의 동적 점차트와 동적 단어출현빈도행렬그림을 그리면 다음 Figure 3.5와 같다. Figure 3.6은 출현빈도수 순위 2,433위에서 2,462위(마지막 순위)까지의 동적 점차트와 동적 단어출현빈도행렬그림을 그린 그림들이다.



(a) Dynamic dot chart (b) Dynamic word frequency matrix plot

Figure 3.6. Dynamic dot chart and dynamic word frequency matrix plot(Rank: 2,433–2,462) for 28 chapters of The gospel according to Matthew

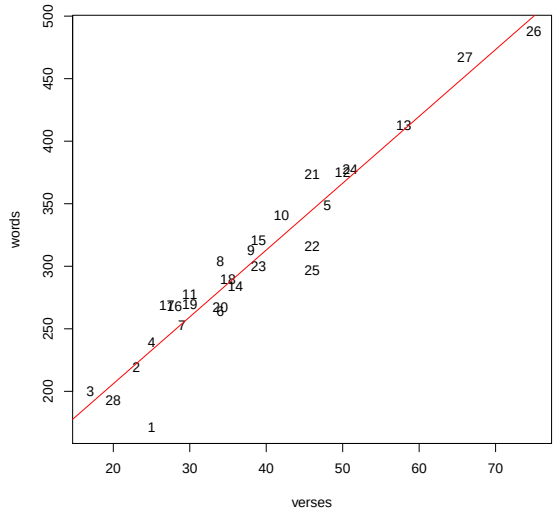


Figure 4.1. Scatterplot and regression line for 28 chapters of The gospel according to Matthew

4. 부차적인 텍스트분석

마태복음 28장 각각의 장에 대하여 절의 개수와 단어개수(중복배제)를 조사한 후 단순선형회귀식을 구하니 다음 Figure 4.1과 같이 나타났다. 단순선형회귀식은 단어개수 $= 99.193 + 5.344 * (\text{절의 개수})$이고 분산분석 결과 p-값이 5.283×10^{-16}으로서 매우 유의하였다. 한 절 증가할 때마다 약 다섯 단어씩 증가함을 알 수 있다.

마태복음 28장 각각 2,462개 단어별 출현 빈도수를 구한 후 28장 서로의 상관계수들을 구한 후 상관계수행렬그림을 그리니 다음 Figure 4.2와 같았다. 전체적으로는 높은 양의 상관관계가 있음을 알 수 있

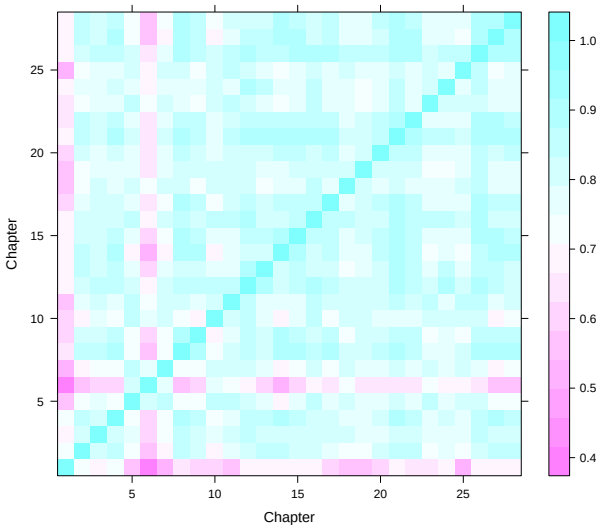


Figure 4.2. Correlation matrix plot for 28 chapters of The gospel according to Matthew

다. 그러나 1장(예수의 족보와 예수의 탄생)에서 사용되는 단어들이 다른 장들과 상대적으로 다른 패턴을 보이고 5, 6, 7장(산상설교)에서 사용되는 단어들이 다른 장들과 상대적으로 다른 패턴을 보이는 것을 확인할 수 있다.

5. 결론

단어 구름은 문자 텍스트 상의 복수개의 단어들을 대상으로 그 단어들의 출현 빈도에 비례하는 글자 크기나 글자 색깔로 중요도를 나타내는 텍스트 시각화 방법이다. 한 예로 성경을 텍스트 삼아 단어 구름을 그려 보면 단어 구름이 텍스트 상의 핵심단어를 잘 각인시켜 주는 그래픽 도구임을 알 수 있었다. 또한 동적 그래픽스를 통하여 단어들의 출현빈도수의 패턴을 동적으로 확인할 수 있었다. 이런 단어 구름과 동적 그래픽스를 병행하여 텍스트 시각화 도구들로 활용한다면 우리는 시너지 효과를 얻을 수 있다.

References

English Bible (2013). New International Version(NIV), *http://www.ctmbible.net/ibible/read.asp#Bible_ Seleted*

Gottron, T. (2009). Document word clouds: Visualising web documents as tag clouds to aid users in relevance decisions, *Research and Advanced Technology for Digital Libraries-Lecture Notes in Computer Science*, **5714**, 94–105.

Huh, M. H. (2013). Moving data pictures, *The Korean Journal of Applied Statistics*, **26**, 999–1007.

Kaptein, R. and Kamps, J. (2011). Word clouds of multiple search results, *Multidisciplinary Information Retrieval-Lecture Notes in Computer Science*, **6653**, 78–93.

Riggs, R. J. and Hu, S. J. (2013). Disassembly liason graphs inspired by word clouds, *Procedia CIRP*, **7**, 521–526.

Seifert, C., Ulbrich, E. and Granitzer, M. (2011). Word clouds for efficient document labeling, *Discovery Science-Lecture Notes in Computer Science*, **6926**, 292–306.

Wikipedia (2013). Word cloud, *http://en.wikipedia.org/wiki/Word_cloud*

단어 구름과 동적 그래픽스 기법을 이용한 영어성경 텍스트 시각화

장대흥[a,1]

[a]부경대학교 통계학과

(2013년 10월 30일 접수, 2014년 03월 14일 수정, 2014년 04월 07일 채택)

요 약

단어 구름은 문자 텍스트 상의 복수개의 단어들을 대상으로 그 단어들의 출현 빈도에 비례하는 글자의 크기나 글자의 색깔로 중요도를 나타내는 텍스트 시각화 방법이다. 이 그림은 텍스트 상의 핵심단어를 재빨리 인지하고 단어들의 상대적 출현빈도수에 맞추어 배열하는 데 유용하다. 동적 그래픽스를 이용하여 텍스트 장들의 변화에 따른 핵심단어와 단어출현빈도의 패턴의 변하는 모습을 살필 수 있다. 행들이 텍스트 상의 장들이고 열들이 텍스트에 출현하는 단어들의 출현빈도수 순위들인 단어출현빈도행렬을 정의할 수 있고 이 행렬을 이용하여 단어출현빈도행렬그림을 그릴 수 있다. 동적 그래픽스를 이용하여 출현빈도수 순위의 변화에 따른 단어출현빈도행렬의 패턴의 변하는 모습을 살필 수 있다. 우리는 단어 구름과 동적 그래픽스 기법을 사용하여 영어성경 텍스트 시각화를 수행할 수 있다.

주요용어: 단어 구름, 단어출현빈도행렬, 동적 그래픽스, 성경.

[1](608-737) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: dhjang@pknu.ac.kr.