(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년08월18일
(11) 등록번호 10-1769035
(24) 등록일자 2017년08월10일

(51) 국제특허분류(Int. Cl.)
G06F 17/27 (2006.01) *G06F 17/21* (2006.01)
(52) CPC특허분류
G06F 17/2705 (2013.01)
G06F 17/21 (2013.01)
(21) 출원번호 10-2016-0036552
(22) 출원일자 2016년03월28일
심사청구일자 2016년03월28일
(56) 선행기술조사문헌
JP2009223463 A*
JP4547500 B2
JP2010146171 A*
KR1020130036863 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
장봉수
울산광역시 울주군 언양읍 유니스트길 50
김경훈
울산광역시 울주군 언양읍 유니스트길 50
(74) 대리인
특허법인 프렌즈드림

전체 청구항 수 : 총 4 항 심사관 : 경연정

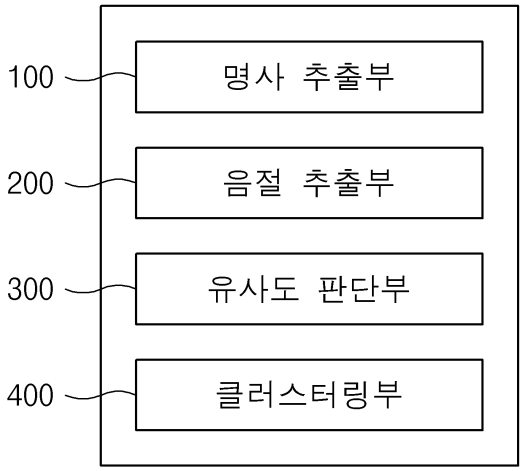
(54) 발명의 명칭 **한국어 텍스트 클러스터링 시스템 및 방법**

(57) 요 약

한국어 텍스트 클러스터링 시스템 및 방법이 개시된다.

본 발명은 텍스트의 명사를 형태소 분석을 통해 추출한 이후 사용자의 의도에 따라 임의의 음절을 다시 추출하여 추출된 음절을 이용하여 텍스트의 유사도 판단 및 클러스터링을 하는 것으로 명사만을 이용한 유사도 판단 및 클러스터링에 비해 정확도는 유지시키되 메모리 소모량은 감소하는 효과가 있다.

대 표 도 - 도1



(52) CPC특허분류
G06F 17/2755 (2013.01)

명 세 서

청구범위

청구항 1

형태소 분석을 이용하여 텍스트의 명사를 추출하는 명사 추출부;

상기 명사 추출부가 추출한 명사에 대하여 사용자의 의도에 따라 선택된 개수의 음절을 추출하는 음절 추출부;

상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 유사도 판단부;

상기 유사도 판단부가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 상기 텍스트를 클러스터링하는 클러스터링부;

를 포함하되,

[수학식 1]

$$TF(t,d) = f_{t,d}$$

상기 수학식 1에서 t는 용어, d는 문서를 의미하고,

[수학식 2]

$$IDF(t,D) = \log \frac{n}{|\{d \in D : t \in d\}|}$$

상기 수학식 2에서 n은 문서 전체의 개수이고, 분모는 용어 t가 포함된 문서의 개수이다. 또한, d 는 문서, D는 문서의 집합을 의미하며,

[수학식 3]

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

[수학식 4]

$$\text{코사인 유사도} = \cos(\theta) = \frac{d_2 \cdot d_1}{\|d_2\| \|d_1\|}$$

상기 유사도 판단부는

상기 텍스트, 상기 음절 및 상기 수학식 1 내지 3을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축하며, 상기 수학식 4를 통해 코사인 유사도 도출하고, 도출된 코사인 유사도 이용하여 상기 텍스트의 유사도를 판단하되,

상기 수학식 1 내지 3의 t는 용어로서, 상기 음절 추출부가 추출한 음절을 의미하는 것을 특징으로 하는 한국어 텍스트 클러스터링 시스템.

청구항 2

제1항에 있어서,

상기 클러스터링부는

특이값 분해(SVD, Singular Value Decomposition) 또는 비음수행렬인수분해(NMF, Non-negative Matrix Factorization) 중 어느 하나를 이용하여 상기 텍스트를 클러스터링하는 것을 특징으로 하는 한국어 텍스트 클러스터링 시스템.

청구항 3

명사 추출부가 형태소 분석을 이용하여 텍스트의 명사를 추출하는 단계;

음절 추출부가 상기 명사 추출부가 추출한 명사에 대하여 사용자의 의도에 따라 선택된 개수의 음절을 추출하는 단계;

유사도 판단부가 상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 단계;

클러스터링부가 상기 유사도 판단부가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 상기 텍스트를 클러스터링하는 단계;

를 포함하되,

상기 유사도 판단부가 상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 단계는

상기 유사도 판단부가 상기 텍스트와 상기 음절을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축하는 단계;

를 포함하고,

상기 유사도 판단부가 상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 단계는

상기 유사도 판단부가 코사인 유사도를 이용하여 상기 텍스트의 유사도를 판단하는 단계;

를 포함하되,

[수학식 1]

$$TF(t,d) = f_{t,d}$$

상기 수학식 1에서 t는 용어, d는 문서를 의미하고,

[수학식 2]

$$IDF(t,D) = \log \frac{n}{|\{d \in D : t \in d\}|}$$

상기 수학식 2에서 n은 문서 전체의 개수이고, 분모는 용어 t가 포함된 문서의 개수이다. 또한, d 는 문서, D는 문서의 집합을 의미하며,

[수학식 3]

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

[수학식 4]

$$\text{코사인 유사도} = \cos(\theta) = \frac{d_k \cdot d_i}{\|d_k\| \|d_i\|}$$

상기 유사도 판단부는

상기 텍스트, 상기 음절 및 상기 수학식 1 내지 3을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축하며, 상기 수학식 4를 통해 코사인 유사도 도출하고, 도출된 코사인 유사도 이용하여 상기 텍스트의 유사도를 판단하되,

상기 수학식 1 내지 3의 t는 용어로서, 상기 음절 추출부가 추출한 음절을 의미하는 것을 특징으로 하는 한국어 텍스트 클러스터링 방법.

청구항 4

제3항에 있어서,

상기 클러스터링부가 상기 유사도 판단부가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 상기 텍스트를 클러스터링하는 단계는

상기 클러스터링부가 특이값 분해(SVD, Singular Value Decomposition) 또는 비음수행렬인수분해(NMF, Non-negative Matrix Factorization) 중 어느 하나를 이용하여 상기 텍스트를 클러스터링하는 단계;

를 포함하는 한국어 텍스트 클러스터링 방법.

청구항 5

삭제

청구항 6

삭제

청구항 7

삭제

청구항 8

삭제

발명의 설명

기 술 분 야

[0001] 본 발명은 문서에 대한 유사도를 이용하여 클러스터링하는 시스템 및 방법에 관한 것으로서, 보다 상세하게는 문서를 형태소 분석하여 명사를 추출하고 추출된 명사로부터 임의의 음절을 추출하여 추출된 음절을 이용하여 텍스트의 유사도를 판단하고 이를 이용해 문서 클러스터링을 하는 한국어 텍스트 클러스터링 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 최근 정보통신 기술의 발전에 따라 기존의 활자 매체가 아닌 전자문서를 통해 다양한 정보가 온라인을 통해 제공되고 있다. 이러한 정보는 다양하게 재생산되어 전자문서를 통한 정보량은 기하급수적으로 늘어나고 있는 추세이다.

[0003] 전자문서가 가지고 있는 정보 중에서 사용자가 찾고자 하는 정보를 찾기 위해 해당 정보를 가진 전자문서를 찾

는 일은 쉬운 일이 아니며, 이를 위해 다양한 방법이 개발되고 현재 사용되고 있다.

[0004] 한국공개특허공보 제10-2011-0058593호(2011.06.01 공개)는 노출도 분석을 이용한 유사문서 분류 장치를 개시하고 있으며, 이러한 분류 장치는 형태소 분석을 이용하여 명사별 출현빈도를 통해 유사문서를 분류하고 있다.

[0005] 하지만 명사별 출현빈도의 경우 영어에 적합한 것으로 상술한 한국어의 특성을 반영하지 못한 단점이 있다.

[0006] 보다 구체적으로 설명하자면 전자문서에 기재된 한국어는 영어와 달리 단어 또는 구를 통한 용어 정의가 어렵고, 정지 단어와 같은 전처리 과정을 통해 쉽게 차원 축소할 수 없다는 차이가 있다.

[0007] 즉, 한국어는 단어를 용어로 정의하기에 '대한민국은'과 같은 띄어쓰기로 단어가 구분되지 않으며, '대한 민국','대한민국' 등 용어의 형태가 일정하지 않는 특징이 있다.

[0008] 이러한 한국어의 특징을 한국공개특허공보 제10-2011-0058593호(2011.06.01 공개)의 유사문서 분류 장치는 반영하지 못하고 있다는 것이다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 한국공개특허공보 제10-2011-0058593호(2011.06.01 공개)

발명의 내용

해결하려는 과제

[0010] 따라서, 이러한 문제점을 해결하기 위한 본 발명의 첫 번째 목적은 한국어의 특징을 반영하여 문서의 유사도를 판단하고, 일정 기준 값 이상의 문서를 클러스터링하는 한국어 텍스트 클러스터링 시스템을 제공하는 것이다.

[0011] 또한, 두 번째 목적은 한국어의 특징을 반영하여 문서의 유사도를 판단하고, 일정 기준 값 이상의 문서를 클러스터링하는 한국어 텍스트 클러스터링 방법을 제공하는 것이다.

과제의 해결 수단

[0012] 상기 첫 번째 목적을 달성하기 위하여 본 발명은 형태소 분석을 이용하여 텍스트의 명사를 추출하는 명사 추출부, 상기 명사 추출부가 추출한 명사에 대하여 사용자의 의도에 따라 선택된 개수의 음절을 추출하는 음절 추출부, 상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 유사도 판단부, 상기 유사도 판단부가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 상기 텍스트를 클러스터링하는 클러스터링부를 포함하는 한국어 텍스트 클러스터링 시스템을 제공한다.

[0013] 상기 유사도 판단부는 상기 텍스트와 상기 음절을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축하는 것을 특징으로 할 수 있다.

[0014] 상기 유사도 판단부는 코사인 유사도를 이용하여 상기 텍스트의 유사도를 판단하는 것을 특징으로 할 수 있다.

[0015] 상기 클러스터링부는 특이값 분해(SVD, Singular Value Decomposition) 또는 비음수행렬인수분해(NMF, Non-negative Matrix Factorization) 중 어느 하나를 이용하여 상기 텍스트를 클러스터링하는 것을 특징으로 할 수 있다.

[0016] 상기 두 번째 목적을 달성하기 위하여 본 발명은 명사 추출부가 형태소 분석을 이용하여 텍스트의 명사를 추출하는 단계, 음절 추출부가 상기 명사 추출부가 추출한 명사에 대하여 사용자의 의도에 따라 선택된 개수의 음절을 추출하는 단계, 유사도 판단부가 상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 단계, 클러스터링부가 상기 유사도 판단부가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 상기 텍스트를 클러스터링하는 단계를 포함하는 한국어 텍스트 클러스터링 방법을 제공한다.

[0017] 상기 유사도 판단부가 상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 단계는 상기 유사도 판단부가 상기 텍스트와 상기 음절을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축하는 단계를 포함할 수 있다.

[0018] 상기 유사도 판단부가 상기 음절 추출부가 추출한 음절을 이용하여 상기 텍스트의 유사도를 판단하는 단계는 상기 유사도 판단부가 코사인 유사도를 이용하여 상기 텍스트의 유사도를 판단하는 단계를 포함할 수 있다.

[0019] 상기 클러스터링부가 상기 유사도 판단부가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 상기 텍스트를 클러스터링하는 단계는 상기 클러스터링부가 특이값 분해(SVD, Singular Value Decomposition) 또는 비음수행렬인수분해(NMF, Non-negative Matrix Factorization) 중 어느 하나를 이용하여 상기 텍스트를 클러스터링 하는 단계를 포함할 수 있다.

발명의 효과

[0020] 상기에서 설명한 본 발명의 한국어 텍스트 클러스터링 시스템 및 방법에 의하면, 유사도의 정확도는 유지하되 기존에 명사만을 비교하여 유사도를 판단하여 클러스터링하는 방법에 비하여 클러스터링에 사용되는 메모리가 현저하게 감소하는 효과가 있다.

도면의 간단한 설명

[0021] 도 1은 본 발명의 일 실시예인 한국어 텍스트 클러스터링 시스템의 개략적인 구성도이다.

도 2는 본 발명의 일 실시예인 한국어 텍스트 클러스터링 방법의 개략적인 흐름도이다.

도 3은 본 발명의 일 실시예에 따를 때 문서의 수에 따른 TF-IDF 행렬 차원의 크기를 나타낸 그래프이다.

도 4는 본 발명의 일 실시예에 따를 때 명사를 추출한 때 행렬의 값을 채도로 나타낸 그래프이다.

도 5는 본 발명의 일 실시예에 따를 때 명사 앞 한 음절을 추출한 때 행렬의 값을 채도로 나타낸 그래프이다.

도 6은 본 발명의 일 실시예에 따를 때 명사 앞 두 음절을 추출한 때 행렬의 값을 채도로 나타낸 그래프이다.

도 7은 본 발명의 일 실시예에 따를 때 명사 앞 세 음절을 추출한 때 행렬의 값을 채도로 나타낸 그래프이다.

도 8은 본 발명의 일 실시예에 따를 때 명사를 추출한 때 2개의 문서에 대한 조인트 플롯이다.

도 9는 본 발명의 일 실시예에 따를 때 명사 앞 한 음절을 추출한 때 2개의 문서에 대한 조인트 플롯이다.

도 10은 본 발명의 일 실시예에 따를 때 명사 앞 두 음절을 추출한 때 2개의 문서에 대한 조인트 플롯이다.

도 11은 본 발명의 일 실시예에 따를 때 명사 앞 세 음절을 추출한 때 2개의 문서에 대한 조인트 플롯이다.

발명을 실시하기 위한 구체적인 내용

[0022] 본 명세서 및 청구범위에 사용된 용어나 단어는 통상적이거나 사전적인 의미로 한정 해석되지 아니하며, 발명자는 그 사용자의 발명을 가장 최선의 방법으로 설명하기 위해 용어의 개념을 적절하게 정의할 수 있다는 원칙에 입각하여 본 발명의 기술적 사상에 부합하는 의미와 개념으로 해석되어야만 한다.

[0023] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0024] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 포함한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "…부", "…기", "모듈", "장치" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 및/또는 소프트웨어의 결합으로 구현될 수 있다.

[0025] 본 발명에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

[0026] 도 1은 본 발명의 일 실시예인 한국어 텍스트 클러스터링 시스템의 개략적인 구성도이다.

[0027] 도 1을 참고하면, 한국어 텍스트 클러스터링 시스템은 명사 추출부, 음절 추출부, 유사도 판단부 및 클러스터링부를 포함한다.

[0028] 명사 추출부(100)는 형태소 분석을 이용하여 텍스트의 명사를 추출한다.

[0029] 여기서 형태소 분석은 문장을 분해 가능한 최소한의 단위로 분리하는 것으로 일반적인 기술이다.

[0030] 음절 추출부(200)는 명사 추출부(100)가 추출한 명사에 대하여 사용자의 의도에 따라 선택된 개수의 음절을 추출한다.

[0031] 여기서 사용자의 의도에 따라 선택된 개수의 음절은 N 개로서, N은 1 이상을 의미한다.

[0032] 유사도 판단부(300)는 음절 추출부(200)가 추출한 음절을 이용하여 텍스트의 유사도를 판단한다.

[0033] 이러한 유사도 판단을 위해 벡터 공간 모델을 이용하며, 벡터 공간 모델에서 텍스트는 용어(term)에 대한 벡터로 표현될 수 있다.

[0034] 보다 구체적으로 유사도 판단부(300)는 텍스트와 음절 추출부(200)가 추출한 음절을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축할 수 있다.

[0035] TF-IDF(Term Frequency - Inverse Document Frequency)는 용어 빈도(Term Frequency 이하 'TF')와 역문서빈도(Inverse Document Frequency 이하 'IDF')를 사용하는 방법을 말한다.

[0036] TF는 특정한 용어가 문서 내에 얼마나 자주 등장하는지를 나타내는 값으로, 이 값이 높을수록 문서에서 중요하다고 생각할 수 있다.

[0037] 하지만, 용어 자체가 문서군 내에서 자주 사용되는 경우, 이것은 그 용어가 흔하게 등장한다는 것을 의미한다. 이것을 DF(문서 빈도, document frequency)라고 하며, 이 값의 역수를 IDF 라고 하며, TF-IDF는 TF와 IDF를 곱한 값이다.

[0038] 보다 구체적으로, TF는 아래와 같은 수학식 1로 표현될 수 있으며, IDF는 수학식 2로 표현될 수 있다.

수학식 1

[0039] $$TF(t,d) = f_{t,d}$$

[0040] 수학식 1에서 t는 용어, d는 문서를 말한다.

수학식 2

[0041] $$IDF(t,D) = \log \frac{n}{|\{d \in D : t \in d\}|}$$

[0042] 수학식 2에서 n은 문서 전체의 개수이고, 분모는 용어 t가 포함된 문서의 개수이다. 또한, d 는 문서, D는 문서의 집합을 말한다.

[0043] 보다 구체적으로, TF와 IDF는 아래의 수학식 3, 4와 같이 표현될 수 있다.

수학식 3

[0044] $$\mathrm{TF} = \frac{(\text{문서 내 특정 용어의 개수})}{(\text{문서 내 모든 용어의 개수})}$$

수학식 4

$$IDF = \frac{(문서\ 전체\ 개수)}{(특정\ 용어를\ 포함한\ 문서의\ 개수)}$$

[0045]

[0046] 즉, 예를 들어 문서 전체 개수는 500개이고, 문서 내 모든 용어의 개수가 1000개이며, 사용자가 원하는 특정 용어를 포함한 문서의 개수가 4개이고, 사용자가 원하는 특정 용어의 개수가 3개라면 TF는 3/1000 이며 IDF는 500/4 이 되는 것이다.

[0047] 그리고 TF-IDF는 수학식 5와 같이 표현될 수 있다.

수학식 5

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

[0048]

[0049] 상술한 TF, IDF 및 TF-IDF를 이용하여 유사도 판단부(300)는 텍스트와 음절 추출부(200)가 추출한 음절을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축할 수 있다.

[0050] 또한, 유사도 판단부(300)는 코사인 유사도를 이용하여 텍스트의 유사도를 판단할 수 있다.

[0051] 보다 구체적으로 코사인 유사도는 내적 공간의 두 벡터 간 각도의 코사인 값을 이용하여 측정된 벡터 간의 유사한 정도를 의미한다.

[0052] 이러한 코사인 유사도는 아래와 같은 수학식 6으로 표현될 수 있다.

수학식 6

$$\text{코사인 유사도} = \cos(\theta) = \frac{d_k \cdot d_j}{\|d_k\| \|d_j\|}$$

[0053]

[0054] 클러스터링부(400)는 유사도 판단부(300)가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 텍스트를 클러스터링한다.

[0055] 보다 구체적으로, 클러스터링부(400)는 유사도 판단부(300)가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 특이값 분해(SVD, Singular Value Decomposition) 또는 비음수행렬인수분해(NMF, Non-negative Matrix Factorization) 중 어느 하나를 이용하여 텍스트를 클러스터링할 수 있다.

[0056] 도 2는 본 발명의 일 실시예인 한국어 텍스트 클러스터링 방법의 개략적인 흐름도이다.

[0057] 도 2를 참고하면, 명사 추출부(100)가 형태소 분석을 이용하여 텍스트의 명사를 추출한다.(S230)

[0058] 음절 추출부(200)가 명사 추출부(100)가 추출한 명사에 대하여 사용자의 의도에 따라 선택된 개수의 음절을 추출한다.(S231)

[0059] 유사도 판단부(300)가 음절 추출부(200)가 추출한 음절을 이용하여 텍스트의 유사도를 판단한다.(S232)

[0060] 보다 구체적으로 유사도 판단부(300)는 텍스트와 음절을 이용하여 TF-IDF(Term Frequency - Inverse Document Frequency) 행렬을 구축할 수 있다.

[0061] 그리고 유사도 판단부(300)는 코사인 유사도를 이용하여 텍스트의 유사도를 판단할 수 있다.

[0062] 유사도 판단부(300)가 판단한 텍스트의 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에는 S234 단계를 수행하고, 기준 값 미만인 때에는 S231 단계를 수행할 수 있다.(S233)

[0063] 클러스터링부(400)는 유사도 판단부(300)가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 텍스트를 클러스터링한다.(S234)

[0064] 보다 구체적으로 클러스터링부(400)는 유사도 판단부(300)가 판단한 유사도가 사용자의 의도에 따라 정해진 기준 값 이상인 때에 특이값 분해(SVD, Singular Value Decomposition) 또는 비음수행렬인수분해(NMF, Non-negative Matrix Factorization) 중 어느 하나를 이용하여 텍스트를 클러스터링할 수 있다.

[0065] 상술한 본 발명의 일 실시예인 한국어 텍스트 클러스터링 시스템 또는 본 발명의 또 다른 실시예인 한국어 텍스트 클러스터링 방법을 이용하여 유사도 판단을 하였을 때 결과를 하기 설명하겠다.

[0066] 텍스트의 클러스터링을 위한 데이터 집합은 NTCIR-6 CLIR test collection(Chosunilbo(A))을 사용하였다.

[0067] 또한 한국어 형태소 분석을 위해서 https://bitbucket.org/eunjeon/mecab-ko/ 의 mecab-ko를 이용하여 기존의 방식인 명사를 통한 유사도 판단하는 것과 본 발명의 일 실시예인 명사로부터 임의의 음절을 추출해 유사도 판단하는 것을 비교하였다.

실 시 예 1

[0069] TF-IDF 행렬은 하기 표 1과 같은 기준으로 구축될 수 있다.

표 1

[0070]

	행	열
Noun	문서 d_i	명사
Syllable-1	문서 d_i	명사 앞 한 음절
Syllable-2	문서 d_i	명사 앞 두 음절
Syllable-3	문서 d_i	명사 앞 세 음절

[0071] 표 1에 대한 그래프는 도 3과 같다.

[0072] 도 3은 본 발명의 일 실시예에 따를 때 문서 개수에 따른 TF-IDF행렬 차원의 크기를 나타낸 그래프이다.

[0073] 보다 구체적으로 도 3은 늘어나는 문서 수(x축, The number of documents)에 대해 증가하는 행렬의 크기를 나타낸다.

[0074] 일반적으로 문서의 수가 증가할수록 구분되는 용어의 수도 증가하기 때문에 행렬의 크기는 급격하게 커진다.

[0075] 따라서, 명사를 추출한 때 행렬의 크기, 즉 구분되는 용어의 수가 가장 큰 값을 가지게 된다.

[0076] 명사 앞 세 음절을 추출한 때는 명사를 추출한 때와 유사한 행렬의 크기를 갖는다.

[0077] 명사 앞 두 음절을 추출한 때와 명사 앞 한 음절을 추출한 때는 행렬의 크기가 급격하게 줄어든 것을 확인할 수 있는데, 이를 근거로 하여 본 발명에 의할 때 문서의 유사도를 판단하는 때에 소모되는 메모리가 절반 이상으로 줄어드는 효과가 있음을 알 수 있다.

실 시 예 2

[0079] 도 4는 본 발명의 일 실시예에 따를 때 명사를 추출한 때 행렬의 값을 채도로 나타낸 그래프이며, 도 5는 본 발명의 일 실시예에 따를 때 명사 앞 한 음절을 추출한 때 행렬의 값을 채도로 나타낸 그래프이고, 도 6은 본 발명의 일 실시예에 따를 때 명사 앞 두 음절을 추출한 때 행렬의 값을 채도로 나타낸 그래프이며, 도 7은 본 발명의 일 실시예에 따를 때 명사 앞 세 음절을 추출한 때 행렬의 값을 채도로 나타낸 그래프이다.

[0080] 행렬의 값에 대한 채도로써 유사함을 판단하는 방법은 해당 분야에서 일반적인 방법이다.

[0081] 도 4 내지 도 7을 참고하면, 명사를 추출한 때의 행렬의 값에 대한 채도와 명사 앞 두 음절을 추출한 때의 행렬의 값에 대한 채도 및 명사 앞 세 음절을 추출한 때의 행렬의 값에 대한 채도는 유사한 것을 알 수 있다.

[0082] 따라서, 명사를 추출한 때의 행렬과 명사 앞 두 음절을 추출한 때의 행렬 및 명사 앞 세 음절을 추출한 때의 행렬은 유사함을 알 수 있다.

실 시 예 3

[0084] 도 8은 본 발명의 일 실시예에 따를 때 명사를 추출한 때 2개의 문서에 대한 조인트 플롯이며, 도 9는 본 발명의 일 실시예에 따를 때 명사 앞 한 음절을 추출한 때 2개의 문서에 대한 조인트 플롯이고, 도 10은 본 발명의 일 실시예에 따를 때 명사 앞 두 음절을 추출한 때 2개의 문서에 대한 조인트 플롯이며, 도 11은 본 발명의 일 실시예에 따를 때 명사 앞 세 음절을 추출한 때 2개의 문서에 대한 조인트 플롯이다.

[0085] 조인트 플롯은 주어진 두 비교 대상의 데이터 분포를 보기 위한 그림을 말한다.

[0086] 그리고 피어슨 상관계수(pearsonr)는 두 변수 간의 관련성을 구하기 위해 계산하는 값이고, 유의확률(p)은 통계적 가설 검정에서 계산하는 확률 값이다.

[0087] 피어슨 상관계수와 유의확률은 해당 분야에서 일반적으로 쓰는 방법이다.

[0088] 보다 구체적으로, 도 8을 참고하면, 명사를 추출한 때 2개의 문서에 대한 조인트 플롯에서 피어슨 상관계수는 -0.056이며, 유의확률은 0.7이다.

[0089] 도 9를 참고하면, 명사 앞 한 음절을 추출한 때 2개의 문서에 대한 조인트 플롯에서 피어슨 상관계수는 0.29이며, 유의확률은 0.049이다.

[0090] 도 10을 참고하면, 명사 앞 두 음절을 추출한 때 2개의 문서에 대한 조인트 플롯에서 피어슨 상관계수는 -0.058이며, 유의확률은 0.69이다.

[0091] 도 11을 참고하면, 명사 앞 세 음절을 추출한 때 2개의 문서에 대한 조인트 플롯에서 피어슨 상관계수는 -0.056이며, 유의확률은 0.7이다.

[0092] 따라서, 조인트 플롯에서 나타난 피어슨 상관계수 및 유의확률을 고려할 때 명사를 추출한 때와 명사 앞 두 음절 및 명사 앞 세 음절을 추출한 때는 데이터 분포가 유사함을 알 수 있다.

[0094] 실시예 1, 실시예 2 및 실시예 3과 관련하여 행렬 노름은 하기 표 2와 같다.

표 2

[0095]

	$\|A_{noun} - A_{Syllable-1}\|$	$\|A_{noun} - A_{Syllable-2}\|$	$\|A_{noun} - A_{Syllable-3}\|$
Frobenius Norm	14.77	0.64	0.06
Nuclear Norm	33.71	2.90	0.19
Inf Norm	17.10	0.90	0.12
-Inf Norm	9.69	0.10	0.0010
max(sum(abs(x), axis=0))	17.10	0.90	0.12
min(sum(abs(x), axis=0))	9.69	0.10	0.0010
2-Norm	14.37	0.42	0.030
Smallest singular value	0.0084	0.00049	7.65e^-22

[0096] 여기서 노름(Norm)은 비교 대상의 차이를 측정하기 위한 방법으로, 명사를 추출한 때의 행렬과 임의의 음절을 추출한 때의 행렬, 즉 명사 앞 한 음절을 추출한 때의 행렬, 명사 앞 두 음절을 추출한 때의 행렬 및 명사 앞 세 음절을 추출한 때의 행렬 간의 차이를 계산하기 위해 사용하였다.

[0097] 계산된 값이 0에 가까울수록 비교 대상의 차이가 작다는 것을 의미한다.

[0098] 상기 표 2를 참고하면, 명사 앞 두 음절을 추출한 때의 행렬 A_syllable-2 및 명사 앞 세 음절을 추출한 때의 행렬 A_syllable-3 을 명사를 추출한 때의 행렬A_noun과 비교하였을 때 그 차이가 없음을 알 수 있다.

[0099] 상술한 결과를 고려할 때 실시예 1을 통해서 명사를 추출하여 문서의 유사도를 판단하는 경우와 비교하였을 때 명사 앞 한 음절을 추출하여 문서의 유사도를 판단하는 경우 및 명사 앞 두 음절을 추출하여 문서의 유사도를 판단하는 경우는 문서의 유사도 판단 과정에서 소모되는 메모리의 감소 효과가 있음을 알 수 있다.

[0100] 그리고 실시예 2, 실시예 3 및 표 2를 통해서 명사를 추출하여 문서의 유사도를 판단하는 경우와 비교하였을 때

명사 앞 두 음절을 추출하여 문서의 유사도를 판단하는 경우 및 명사 앞 세 음절을 추출하여 문서의 유사도를 판단하는 경우는 정확도가 유사하다는 것을 알 수 있다.

[0101] 그러므로 본 발명에 의할 때 사용자의 의도에 따라 문서의 유사도 판단 시 사용되는 메모리의 소모량을 감소시킬 것인지, 문서의 유사도 판단의 정확도를 유지시킬지 선택할 수 있는 장점이 있다.

[0102] 또한, 사용자의 선택에 의하여 명사 앞 두 음절을 추출하여 문서의 유사도를 판단하는 경우에 메모리 소모량의 감소와 더불어 문서의 유사도 판단의 정확도를 유지할 수 있는 효과가 있다.

[0103] 상기와 같이 본 발명의 실시예에 따른 한국어 텍스트 클러스터링 시스템 및 방법의 구성 및 동작이 이루어질 수 있으며, 한편 상기 본 발명의 설명에서는 구체적인 실시예에 관해 설명하였으나 여러 가지 변형이 본 발명의 범위를 벗어나지 않고 실시될 수 있다.

[0104] 이상에서 본 발명은 비록 한정된 실시예와 도면에 의해 설명되었으나, 본 발명은 이것에 의해 한정되지 않으며 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에 의해 다양한 수정 및 변형이 가능함은 물론이다.

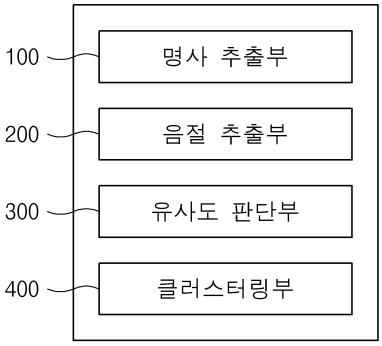
부호의 설명

[0105] 100: 명사 추출부 200: 음절 추출부

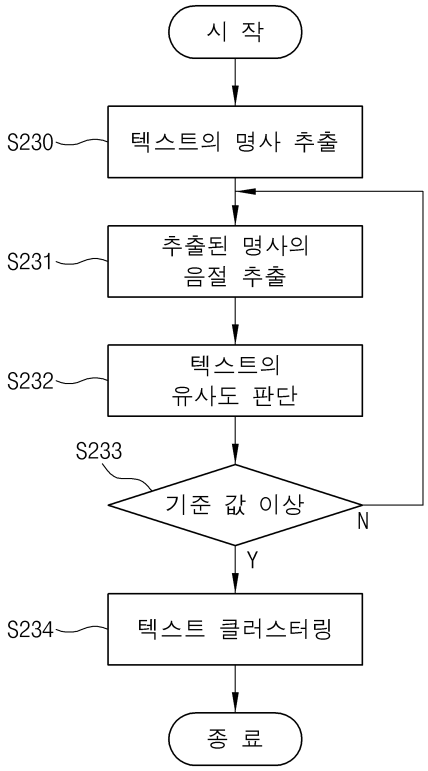
300: 유사도 판단부 400: 클러스터링부

도면

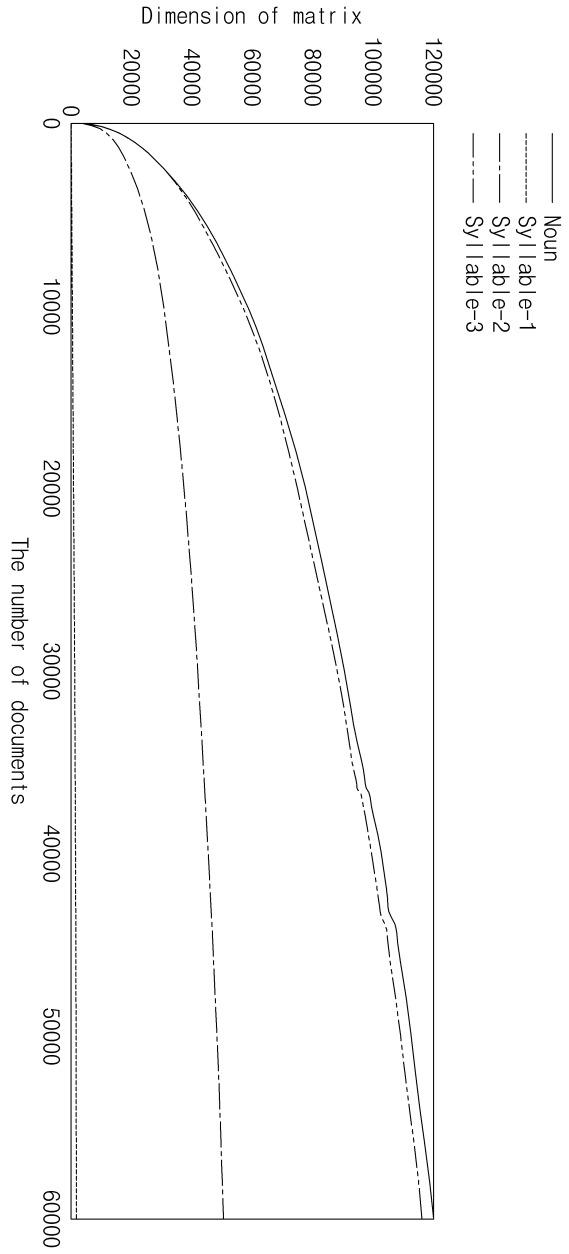
도면1



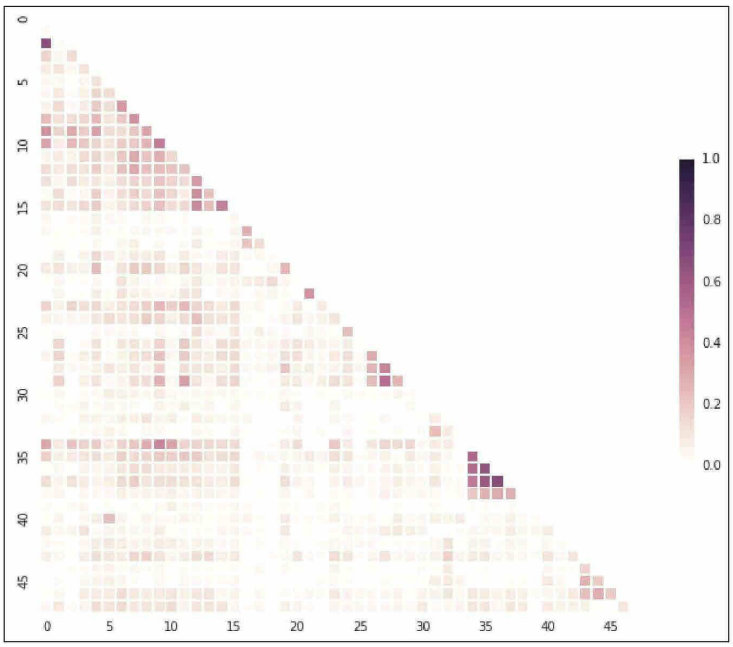
도면2


시 작
S230
텍스트의 명사 추출
S231
추출된 명사의
음절 추출
S232
텍스트의
유사도 판단
S233
기준 값 이상
N
Y
S234
텍스트 클러스터링
종 료

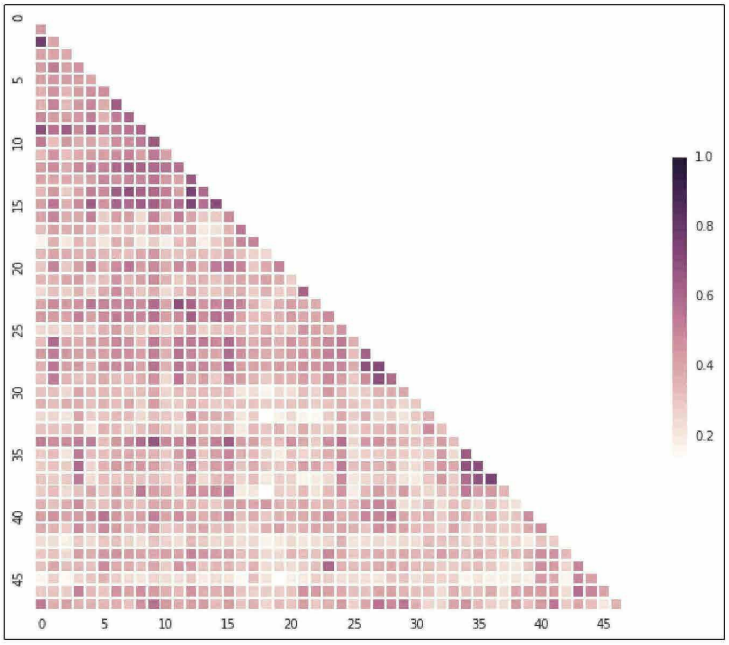
도면3


Dimension of matrix
0
20000
40000
60000
80000
100000
120000
0
10000
20000
30000
40000
50000
60000
The number of documents
Noun
Syllable-1
Syllable-2
Syllable-3

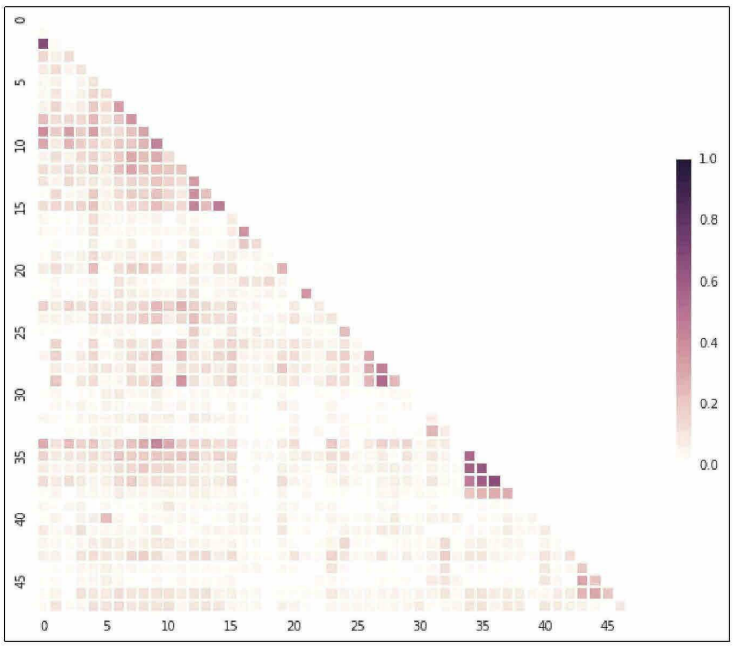
도면4


0
5
10
15
20
25
30
35
40
45
1.0
0.8
0.6
0.4
0.2
0.0

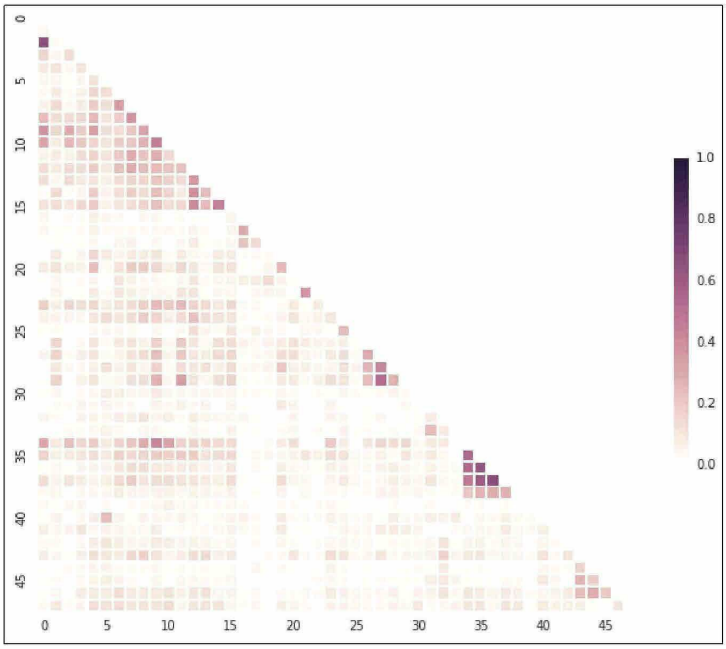
도면5


0
5
10
15
20
25
30
35
40
45
1.0
0.8
0.6
0.4
0.2

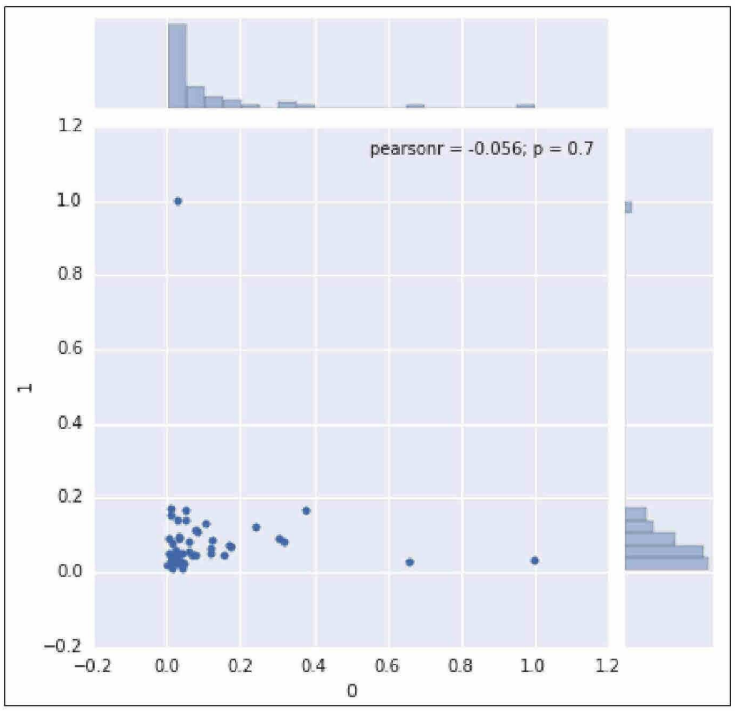
도면6


0
5
10
15
20
25
30
35
40
45
1.0
0.8
0.6
0.4
0.2
0.0

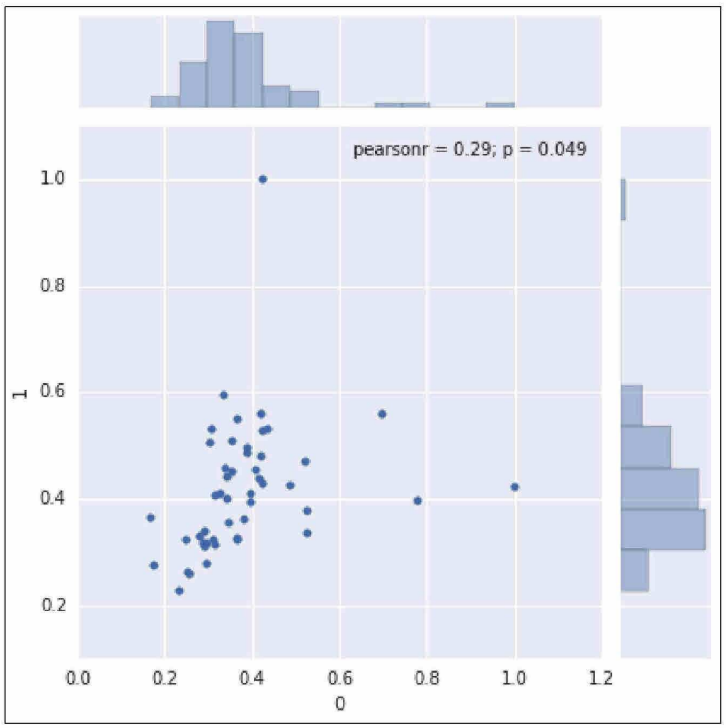
도면7


0
5
10
15
20
25
30
35
40
45
1.0
0.8
0.6
0.4
0.2
0.0

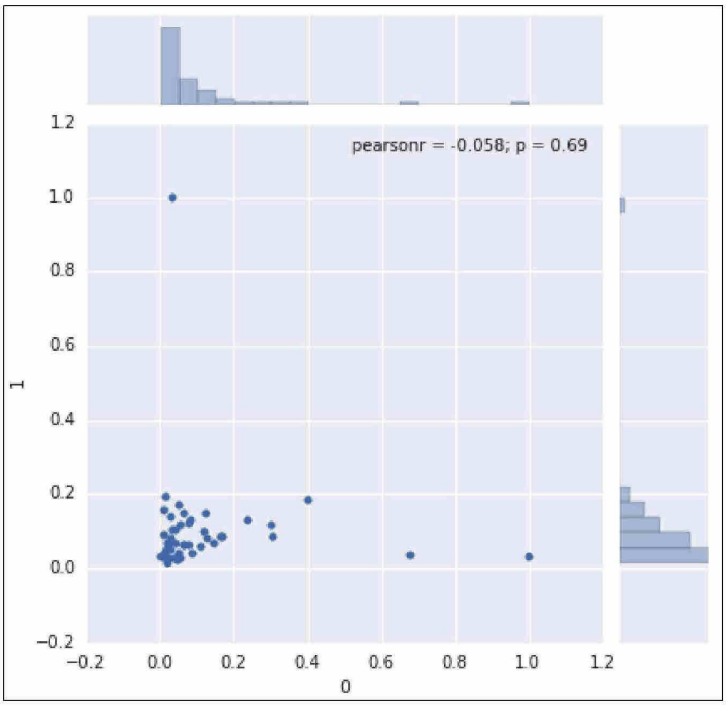
도면8


pearsonr = -0.056; p = 0.7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
1
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0

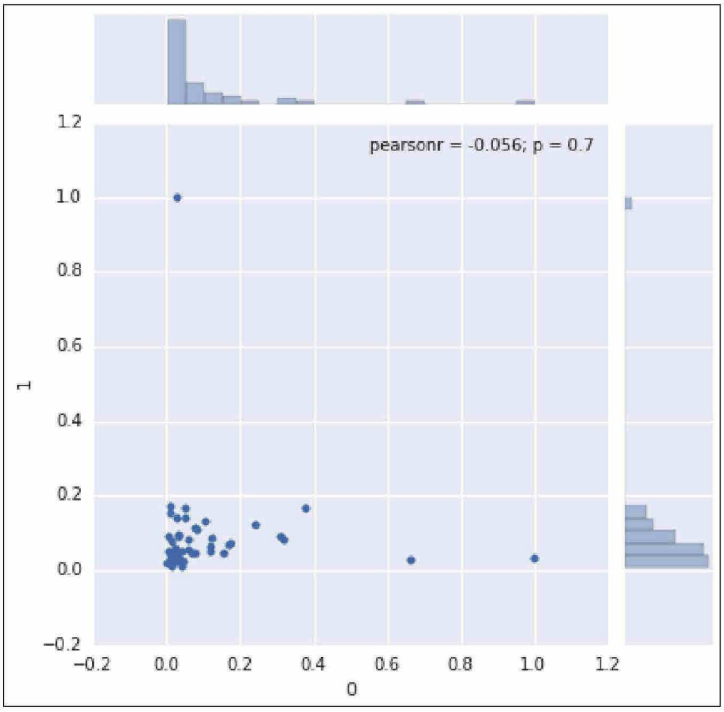
도면9


pearsonr = 0.29; p = 0.049
1.0
0.8
0.6
0.4
0.2
1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0

도면10


pearsonr = -0.058; p = 0.69
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
1
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0

도면11


pearsonr = -0.056; p = 0.7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
1
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0