통계 기반 기법 개선하기

2019.07.09

이름	신정아
E-mail	jeongah.h.shin@gmail.com

PMI와 PPMI

통계 기반 기법 개선하기

PMI(Pointwise Mutual Information) 점별 상호정보량

x와 y가 corpus 내에 동시에 등장할 확률

x와 y가 corpus 내에 동시에 등장하는 횟수

$$\mathrm{PMI}(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{\frac{C(x,y)}{N}}{\frac{C(x)}{N}\frac{C(y)}{N}} = \log_2 \frac{C(x,y)\cdot N}{C(x)C(y)}$$

x가 corpus 내에 등장할 확률

x가 corpus 내에 등장하는 횟수

전체 corpus 갯수

0이 되어버린다면?

PPMI(Positive PMI)

PMI와 PPMI

통계 기반 기법 개선하기

PPMI(Positive PMI)　양의 상호정보량

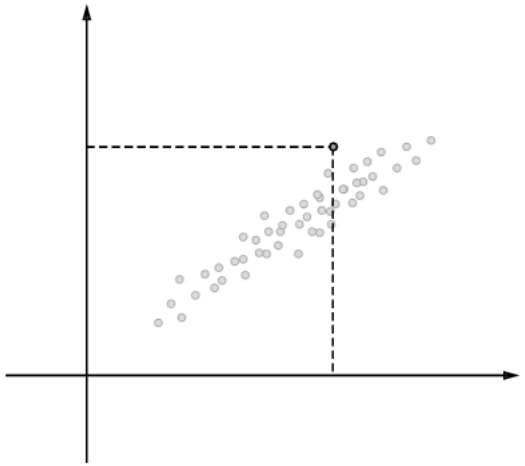
```
def ppmi(C, verbose=False, eps=1e-8):
    M = np.zeros_like(C, dtype=np.float32)
    N = np.sum(C)
    S = np.sum(C, axis=0)
    total = C.shape[0] * C.shape[1]
    cnt = 0

    for i in range(C.shape[0]):
        for j in range(C.shape[1]):
            pmi = np.log2(C[i, j] * N / (S[j]*S[i]) + eps)
            M[i, j] = max(0, pmi)

            if verbose:
                cnt += 1
                if cnt % (total//100) == 0:
                    print('%.1f%% 완료' % (100*cnt/total))
    return M
```

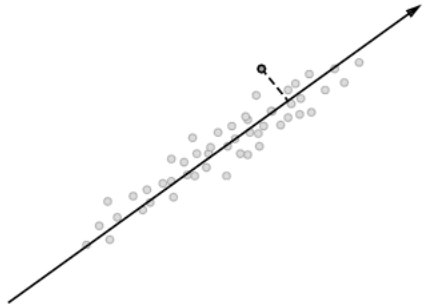
pmi < 0 라면 0값이 대입되게 됨.

차원 감소와 SVD

통계 기반 기법 개선하기

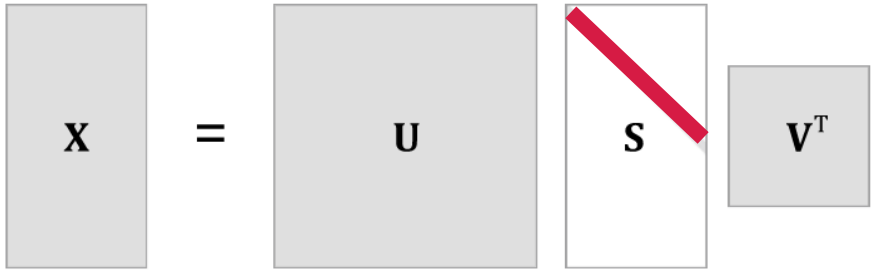
2차원 기준(2개의 축 기준) 분포를 1차원(1개의 축 기준) 기준 분포로!





기준 축은 어떻게 선정?

SVD(Single Value Decomposition)



특잇값(singular value)이 큰 순서대로 나열된 대각 행렬

$A\ {}^tA = {}^tA\ A = I$인 행렬,

$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$이라 두면,

$$A\ {}^tA = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I$$

U, V는 직교 행렬 (전치행렬을 곱했을 때, 단위행렬이 되는 경우)

차원 감소와 SVD

통계 기반 기법 개선하기

특잇값(singular value)이 작다면 중요도가 낮다는 뜻

차원 감소와 SVD

통계 기반 기법 개선하기

numpy.linalg.svd

numpy.linalg.**svd**(*a*, *full_matrices=True*, *compute_uv=True*) [source]

Singular Value Decomposition.

When *a* is a 2D array, it is factorized as `u @ np.diag(s) @ vh = (u * s) @ vh`, where *u* and *vh* are 2D unitary arrays and *s* is a 1D array of *a*'s singular values. When *a* is higher-dimensional, SVD is applied in stacked mode as explained below.

Parameters:

a : *(..., M, N) array_like*

A real or complex array with `a.ndim >= 2`.

full_matrices : *bool, optional*

If True (default), *u* and *vh* have the shapes `(..., M, M)` and `(..., N, N)`, respectively. Otherwise, the shapes are `(..., M, K)` and `(..., K, N)`, respectively, where `K = min(M, N)`.

compute_uv : *bool, optional*

Whether or not to compute *u* and *vh* in addition to *s*. True by default.

Returns:

u : *{ (..., M, M), (..., M, K) } array*

Unitary array(s). The first `a.ndim - 2` dimensions have the same size as those of the input *a*. The size of the last two dimensions depends on the value of *full_matrices*. Only returned when *compute_uv* is True.

s : *(..., K) array*

Vector(s) with the singular values, within each vector sorted in descending order. The first `a.ndim - 2` dimensions have the same size as those of the input *a*.

vh : *{ (..., N, N), (..., K, N) } array*

Unitary array(s). The first `a.ndim - 2` dimensions have the same size as those of the input *a*. The size of the last two dimensions depends on the value of *full_matrices*. Only returned when *compute_uv* is True.

```
U, S, V = np.linalg.svd(W)
```

PTB 데이터셋

통계 기반 기법 개선하기

펜 트리뱅크(Penn Treebank, PTB)

너무 크지도 않고 적당한(?) 말뭉치, 주어진 기법의 품질을 측정 하는 데에 빈번하게 쓰임.

http://www.fit.vutbr.cz/~imikolov/rnnlm/simple-examples.tgz