|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | global\_JJ | classic\_JJ | ancient\_JJ | liberal\_JJ |  |
| politician\_NN | 0 | 5 | 0 | 3 | 8 |
| agenda\_NN | 1 | 1 | 1 | 4 | 7 |
| conservative\_NN | 0 | 4 | 0 | 1 | 5 |
| liberal\_NN | 1 | 6 | 0 | 1 | 8 |

 2 16 1 9 |28

N = 2 + 16 + 1 + 9 = 28

**Two ways of looking at PMI**

1. PMI( I, j) = Log2 ( P( I, j )/( P( I ) \* P( j ) ) )

P( I , j ) = P( I ) \* P( j | I ). Chain Rule

P( i , j ) / P( i ) = P( j | i )

PMI( i , j ) = Log2 ( P( i, j ) / ( P ( i ) \* P ( j ) ) )

2. PMI ( i , j). = Log2 ( P ( j | i ) / P ( j ) )

---------------------------------------------------------------

For word vectors:

PMI(c = i, w = j) = Log2 ( P(w = j | c = i) / P(w = j))

In the example below, we use this version.

**PMI(c = i, w = j) = Log2 ( P(w = j, c = i) / (P(c = i) \*P(w = j)))**

**Part One**

**The Three terms in the PMI formula**

1. P( w = agenda\_NN , c = global\_JJ )

1 / 28

 2. P( w = agenda\_NN )

7/28

 3. P( c = global\_JJ )

2/ 28

4. PMI( w = agenda\_NN, c = global\_JJ ) =

Log2 ( P( w = agenda\_NN , c = global\_JJ )/ (P( w = agenda\_NN ) \* P( c = global\_JJ )). =

Log2 ( 1/28/(7/28 \* 2/28) ) =

Log2 (1/28 \* ((28\* 28)/(7 \* 2)). = (1/28) \* (28 \* 2)) = Log2 (2.0)

Log2 (2.0) = 1.0

In Python

>>> from math import log

>>> log(2.)

0.6931471805599453

>>> log(4.)

1.3862943611198906

>>> log(8.)

2.0794415416798357

>>> log(2)/log(2)

1.0

>>> log(4)/log(2)

2.0

>>> log(8)/log(2)

3.0

>>> log(2,2)

1.0

>>> log(8,2)

3.0

**Part Two**

**Count vectors**

cos ( agenda\_NN, liberal\_NN ). =

 (agenda\_NN\_vec **DOT** liberal\_NN\_vec)/ (| agenda\_NN\_vec | \* (| liberal\_NN\_vec |)

agenda\_NN\_vec = (1, 1, 1, 4)

liberal\_NN\_vec =. (1, 6, 0, 1)

| agenda\_NN\_vec |. = sqrt\_of ((1 \* 1) + (1 \* 1) + (1 \* 1) + (4 \* 4)) = sqrt\_of(19) = 4.36

|liberal\_NN\_vec| =. sqrt\_of((1\* 1) + (6 \* 6) + (0 \* 0) + (1 \* 1)) = sqrt\_of(38) = 6.16

agenda\_NN\_vec **DOT** liberal\_NN\_vec) =. (1 \*1) + (1 \* 6) + (1 \* 0) + (4 \* 1) = 11

cos ( agenda\_NN, liberal\_NN ). = 11/(4.36 \* 6.16). = .410

PPMI Vectors

See the Word Vector Python notebook