Mutual information $I(X;Y)$

Two random variables X, Y are **independent**
iff their joint distribution is equal to the product of their individual distributions:

$$p(X, Y) = p(X)p(Y)$$

That is, for all outcomes x, y:

$$p(X=x, Y=x) = p(X=x)p(Y=y)$$

$I(X;Y)$, the **mutual information** of two random variables X and Y is defined as

$$I(X;Y) = \sum_{X,Y} p(X=x, Y=y) \log \frac{p(X=x, Y=y)}{p(X=x)p(Y=y)}$$

Pointwise mutual information (PMI)

Recall that two **events** x, y are **independent** if their joint probability is equal to the product of their individual probabilities:

x,y are independent iff $\mathrm{p(x,y) = p(x)p(y)}$

x,y are independent iff $\mathrm{p(x,y) / p(x)p(y) = 1}$

In NLP, we often use the pointwise mutual information (PMI) of two outcomes/events (e.g. words):

$$PMI(x,y) = \log \frac{p(X=x, Y=y)}{p(X=x)p(Y=y)}$$

Using PMI to cluster words

Using PMI to find related words

Find pairs of words w_i, w_j that have high pointwise mutual information:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

Different ways of defining $p(w_i, w_j)$
give different answers.

Using PMI to find "sticky pairs"

$p(w_i, w_j)$: probability that w_i, w_j are adjacent

Define $p(w_i, w_j) = p(\text{“}w_i w_j\text{”})$

High PMI word pairs under this definition:

Humpty Dumpty, Klux Klan, Ku Klux, Tse Tung,
avant garde, gizzard shad, Bobby Orr, mutatis mutandis,
Taj Mahal, Pontius Pilate, ammonium nitrate,
jiggery pokery, anciens combattants, fuddle duddle,
helter skelter, mumbo jumbo
(and a few more)

Using PMI to find "semantic clusters"

$p(w_i, w_j)$: probability that w_i, w_j are near each other

Define $p(w_i, w_j) = p(w_i)p(w_j \mid w_i)$, and $p(w_j \mid w_i)$ as the probability of w_j occurring within a window around w_i

(window = 500 words to the left or right of w_i, excluding two words before w_i and two words after w_i)

Resulting word clusters:

- *we our us ourselves ours*
- *question questions asking answer answers answering*
- *tie jacket suit*
- *attorney counsel trial court judge*
- *morning noon evening night nights midnight bed*
- *wall ceiling walls enclosure roof*
- *sell buy selling buying sold*