Associations

TM the Great Unread

DTL|Digital Arts Initiative
Interacting Minds Centre|Aarhus University

July 27, 2016

Associations between words

“Thou shall know a word by the company it keeps” (Firth, 1975)

Words acquire meaning from their mutual dependencies to or co-occurrence with other words (distributional semantics)

```
[1] "it hath been said thou shalt love thy neighbour and hate
thine enemy but i say unto you love your enemies"
[2] "no man can serve two masters for either he will hate the
one and love the other or else he will"
```

Identify words that collocate with a specific probe or node word beyond chance level

Used for tracking consumer preferences, text retrieval, recommender systems, OCR ...

Pointwise Mutual Information

PMI is still the industry standard for measuring word associations

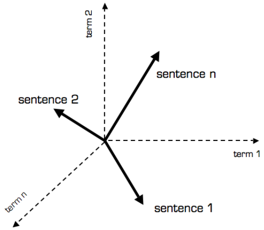
Compares the probability of observing two words together (their joint probability) with the probabilities of observing them independently (chance level)

$$PMI = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

An association is when the joint probability is larger than chance: $P(w_1, w_2) > P(w_1)P(w_2)$.
The words are independent when the joint probability is at chance level: $P(w_1, w_2) = P(w_1)P(w_2)$.
PMI can be negative when the co-occurrence is smaller than expected: $P(w_1, w_2) < P(w_1)P(w_2)$

Distance between word vectors

What does it take for two words in a document term matrix to be related?



Distance between documents: $Dist(d_1, d_2) = \sqrt{\sum_{i=1}^{n}(w_{i,d_2} - w_{i,d_1})^2}$

Distance between words: $Dist(w_1, w_2) = \sqrt{\sum_{i=1}^{n}(d_{i,w_2} - d_{i,w_1})^2}$