Text Similarity

Thesaurus-based Word Similarity Methods
Quiz

• Which pair of words exhibits the greatest similarity?
  – 1. Deer–elk
  – 2. Deer–horse
  – 3. Deer–mouse
  – 4. Deer–roof
Quiz Answer

• Which pair of words exhibits the greatest similarity?
  – 1. Deer–elk
  – 2. Deer–horse
  – 3. Deer–mouse
  – 4. Deer–roof

• Why?
Remember Wordnet

- **ungulate**
  - even-toed ungulate
    - ruminant
      - okapi
      - deer
      - giraffe
    - elk
    - wapiti
    - caribou
  - odd-toed ungulate
    - mule
    - horse
      - pony
    - zebra
- equine
Path Similarity

• Version 1
  – \( \text{Sim} (v,w) = - \text{pathlength} (v,w) \)

• Version 2
  – \( \text{Sim} (v,w) = - \log \text{pathlength} (v,w) \)
Problems with this Approach

- There may be no tree for the specific domain or language
- A specific word (e.g., a term or a proper noun) may not be in any tree
- IS–A (hyponym) edges are not all equally apart in similarity space
Path similarity between two words

• Version 3 (Philip Resnik)
  \[
  \text{Sim} (v,w) = - \log P(\text{LCS}(v,w))
  \]
  where LCS = lowest common subsumer, e.g.
  ungulate for deer and horse
  deer for deer and elk
Information content

• Version 4 (Dekang Lin)
  – Wordnet augmented with probabilities (Lin 1998)
  – $IC(c) = -\log P(c)$
  – $Sim(v,w) = 2 \times \log \frac{P(LCS(v,w))}{\log P(v) + \log P(w)}$

\[
\text{sim}(\text{Hill, Coast}) = \frac{2 \times \log P(\text{Geological-Formation})}{\log P(\text{Hill}) + \log P(\text{Coast})} = 0.59
\]
Wordnet Similarity in NLTK

- NLTK
  
  >>> dog.lin_similarity(cat, brown_ic)
  0.879
  >>> dog.lin_similarity(elephant, brown_ic)
  0.531
  >>> dog.lin_similarity(elk, brown_ic)
  0.475
NLP