

HW #7

Distributional Semantics

- Goals:
 - Explore distributional semantic models
 - Compare effects of differences in context
 - Evaluate qualitatively & quantitatively

Task

- Construct distributional similarity models
- Use fixed data resources
 - Brown corpus data
- Compare similarity measures under models
- Compare correlation with human judgments

Mechanics

- Corpus Reader

- Loading Brown corpus via NLTK:

```
brown_words = nltk.corpus.brown.words()  
brown_sents = nltk.corpus.brown.sents()
```

- ~1.2M words

- May want to develop on subset
- e.g. `brown_words = brown_words[0:10000]`
 - Caveat: lexical Gaps

Mechanics

- Correlation:
 - `from scipy.stats.stats import spearmanr`
 - `A = spearmanr(list1, list2)`
 - Return correlation coefficient, p-value
`A.correlation`

Use Condor in Development!

- Don't run any non-trivial scripts on the patas head-node
- Lots of fighting for small resource
- Can wind up locking people out
- Use condor!

Details

- Windows:
 - “2” means two words before or after the modeled word
 - The quick brown fox jumped over the lazy dog
- Weights:
 - “FREQ”: straight co-occurrence count (“term frequency”)
 - “PMI”: (positive) point-wise mutual information

(P)PMI

- Positive Pointwise Mutual Information (PPMI)
- Given the tabulated context vectors:

$$PPMI_{ij} = \max(\log_2 \frac{p_{ij}}{p_{i*} \cdot p_{*j}}, 0)$$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^W \cdot \sum_{j=1}^C \cdot f_{ij}} \quad p_{i*} = \frac{\sum_{j=1}^C \cdot f_{ij}}{\sum_{i=1}^W \cdot \sum_{j=1}^C \cdot f_{ij}} \quad p_{*j} = \frac{\sum_{i=1}^W \cdot f_{ij}}{\sum_{i=1}^W \cdot \sum_{j=1}^C \cdot f_{ij}}$$

Word2Vec

- Compare results to (CBOW) word2vec

- Python package **gensim**

```
model = gensim.models.word2vec.Word2Vec(sents,  
    vector_size=100, window=2, min_count=1, workers=1)
```

- Sents is a list of lists of strings

```
model.wv.similarity('man', 'woman')
```