Natural Language Processing
CSE 498 Fall 2017
Project 4
Sachin Joshi

Exercises:

• Describe the benefits of lower casing and lemmatizing words in context and signature.

Lemmatization involves converting the words of the sentence to its dictionary form, which allows us to query the WordNet dictionary for information retrieval. The main advantage of lemmatization is that it takes into consideration the context of the word to determine which is the intended meaning the user is looking for. This process allows to decrease noise and speed up the user's task. Lower casing resolves the issue of case sensitivity, where words such as text and Text are treated as two completely different words.

• Prove that the Jaccard similarity is the ratio of the number of common words to N.

Since the Jaccard similarity uses the min function as the numerator, it is essentially computing the (weighted) number of overlapping features, since if either vector has a zero-association value for an attribute, the result will be zero and hence we can say that it is the number of common words. On the other hand, it uses the max function as the denominator which is viewed as the normalizing factor and consists of all non-repeating words between the vectors which is N. Hence Jaccard similarity is the ratio of the number of common words to N.

• **Prove that both the similarity metrices are symmetric.**

Similarity Metrices are symmetric, i.e.,

$$sim_{cosine}(v, w) = sim_{cosine}(w, v)$$

$$sim_{cosine}(v, w) = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i} \sqrt{\sum_{i=1}^{N} w_i}}$$

$$sim_{cosine}(w, v) = \frac{\sum_{i=1}^{N} w_i \times v_i}{\sqrt{\sum_{i=1}^{N} w_i} \sqrt{\sum_{i=1}^{N} v_i}}$$

$$= \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i} \sqrt{\sum_{i=1}^{N} w_i}}$$

$sim_{cosine}(w, v) = sim_{cosine}(v, w)$

Hence Proved

$$sim_{Jaccard}(v, w) = \frac{\sum_{i=1}^{N} \min\{v_i, w_i\}}{\sum_{i=1}^{N} \max\{v_i, w_i\}}$$

$$= \frac{\sum_{i=1}^{N} \min\{w_i, v_i\}}{\sum_{i=1}^{N} \max\{w_i, v_i\}}$$

$sim_{Jaccard}(v, w) = sim_{Jaccard}(w, v)$