Billy DeLucia
wld217@lehigh.edu
CSE498 Natural Language Processing Project 4

Exercises:

- The benefits of lower case casting and lemmatizing the wordforms found in the context and signature basically reduce the sparsity of the similarity vectors without significant data loss. This is because we are removing irrelevant data such as the conjugation of a verb (drank, drink, drinks => drink) and the capitalization of a word due to placement in a sentence or any other reason (Drink, drink => drink). Neither of these transformations change the meaning of the word drink in the similarity vector. If I can drink beer; then she drinks beer; and either way beer was drank. In all these cases beer is used in the context of drink and the similarity vector should reflect that regardless of conjugation or capitalization.

- Jaccard similarity is defined as the intersection between the two vectors divided by the union of the two vectors. In this case the union will always be N, and the intersection will be the lemmas the vectors have in common. Therefore Jaccard similarity will always be the ratio of the # of common words to N.

- Jaccard:

  $\mathrm{Sim}_{\mathrm{Jaccard}}(v, w) = (v \cap w) / (v \cup w)$

  Given the intersection of two vectors is symmetric:

  $X = v \cap w = w \cap v$

  Given the union of two vectors is symmetric:

  $Y = v \cup w = w \cup v$

  Therefore $\mathrm{Sim}_{\mathrm{Jaccard}}$ is symmetric:

  $\mathrm{Sim}_{\mathrm{Jaccard}}(v, w) = \mathrm{Sim}_{\mathrm{Jaccard}}(w, v) = X / Y$

- Cosine:

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

  Given that summation and multiplication are symmetric functions:

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

  And

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

  Therefore $\mathrm{Sim}_{\mathrm{Cosine}}$ is symmetric:

  $\mathrm{Sim}_{\mathrm{Cosine}}(v, w) = \mathrm{Sim}_{\mathrm{Cosine}}(w, v) = X / Y$