Problem Definition

I have defined the problem of this assignment as the following:

Given a text corpus of a specific category (e.g. news, editorial etc.), identify within the corpus pairs of a verb or an adjective matched with an intensified form by an attachment of an adverb. Assume that adverb only describes an adjective or a verb, it may be attached both before and after the word being described, and it may not be chained to another adverb. Adverbs attached to words intensifies the word in a sense that it specifies the meaning. Verbs and adjectives used more frequently are those that are weaker in intensity than their respective counterparts.

Output Justification

The output is formed by finding the occurrences of a verb or an adjective in the corpus that are most similar to the verb or the adjective that is described by an intensity modifying adverb.

The reasonable quality of the results are achieved through several factors. First, numerous examples of intensity modifying adverbs are covered by generating a set of them from synonyms of "seed" intensity modifying adverbs. Through this method, more adverbs can be covered even if the seed adverbs are not enough. Thus, the seeds are manually provided, but more of them are automatically generated. Also, the synonyms of the word described by an adverb are found accurately from the provided corpus by calculating the path similarity between words in synsets of two words. Words have meanings in different parts of speech and may affect the similarity if it is not taken into account. Limiting the synsets by parts of speech such as verb or adjective improved accuracy of similarity calculation. Because the corpus is of a specific category, it is likely that there are a plenty of synonyms for any verb or adjective in the same corpus. As shown in the results, the pairs between a verb or an adjective plus an adverb and the corresponding verb or adjective have close meanings. This part satisfies the requirement that the adverb should intensify the verb or the adjective while preserving the general meaning. Furthermore, the last factor is how closely the intensity modifying adverb alters the original verb or adjective to the corresponding word. As shown in the code, I compared the frequency of the two words to determine which is the one with "weaker" intensity and thus requires the adverb. I decided that the one used less frequently is the one with stronger intensity because people would use the more general, weaker version more readily. Following such logic, I found pairs such as (skidded completely, sliding) and (slightly damaged, impair), which show one that is used more frequently is weaker and used with an adverb.

However, there was some room for improvement as well. I used a stemmer in nltk to compare the stems of words and ignore those that have same stem even if they had high similarity, because they are likely different tenses of the same verb etc. But since the English language has many exceptions, the stemmer is not perfect. Thus, the results contain pairs such as (begin, began). If I could eliminate such cases, the results could have been more accurate. In addition, I assumed that verbs and adjectives that are attached to an adverb and used more frequently are the weaker ones. This may be true in that adverb specifies the intensity of the weaker word and make it mean something stronger, but the frequency does not always signify the original strength of the word. If another more accurate metric to calculate intensity of a verb or andjective could be found, the results might have been better. Also, adjectives do not work well with WordNet as described in its specifications. The similarities calculated for adjectives are not accurate relative to those of verbs. Finally, the adverbs could be manually assigned values that describes its role: strengthening, weakening or maintaining for better comparison. However, such method that could be automated was not found. Somehow applying such method could have given better results.