# CS-660 Homework 5

Zijing Huang 10414952

**3. Take any review from your dataset, find the top 10 documents most similar to the selected review.** **Analyze why these 10 documents are selected.**

I choose the first review as my selected review.

*This sound track was beautiful! It paints the senery in your mind so well I would recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars and soulful orchestras. It would impress anyone who cares to listen! ^\_^*

Before lemmatization, the top10 similarity is as follows:

[(4, 0.17339273676970235), (2, 0.1687852896961608), (212, 0.15653925288912274), (210, 0.15548278056338183), (69, 0.11769812548559777), (3, 0.1171364917683797), (119, 0.1125800321404632), (5, 0.10799774356351222), (216, 0.1075584236096584), (217, 0.09808556933439538)]

After Comparing the top10 similarity reviews with the selected reviews, I found that they are similar not because they are both positive/negative or they express the same meaning. They are similar because they have similar “weight” of words. If a word appears frequently within a document and seldom appear in other documents, then this word has very high weight, which means the sentence is high depended on this word. So these top10 reviews are selected because the weight of words they have are the closest.

**5. Test the modified function by setting “lemmatized” to True and redo Step 3 to see if lemmatization help you find similar documents better. Also note down the change in size of your tf-idf matrix.**

Yes, lemmatization is to reduce inflectional which improve text matching and reduce feature space. Because it helps to accurate the weight of words, it helps to find similar documents better.

Before lemmatization, the tokens of the first review would be:

['sound', 'track', 'beautiful', 'paints', 'senery', 'mind', 'well', 'would', 'recomend', 'even', 'people', 'hate', 'vid', 'game', 'music', 'played', 'game', 'chrono', 'cross', 'games', 'ever', 'played', 'best', 'music', 'backs', 'away', 'crude', 'keyboarding', 'takes', 'fresher', 'step', 'grate', 'guitars', 'soulful', 'orchestras', 'would', 'impress', 'anyone', 'cares', 'listen', '^\_^'],

After lemmatization, the tokens would be:

['sound', 'track', 'beautiful', u'paint', 'senery', 'mind', 'well', 'would', 'recomend', 'even', 'people', 'hate', 'vid', 'game', 'music', u'play', 'game', 'chrono', 'cross', u'game', 'ever', u'play', 'best', 'music', u'back', 'away', 'crude', 'keyboarding', u'take', u'fresh', 'step', 'grate', u'guitar', 'soulful', u'orchestra', 'would', 'impress', 'anyone', u'care', 'listen', '^\_^']

Before lemmatization, the top10 similarity is as follows:

[(4, 0.17339273676970235), (2, 0.1687852896961608), (212, 0.15653925288912274), (210, 0.15548278056338183), (69, 0.11769812548559777), (3, 0.1171364917683797), (119, 0.1125800321404632), (5, 0.10799774356351222), (216, 0.1075584236096584), (217, 0.09808556933439538)]

After lemmatization, the top10 similarity is as follows:

[(69, 0.2141006635204249), (4, 0.1856263304750434), (195, 0.18053052570390327), (2, 0.17950120095468614), (212, 0.1742544269124907), (216, 0.17099243959703625), (210, 0.16810560204703884), (220, 0.1495523013617911), (71, 0.14619436627508675), (70, 0.139819177088212)]

The size of tf-idf matrix after lemmatization is getting smaller. Because the duplicate of the words with the same lemma would be removed. For example, “say”and “saying” with the same lemma “say”will be just keep as “say”in the token list.