1. Why choose nouns only?

When using word-based model(especially unigram model), the feature words we extracted are usually discrete and it is hard to tell the meaning behind these words without their contexts. In such a situation, I think nouns usually convey more meanings, especially emotional meanings, than verbs. In fact, Classical Chinese poetry is a concise language for depicting images and describing emotions. Imagery can be the spirit of Chinese poems, for example:

“大漠孤烟直，长河落日圆" is wildly known for Chinese, which can be directly translated into: Dessert with lonely smoke, long river under sunset.

or

"枯藤老树昏鸦，小桥流水人家，古道西风瘦马" can be translated as: Withered vines, old trees and drowning crows, small bridges and flowing water near homes, ancient roads with west wind and thin horses(translated from Google Translate)

The nouns can be more intuitive, sometimes even conveying emotions, thus helping us to find the pattern behind our feature words.

2. What is Glove?

Glove(Global Vectors for Word Representation) techique is based on co-occurance matrix, which projects different words into vectors, hoping the vectors represent relationship between words in some aspects. The main idea is to minimize the loss $\sum_i \sum_j (w_i^T w_j + bias_{ij} - logX_{ij})$ where w_i is the word vector and $X_{ij} = \#$ co-occurance of word i and j

If word frequency(in the whole training text) in a word set do not fluctuate too much, it is proper for us to assume $w_i^T w_j$ represents the co-occurance of these two words.

There are two shinning points in the result of Glove: the first one is *Nearest Neighbors*, which means similar words gather together under euclidean distance or cosine similarity. The second one is *Linear Substructures*, such as "king is to queen as man is to woman", which corresponds to $\vec{king} - \vec{man} + \vec{woman} \approx \vec{queen}$.

There is another classic NLP model called **LSA**(Latent semantic analysis) which basically means each document is a distribution of topics and each topic is a distribution of words.

Given these two ideas, though each element of Glove vectors may not convey a certain topic, it is possible that a linear combination of dimensions of word vector in Glove actually represent the idea of peacefulness (or "peaceful" direction).

3. Method: Linear SVM

We used Linear SVM to classify the feature words in peace group and less peace group. If our assumption is correct, the normal vector of our hyperplane actually correspond to the peacefulness direction. We used *glove.840B.300d*(computed from Common Crawl dataset) to project our word, and collected 903 words(461 high-peace words, 442 low-peace words). Since Linear SVM is a linear classification model, according to the *10EPV*(Event per Variable) rule of thumb, we used PCA to reduce the dimensionatilty to 90(76.2% explained variance ratio), and the model accuracy is 83.4%.

4. Feature words(First 50 according to SVM distance)

High-peace Group:

ban, crackdown, marijuana, veto, lawsuit, gay, referendum, takeover, cyclist, protester, debate, ice, firefighter, wildfire, pound, legislation, overhaul, senator, downtown, landmark, blaze, snow, coverage, push, mayor, supermarket, backlash, crash, critic, spill, easing, billionaire, storm, cat, bet, blue, weather, mortgage, forecast, rollout, cheese, chocolate, teammate, apartment, spending, ballot, spokeswoman, abortion, slide, bankruptcy

Low-peace Group:

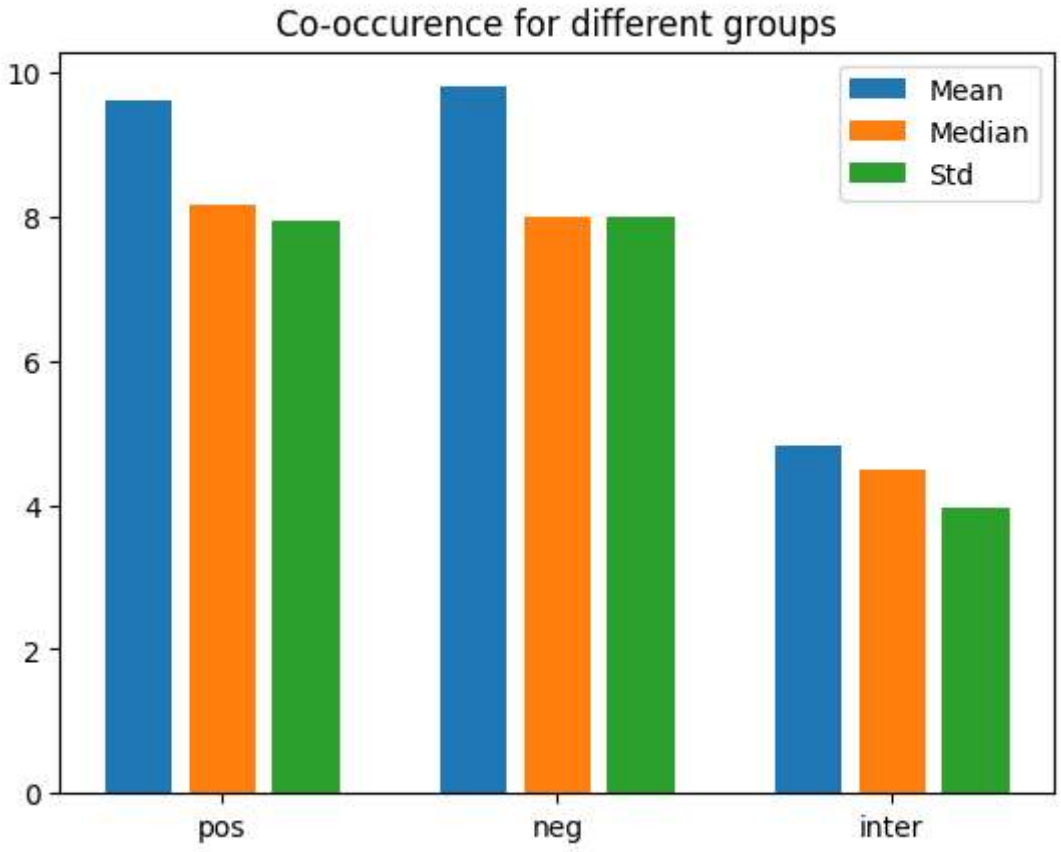
sanitation, prophet, complainant, tribe, empowerment, sin, forgiveness, livelihood, ritual, village, grace, malaria, belonging, labourer, agriculture, virtue, doctrine, realisation, famine, entrepreneurship, wicket, knowledge, mercy, verification, pilgrim, purpose, manufacture, prayer, methodology, governance, assurance, function, verse, receipt, excellence, achievement, cricket, occasion, pray, magistrate, notification, deed, par, sacrifice, dignitary, entity, machinery, delegation, unity, avail

It seems as if high-peace country focus on personal consumption and concrete nouns, while low-peace country focus on religious,government-related, abstract nouns.

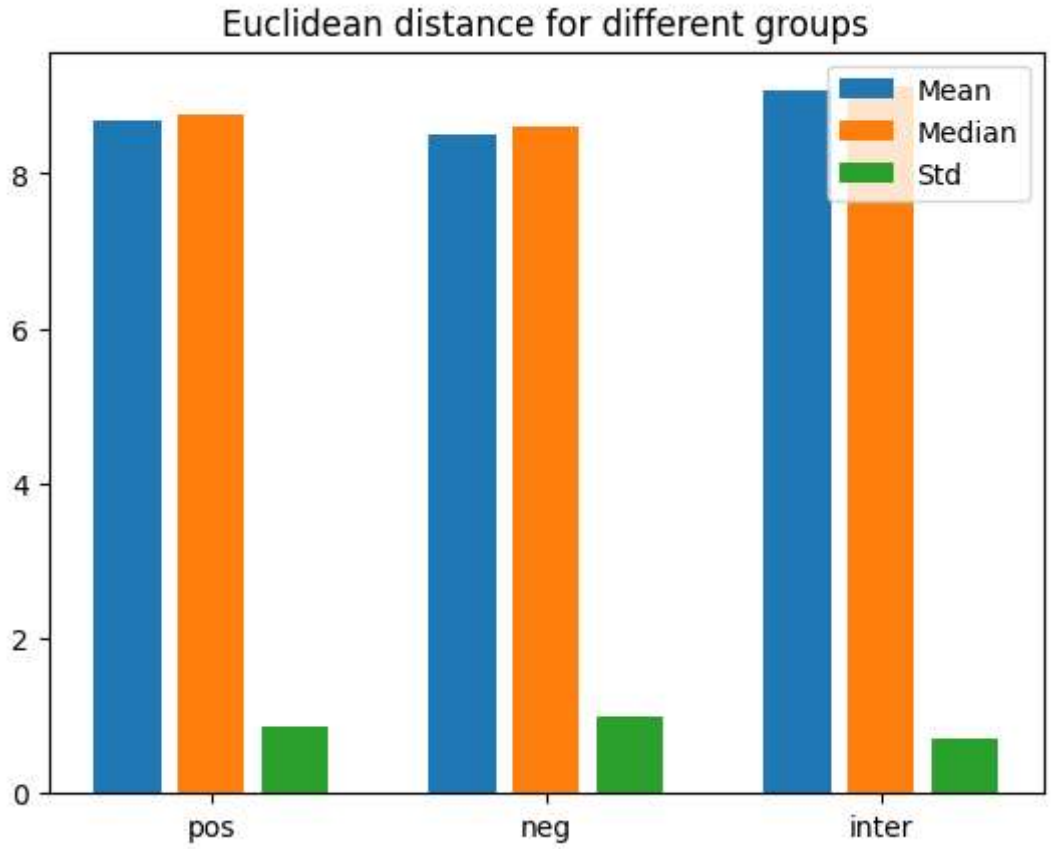
5. Words co-occurance, euclidean distance and cosine similarity

Here we are trying to explore further on our feature words using Glove Embedding. We selected first 50 words collected from SVM distance, calculate words co-occurance, euclidean distance and cosine similarity inside and between groups(called pos,neg,inter group).

For the co-occurance,



For euclidean distance,



For cosine similarity,