Contents


```
vocab : 1 * 125166
bigram : 1 * 1578919

Stop words -- remove some symbolized words (first letter cap)(Microsoft)
-- have some threshold for feature matrix (some bigger frequency)
-- chunk all reviews together and remove <?? occurance words (traning only)
-- this can also remove the dimension matrix

Stop words for bigram
-- remove bigrams that both words are stop words or symbolized words

Look into:
LSA : Latent semantic analysis - HIGH
SentiwordNet and other senti dictionaries -- rank the emotional words
- MEDIUM
Wordnet: Algorithm to find senti levels - LOW

Possible features:
tf.idf for both unigram and bigram
build word/bigram feature space for training & rating label, calculate
cosine similarity
user and dates, usefulness
give larger weights of all capitalized words

hist(double(vertcat(test.word_count)))

-- most word occurs less than 20 (some )

-- hist(cellfun(@length,words),100)

-- intersection kernel
```

1