# CSCI 375 HW 4

Tongyu Zhou

October 12 2017

## 1 Written Exercise [30 points]

1.

P(POS$_{i-2}$ = VERB|s)

Sense 1: 1
Sense 2: $\frac{1}{3}$
Sense 3: 0

P(POS$_{i-2}$ = NOUN|s)

Sense 1: 0
Sense 2: $\frac{2}{3}$
Sense 3: 0

P(POS$_{i-2}$ = NUM|s)

Sense 1: 0
Sense 2: 0
Sense 3: 1

2.

a)

$$|log(\frac{\frac{1}{3}}{\frac{1}{2}})| = 0.1761$$

b)

$$|log(\frac{\frac{1}{6}}{\frac{1}{3}})| \approx 0.3010$$

## 2 Programming Exercise [70 points]

```
Co-occurence features with words only
Accuracy: 0.3218390804597701
Most Informative Features
                      of = True           bank~1 : bank~1 =      7.2 : 1.0
                      in = True           bank~1 : bank~1 =      6.3 : 1.0
                       a = True           bank~1 : bank~1 =      5.9 : 1.0
                      or = True           bank~1 : bank~1 =      5.5 : 1.0
                     and = True           bank~1 : bank~1 =      4.1 : 1.0
Co-occurence features with words + POS
Accuracy: 0.3448275862068966
Most Informative Features
                      of = True           bank~1 : bank~1 =      7.2 : 1.0
                      DT = False          bank~1 : bank~1 =      6.0 : 1.0
                       a = True           bank~1 : bank~1 =      5.9 : 1.0
                       . = False          bank~1 : bank~1 =      5.0 : 1.0
                      DT = True                U : bank~1 =      3.5 : 1.0
Collocational features with words only
Accuracy: 0.42528735632183906
Most Informative Features
                  pos_0 = ' the'          bank~1 : bank~1 =      7.0 : 1.0
                  pos_2 = ' ;'            bank~1 : bank~1 =      6.8 : 1.0
                  pos_1 = ' the'          bank~1 : bank~1 =      5.5 : 1.0
                  pos_0 = ' on'           bank~1 : bank~1 =      5.3 : 1.0
                  pos_1 = ' of'           bank~1 : bank~1 =      5.1 : 1.0
Collocational features with words + POS
Accuracy: 0.39080459770114945
Most Informative Features
                  pos_1 = 'WRB'           bank~1 : bank~1 =     10.2 : 1.0
                  pos_5 = 'JJ'            bank~1 : bank~1 =      8.9 : 1.0
                  pos_1 = 'VBG'           bank~1 : bank~1 =      8.5 : 1.0
                  pos_1 = 'DT'            bank~1 : bank~1 =      7.3 : 1.0
                  pos_0 = ' the'          bank~1 : bank~1 =      7.0 : 1.0
Both features with words only
Accuracy: 0.40229885057471265
Most Informative Features
                     of = True            bank~1 : bank~1 =      7.2 : 1.0
                  pos_0 = ' the'          bank~1 : bank~1 =      7.0 : 1.0
                  pos_2 = ' ;'            bank~1 : bank~1 =      6.8 : 1.0
                     in = True            bank~1 : bank~1 =      6.3 : 1.0
                      a = True            bank~1 : bank~1 =      5.9 : 1.0
Both features with words + POS
Accuracy: 0.39080459770114945
Most Informative Features
                  pos_1 = 'WRB'           bank~1 : bank~1 =     10.2 : 1.0
                  pos_5 = 'JJ'            bank~1 : bank~1 =      8.9 : 1.0
                  pos_1 = 'VBG'           bank~1 : bank~1 =      8.5 : 1.0
                  pos_1 = 'DT'            bank~1 : bank~1 =      7.3 : 1.0
                     of = True            bank~1 : bank~1 =      7.2 : 1.0
```

Co-occurrence w/ words only: 0.50(precision), 0.13(recall), 0.20(f1)
Co-occurrence w/ words + POS: 0.34(precision), 0.67(recall), 0.25(f1)
Collocational w/ words only: 0.67(precision), 0.50(recall), 0.57(f1)
Collocational w/ words + POS: 0.71(precision), 0.63(recall), 0.67(f1)
Both w/ words only: 0.67(precision), 0.50(recall), 0.57(f1)
Both w/ words + POS: 0.75(precision), 0.75(recall), 0.75(f1)

In order best classifiers, the following rank can be made: Both (words+POS), Collocational(words+POS), Both/Collocational(words), Co-occurrence(words+POS), Co-occurrence(words). In general, it seems that collocational features that record specific information about the positions of each context respective to the target word is most effective in producing better results. This makes sense, intuitively, as the context of a word is more determined by its particular location in a set of other words than what other words appear with it (and since the binary feature vector for co-occurrence used a bag of words generated from a large corpus, it is not particularly good at determining sense). Combining both collocational and co-occurrence features do not seem to effect the accuracy that drastically, as a large majority of the most informative features was shared with the other classifiers.

Design choices:
- The window size was set to 2 to avoid overfitting, since that would further lower the accuracy as the algorithm would have too many features to consider and use idiosyncrasies of the training set that don't generalize too well to the testing set.
- Stemming was considered in the process of generating the context vectors, but was removed in the final version as it did not significantly change the accuracy (only improved by 2-3 percent so probably not worth).
- In determining what specific tokens to use for the bag of words for co-occurrence, the top 10 most common words/parts-of-speech tags were chosen (of course, most of these were prepositions such as "a" or "the" so they weren't particularly strong at determining context, hence the low accuracy)

In general, although these evaluation measures seem pretty terrible/sketchy compared to the scores from HW2, which were much higher, this assignment required word sense disambiguation of "bank" from a total of 8 senses instead of just differentiating between movie review/movie plot. In this sense, randomization would generate around 0.125 accuracy instead, so doing Naive Bayes is still comparatively a lot better. To further improve results, alternative classifiers such as weka and svmlight from the nltk package can be considered.