Sentiment Analysis with Bayesian Learning

Dekel Viner s2612925

October 2016

1 Stemming vs Lemmatization

1.1

You do not need a POS tagger to use a stemmer but you do in order to use a lemmatizer.

1.2

The stemmer converted all the form of the word operate into "oper" the lemmatizer on the other hand did not change most of the words because the lemmatizer really requires a POS tagger. The stemmer then considers all of the operate words to belong to the same word type however operate and operative are completely different, operational could also be different. By grouping them together a lot of information gets lost regarding their meaning. The lemmatizer retains all the information, besides dealing with plurals as those tend to be really easy to lemmatize.

1.3

the lemmatize considered "better"(as adj) to be "good". But I would not want to group those together as "better" requires a relationship between two entities while "good" does not.

2 WordNet

2.1

13 different noun synsets.
18 different verb synsets

2.2

synset: call.v.11:
'stop or postpone because of adverse conditions, such as bad weather'. This

was not in the cambridge dictionary. The most similar definition was "to decide on". But the definition given by wordnet is more specific.

synset: call.v.05:
'order, request, or command to come'. This was in the Cambridge dictionary under the definition of "ask someone to come".

synset: call.v.20:
'challenge (somebody) to make good on a statement; charge with or censure for an offense'. This was not in the cambridge dictionary and there was also nothing similar.

2.3

['designate.v.01', 'label.v.01', 'name.v.01']

2.4

for synset(call.v.03): [Synset('dial.v.01')]

2.5

The lowest common hypernym among nurse, doctor and caregiver is 'health professional'. Coach's lowest common hypernym with nurse,doctor and caregiver is 'person'. Clearly doctor, caregiver and nurse are more similar to each other. This is pretty much what we expected, we could note however that doctor has a longer path from the root synsets than nurse, and nurse has a longer path than caregiver.

2.6

coach is most similar to caregiver(0.1428571429) and least similar to doctor(0.1111111111).

3 Sentiment Analysis

3.1

both of the sentences seem positive to me. In the first sentence the word fair makes the sentence seem positive to me. It means that the sound quality was not dissapointing, if the word fair was replaced by another like bad or disappointing then the sentence would be negative. In the second sentence the the word cool makes the sentence positive to me. If the word cool was replaced by another adjective like bad or ugly then the sentence would be negative.

3.2

synset('fair.a.01'); Definition:
'free from favoritism or self-interest or bias or deception; conforming with established standards or rules'
fair.a.01: PosScore=0.625 NegScore=0.0

synset('bright.a.01'); Definition:
'emitting or reflecting light readily or in large amounts'
bright.a.01: PosScore=0.125 NegScore=0.0

synset('cool.a.06'); Definition:
'fashionable and attractive at the time; often skilled or socially adept'
cool.s.06: PosScore=0.375 NegScore=0.0

3.3

I would think that cool would have a higher PosScore than fair and personally i do not think that bright should have a positive or a negative connotation.

3.4

I only found one synset of brassy and it was brassy.s.02, which means 'tastelessly showy'. It is obviously different since it has a negative connotation. brassy.s.02: PosScore=0.0 NegScore=0.5

4 Bayesian Learning

4.1

We cutoff the data at 3/4 for training and 1/4 for testing.

4.2

Train on 1500 instances, test on 500 instances: accuracy: 0.728
Train on 1000, test on 1000: accuracy: 0.811
Train on 1800, test on 200: accuracy: 0.73
When looking at the most importing feature, simply based on intuition I would have to say that the classifiers with the 9/10 and 3/4 training sets seem to have the most descriptive features, for example the half-half classifier seem to have whatsoever, fictional and abilities which i would not think to be necessarily positive or negative. However it did tend to have the highest accuracy after a number of tests.

Table 1: classifying using different size training and testing sets

train(9/10)test(1/10)		train(1/2)test(1/2)		train(3/4)test(1/4)	
word	pos:neg	word	pos:neg	word	pos:neg
outstanding	15.6:1.0	whatsoever	1.0:14.3	magnificent	15.0:1.0
ludicrous	1.0:14.2	inept	1.0:13.7	outstanding	13.6:1.0
avoids	12.3:1.0	stupidity	1.0:11.7	insulting	1.0:13.6
astounding	12.3:1.0	nomination	11.7:1.0	vulnerable	13.0:1.0
idiotic	1.0:11.8	ludicrous	1.0:11.7	ludicrous	1.0:11.7
atrocious	1.0:11.7	vulnerable	11.3:1.0	uninvolving	1.0:11.7
offbeat	11.0:1.0	views	11.3:1.0	avoids	11.7:1.0
fascination	11.0:1.0	fictional	9.7:1.0	astounding	10.3:1.0
animators	10.3:1.0	exceptional	9.7:1.0	fascination	10.3:1.0
symbol	10.3:1.0	abilities	1.0:9.7	idiotinc	1.0:9.8

4.3

I used the classifier with the 1/2 training set and 1/2 test set as it had the highest accuracy.
Positive review text : 'this is an amazing fucking movie'.

Negative review text : 'this movie sucked terribly'