Sentiment Analysis

CSCI5180 - Techniques for Data Mining

Date 2010.12.20

Overview

- Preprocessing
- Analysis
- Discussion
- Conclusion

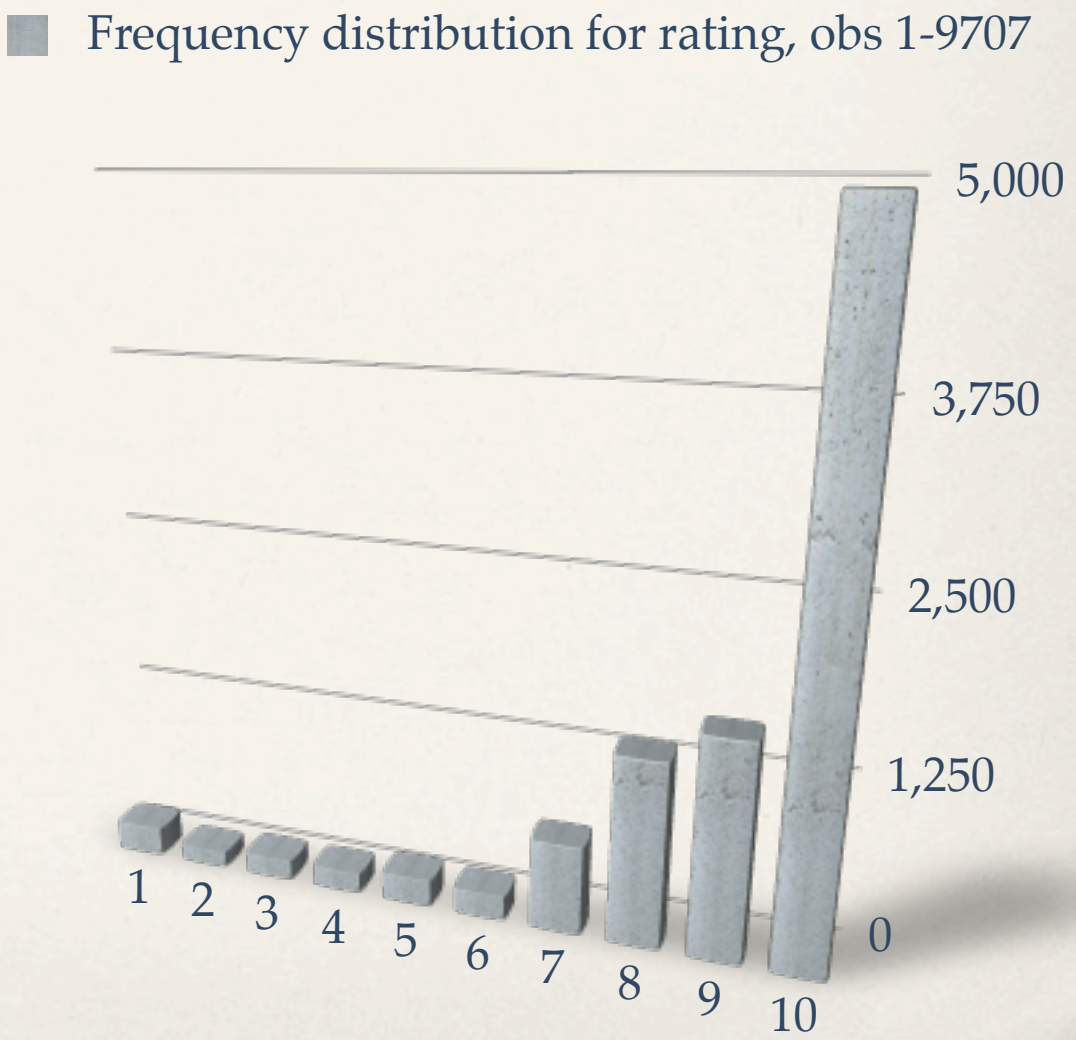
Preprocessing

- Data retrieval from
- Cleaning data

Number of occurrences	Word
144655	the
85009	and
71638	to
70766	a
69071	of
52134	is
39977	in
34675	that
32271	I
30685	it
25182	this
23149	for
21318	with
20179	was
....	....

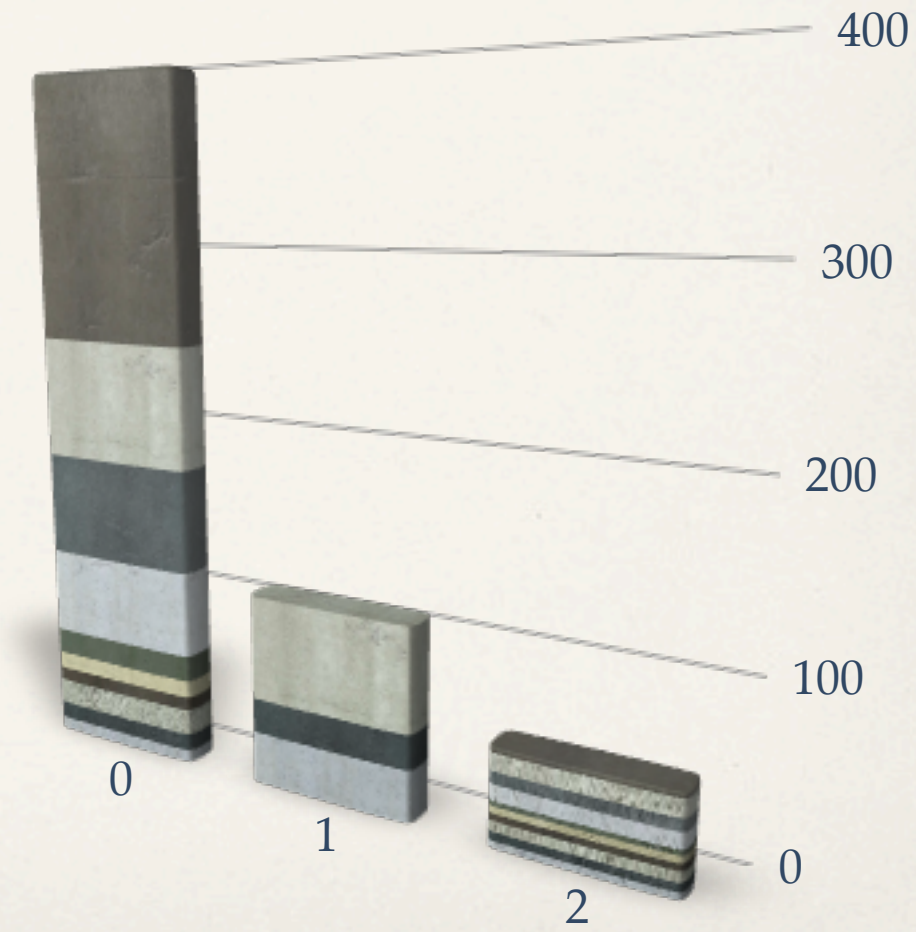
Analysis - biased data

- Biased data set
  - Under sampling class majority
  - Eliminate imbalanced data set



Analysis - choosing candidates

- Attribute selection and candidate words
- Occurrences of the word ‘excellent’



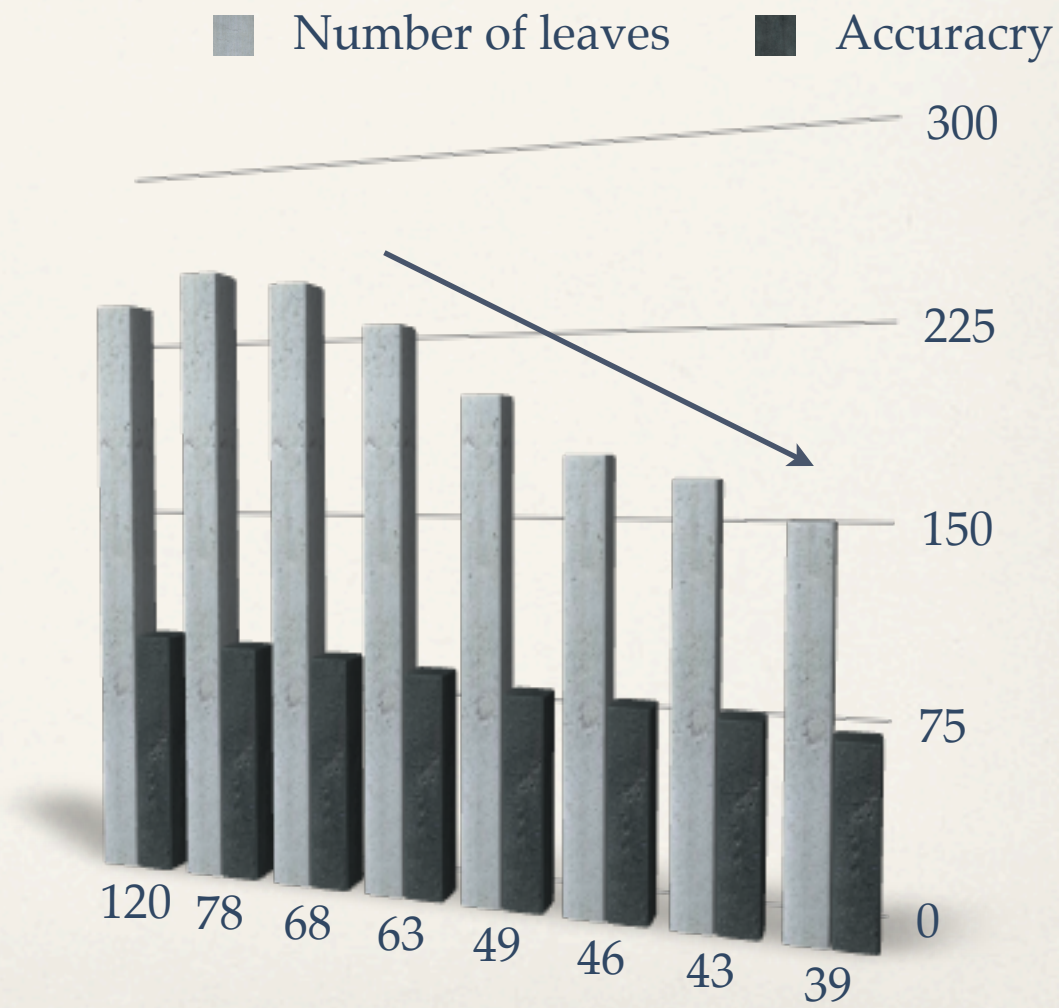
Analysis - candidate words

- Candidate list consisting of 43 emotional words

Positive word candidate	Negative word candidate
amazing	bad
beautiful	bizarre
cleaver	boring
cool	brutal
enjoy	crazy
excellent	cry
fantastic	disappointed
funny	disturbing
hilarious	dry
incredible	gay
intellectual	stupid
romantic	uncomfortable
talented	weak
terrific	wrong
....	....

Discussion

- Decrease in candidate words results in decreased accuracy
- Decision tree for classification



Conclusion

- The nature of natural language problem (negation, double negation, metaphors, synonyms, irony)
- Bag-of-words approach - candidate words consisting of positive and negative emotional emphasis
- Under-sampling

Questions