Semantic Orientation Applied to Unsupervised Classification

Shengxin Qian
ECE 590 Text Analysis

September 23, 2016

1 Pseudo-Code

1). Part-of-speech tagging, select all two-word phrases which match specific patterns in data set

2). For phrase_i in all selected two-word phrases:

$$SO(phrase_i) = \frac{hits\ of\ (phrase_i\ NEAR\ "excellent") * hits\ of\ "poor"}{hits\ of\ (phrase_i\ NEAR\ "poor") * hits\ of\ "excellent"} \tag{1}$$

"hits" represents : # of results search engine returns

("phrase1" NEAR "phrase2") represents : using search engine operator to search documents that contain phrase1 within ten words of phrase2 in either order (e.g. AROUND(10) operator used on Google)

3). Calculate decision statistics of input text:

$$confidence\ of\ input\ text = \frac{\sum_{i=1}^{N} SO(phrase_i)}{N},\ N\ =\ \#\ of\ selected\ phrases \tag{2}$$

4). Make a decision:

$$Decision\ rule : \begin{cases} input\ text\ is\ recommended & confidence > 0 \\ input\ text\ is\ not\ recommended & confidence \leq 0 \end{cases} \tag{3}$$