

FEATURE-BASED OPINION MINING AND RANKING

Purpose of the project

- Since the evolution of social networks, people have started to express their opinions in the form of the blogs or facebook posts or tweets starting from the products people buy to the presidential candidate they support.
- When searched for a particular product on the web, the current day search engines show the list of websites which gives the features of the product and their prices.
- But, the users can be given much more info about the product using the reviews about the searched product.
- Thus, a new type of search engine can be designed which will not only retrieve facts, but will also enable the retrieval of opinions of the users about the product.

Purpose of the project (cont)

- Such a search engine can be used in a number of diverse applications like product reviews to aggregating opinions on a political candidate or issue.
- We create a search engine which, when searched for a product, gives the highlighted features of the product and some of the helpful reviews (instead of the user scrolling through all the reviews to know about the product).

Opinion Mining and Ranking Algorithms

- Basically, two algorithms are used for our purpose:
 - The 1st algorithm (HAC) identifies and extracts the potential features from the reviews of the product.
 - The 2nd algorithm takes these potential features as input, assigns scores to them and finally helps in classifying every review as positive, negative or neutral. The scores obtained for every feature of the product can be used to highlight the good features of the product.

1. The High Adjective Count Algorithm (HAC)

- Instead of using frequency of keywords, our algorithm starts identifying adjectives and nouns.
- The scores of nouns are initialized to zero, each adjective is associated with a noun to which it is closest, this adjective is more likely to describe the noun.
- For each such adjective, score of noun is increased by one. After processing all the reviews, we will have a score associated with each noun, which we call them as opinion scores.
- So nouns with high score have more adjectives to describe them. Then we can have a threshold and nouns having score more than threshold are considered as potential features.

2. Max Opinion Score Algorithm

- This Algorithm takes 3 arguments as input:
 - The first argument is the list of adjectives which are used to express opinions, we refer them as opinion words. We have to choose a value manually between $[-4,4]$ to each opinion word. A high score indicates a stronger opinion than lower score.
 - The second argument is the list of inversion words like 'not' which give a negative sense to opinion, so when these words occur in the left context of opinion words, they change the opinion sense. So when an inversion word appears, we multiply the score by -1.
 - The third argument is the list of potential features obtained by using TF / TF-IDF / HAC algorithm.

MOS Algorithm (cont)

- For each sentence, we look at the opinion words and identify features closest to it. The score of feature is summation of scores of opinion words associated with it.
- The scores of features are further summed up to calculate score of review, so for each review we get a score and reviews are displayed based on this score.
- For each review, we score the features separately for title and body.

$$\text{Review score} = \frac{\alpha \cdot \text{Title score} + \text{Body score}}{\alpha + 1}$$

(α is the title weight co-efficient)

Evaluation Criteria

- For evaluating first we calculate the precision of the top-N features:

$$\text{Precision} = \frac{\text{Number of relevant features}}{N}$$

- The relevant features are those which describe and are related to the review of an item.
- Example:
 - In case of camera, 'g3' and 'canon' are irrelevant features because they are brand terms and do not describe the item and are not helpful.

$$\text{Accuracy} = \frac{\text{Number of correct classifications}}{\text{total number of reviews}}$$

- Initially by the title we can assign each review positive or negative depending on the sentiment.

Evaluation Criteria (cont)

- After calculating the review score using High Adjective Count (HAC) it can be seen whether the class is correct from the initial assigned value.
- If the prediction made by HAC differs from what has been assigned depending on the sentiment than the classification is wrong, if it is same than the classification is correct.
- We can also observe that by giving more priority to the review score of the title than the body, the accuracy can be improved.