New Media and Sentiment Workshop – Aspect Based Sentiment Analysis

The objectives of these exercise are:

1. Perform aspect based sentiment mining using heuristics and enhance understanding
2. Modify an existing CRF aspect sentiment mining model to achieve better performance and understand how different (latent) features may affect the results.

Group yourselves into groups of ***3-5*** for the workshop (e.g. your project team). The exercise has 2 parts.

Part 1: (done in groups)

In the notebook 3_1 _Entity and Aspect Mining – Heuristic.ipynb, aspect sentiment mining was performed for reviews of the restaurant Ding Tai Feng, which were obtained from Yelp. In this exercise we will do the same for the airline sentiment tweets from the dataset airline_sentiment.csv.

Below are the steps to take:

i. Identify what aspects are of interest to airlines. For example:
   a. Service: attendants, food, etc.
   b. Experience: entertainment system, seats, luggage, etc.
   c. Price
   d. Flight journey: flight duration, airplane and general aspects, etc.

   The list is suggestive and not exhaustive.

ii. Create a mapping table to associate words with these aspects

iii. Do sentiment analysis by sentence and associate the aspects with the respective sentiments

iv. Create a visualisation to communicate the results.

In the data file, there are tweets for different airlines. Identify which airline has performed the best in which aspect, and further communicate the results. Summarize on the challenges faced and lessons learnt.

Submit your solutions (pdf converted from ipynb) in Part 1 as a group. Add first names of team members to the file name, e.g. “Group1_Part1_James_Amanda_Muthu.pdf”.

Part 2: (done in groups)

In the notebook 3_2_Entity and Aspect Mining with CRF.ipynb, a CRF model is built to identify entity aspects. This works off an IOB data file that includes the word tokens, the POS, and the aspect tags.

We have seen that by using just POS of the current word, the previous word, and the next word, the model doesn't work well, with f1 at 0.46 for B-A, and 0.32 for I-A.

It is possible to increase the performance of the model by adding new features to the model.

Observe the training data file, Restaurants-train.iob. Create different word features, like pre and post word, case information, whether it's a title word, etc., and try to achieve better precision/ recall.

Report your best set of features and how it affects precision / recall.

Does using more feature lead to better recall/ precision?

What are the most informative features that lead to the best model? Does it tell you anything intuitively?

Identify the challenge and difficulties that you face in doing this assignment. Also suggest how the model can be further improved (with more work).

Submit your solutions (pdf converted from ipynb) in Part 2 as a group. Include first names of team members to the file name, e.g. "Group1_Part2_James_Amanda_Muthu.pdf".