# COM6115: Lab Class 6

## Sentiment Analysis of Tweets

This lab is composed of two parts. The first part aims to implement a corpus-based model based on **pointwise mutual information** (**PMI**) to get overall sentiment about US airline companies from a corpus of tweets. The second part aims to implement a gradable method for sentiment analysis.

## Data Description

The dataset is a corpus of tweets about US airline companies. This dataset is a comma-separated file.

- Each sample is composed of a tweet id (first column), followed by the sentiment and the text of the tweet.
- The text has not been tokenized nor lowercased (you might want to preprocess it before use).

The sentiments are expressed on a 3-value scale: positive, neutral and negative. In the following table you can find several example sentences and their sentiment value.

| | | |
|---|---|---|
| 570270684619923457 | I❤flying @VirginAmerica. | positive |
| 569347934866637345 | @SouthwestAir are you hiring for flight attendants right now? | neutral |
| 569960080734490624 | @united I'm rebooked now, but the line was 300 people deep. | negative |

## Pointwise Mutual Information

In [Turney and Littman, 2003], the semantic orientation of a word (whether it has a positive or negative connotation) is obtain by looking whether it co-occurs more with clearly positive words (e.g. `great`, `fantastic`) or negative words (e.g. bad, `wrong`).

In this lab, PMI will be used to measure the polarity of a word and get the sentiment value.

PMI is defined as follows:

$$\text{PMI}(x, y) = log_2\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

The probabilities will be **estimated** by relative frequency using the raw counts:

- $C(x)$: number of tweets containing word $x$.
- $C(y)$: number of tweets containing word $y$.
- $C(x, y)$: number of tweets where $x$ and $y$ co-occur.
- $N$: total number of tweets

## What do people think of US airline companies?

**Roadmap**:

1. Implement some preprocessing steps.
   You are free to add any preprocessing steps (e.g. lowercasing, tokenization) which you think will be helpful[1].

2. Implement the counting code.

**Question 1** What are the most frequent positive and negative words in this dataset?

3. Implement the Pointwise Mutual Information function from scratch.
   You should **not** use an already-implemented function.

**Question 2** What do positive, zero and negative values of PMI mean?

4. Compute the sentiment for the US airline companies listed in the **companies** list.

**Question 3** What can you conclude?

5. Look at the data and update the lists of positive and negative words. See how this impacts the results.

   - You can get help by looking for word lists on the web. For example:
   - positive: https://www.enchantedlearning.com/wordlist/positivewords.shtml
   - negative: https://www.enchantedlearning.com/wordlist/negativewords.shtml

## Sentiment analysis with gradable method

Use the provided tab-separated file `valence_lexicon_small.tsv` to obtain word polarity. The polarity is computed as the average of ratings, each ranging from -4 to +4, obtained from 10 humans.

1. Implement a gradable method to classify the tweets, as presented in Week 7 lecture 2.

2. Compute and display **confusion matrices** with different thresholds to decide between negative/neutral/positive.

3. Compute the accuracy of your method.

### Going further

1. Add handling of negation, strengthening and weakening words. You might want to update the scores in the provided word lists.

2. Add handling of emoticons and exclamations.

[1] You can use the **NLTK** Python library for this.

## Notes and comments

- Consider using the **Pandas** library to load the data https://pandas.pydata.org/.
- You may search the internet for lists of English punctuation and/or stopwords (also called function words) that you may use in this lab.

## References

[Turney and Littman, 2003] Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst., 21(4):315346*.