# Identifying Salient Named Entity of a Tweet

## Phase 1 report
## Team #8

**Tasks done during phase 1:**

1. Collection of tweet dataset
   raw: 6000+ tweets
   classified: Manually extracted Named Entities from 500+ tweets for evaluation purpose

2. Extraction of Named entities using stanford ner tool
**Language used:** Python

3. Extraction of Named entities using GATE tool
**Language used:** Java

4. Extraction of Named entities using custom built tool.
**Language used:** Java
**Approach:**

- Noise is removed from original tweets by removing non ascii characters and some special characters.
- Tweets are tagged with POS using GATE twittie tagger.
- Tagged tweets are processed to identify patterns of Named entities.
  _USR (user eg @username)
  _HT (hash tag eg #felicity)
  _NNP (eg Boehner)
  _NNP+ ( repeated occurrence of NNP _NNP+ _IN _NNP+ ( two sets of NNP with 'of' or 'for' eg Bank of Thailand)
  _NN or _NNS (single occurrence of singular or plural nouns)

5. Building evaluation tool to evaluate the results generated from different approaches. Evaluation tool calculates the precision and recall for different approaches taken.

**Observations:**

Following table shows the precision and recall values obtained for approaches used.

Precision = (correct  + 0.5 * partially_correct)  / (correct + incorrect + partial)
Recall     =  (correct + 0.5 * partially_correct) / (correct + missing + partial)

| **Approach** | **Precision** | **Recall** |
|---|---|---|
| Stanford_ner | 42.94 | 9.99 |
| Gate_ner | 33.55 | 37.98 |
| Custom_ner | 43.42 | 81.53 |