Social Unrest: Classification and Modeling
Dan Saadati, Farah Uraizee, Tariq Patanam

Motivation
As social media rapidly becomes a podium for political opinions and a tool for the organization and facilitation of protests, a powerful stream of data documenting opinions and actions of individuals becomes readily available. This type of information can provide key social insights in predicting areas at risk of social unrest, which can be significantly useful in scenarios prone to violence.

Methodology

SVM + Presence of Each Word:
- Lacked diversity in training data, used entire vocabulary as feature set
- Results: overfitted, very inaccurate

Algorithm
1. Prepare data for feature extraction
2. **Background subtraction**: downweight words common to both unrest and rest situations;
   \[ f_1(w) = f_1(w) - f_2(w) \]
   where \( f_1(w) \) is the frequency of a word \( w \) in social unrest tweets

Results & Analysis

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>SVM Regression Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of Words</td>
<td>0.0 1.000 0.905 0.662</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>0.0 0.614 0.988 0.984</td>
</tr>
<tr>
<td>Bag of Words</td>
<td>0.0 0.761 0.945 0.792</td>
</tr>
</tbody>
</table>

**Considerations**
- Feature set overfits regionally
- Training Word2Vec model with Twitter data to more accurately reflect sentiment similarity

**Future Work**
- Linear Kernel SVM - less prone to overfitting
- Doc2Vec - maps similar sentiment sentences to similar feature vectors
- TF-idf - finds frequent terms across documents to help determine which words are relevant