# Fake News Identification

Username: gcdk35

February 13, 2018

## 1 Shallow Learning Approach

Given the entire article, we want to parse the text and determine whether the text itself is fake news or not.

### 1.1 Cleaning Data

To help us achieve success through a shallow learning approach, removing any un-necessary data is necessary. The following components of the data have been considered:

**URLs** In the dataset, URLs themselves don't really add anything directly. One could potentially use the number of URLs pointing to a source to determine the validity (in a similar way to Pagerank does with search), but within our fairly small dataset, that won't necessarily work. Filtering out URLs is probably the best option in this case.

**Stop Words** Some words are known to be neutral, and don't add massive amounts to the data, only inflate the content. Therefore, by applying stop words we can filter down the text we want to parse, leaving only the important text that can be put through some classifier.

**Punctuation** Some punctuation could potentially be fairly useful at declaring how reliable a source is, for example, a text with several exclamation marks in a row could potentially indicate a source that is not quite as good as one that uses no exclamation marks at all.

In this instance, I shall leave in punctuation, although when prefiltering the text, it's important that words themselves are not considered to contain punctuation (as this will lead to several invalid matches).