GTBT Report 03/12/2014

**Previous…**

In the previous week an issue with the classification process was noted. This issue meant almost every tweet was being classified with the label ‘Governance’. Fortunately – like suspected – this was not an issue with the classifier itself but rather a bug in the software.

The bug was the mistake of attempting to classify a tweet before tokenisation and stemming had taken place.

After this was fixed by calling the *cleanTweet(tweet)* method, the results of the live classifier were improved, and a log of the different labels produced by the classifier recorded for easier analysis.

Also on the classifier, the live classifier does not seem to be able to classify tweets into the dimensions “Innovation, “Leadership”, “Workplace”, or “Undefined”.

I have two ideas that may be causing this:

* These are the smallest classes in the training set and therefore the results of the classifier are skewed because of this uneven distribution.
* The Classifier just hasn’t met any tweets that would be a likely fit in any one of these classes*.*

The second point I believe is an unlikely situation given the sheer number of tweets coming through, however, with a class like Innovation this may be true.

Again with the second point, the classifier only sees tweets coming from the search endpoint of the Twitter API and this may only be serving what it determines as “relevant” tweets and removing what it determines as background noise. Therefore, the twitter API may actually be the point at fault and not the classifier.

It was also mentioned that a small feature of the application should be able to view not only the tweet text and its class when viewing the twitter data, but to also show all the tweet attributes that could be relevant such as author, location, time, etc.

This small feature has now been implemented.

**User Stories (functional requirements)**

As agreed in the last meeting, a set of user stories should be produced with these being the goals of the project by the end of the year.

Below are a series of ten use cases that are related to a logged in actor:

--------------------------------------------------------------------------------------[0]----

As a user,

I must be notified via email when the reputation of an entity I track is in above a critical threshold

So that,

I can keep aware of critical situations that affect an entity’s reputation dimension.

--------------------------------------------------------------------------------------[1]----

As a user,

I must be able to view the locations from tweets about an entity

So that,

I can understand the influence of a tweet in different geographical markets.

--------------------------------------------------------------------------------------[2]----

As a user,

I must be able to view the changes in an entity’s reputation dimensions over time

So that,

I can understand the growth and decline of an entities reputation over time.

--------------------------------------------------------------------------------------[3]----

As a user,

I must be able to filter and view all the tweets associated with an entity

So that,

I can understand how each tweet is classified.

--------------------------------------------------------------------------------------[4]----

As a user,

I must be able to search for entities that are currently in the system

So that,

I can add existing entities to my Personalised Track List.

--------------------------------------------------------------------------------------[5]----

As a user,

I must be able to add entities to my Personalised Track List

So that,

I only need to view entities that I have an interest in.

--------------------------------------------------------------------------------------[6]----

As a user,

I must be able to add entities that are not currently in the system

So that,

I can continue to track new entities that may appear over time.

---------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------[7]----

As a user,

I must be able to remove entities from my Personalised Track List

So that,

I do not need to see entities that I have lost interest in.

--------------------------------------------------------------------------------------[8]----

As a user,

I must be able to remove entities from my Personalised Track List

So that,

I do not need to see entities that I have lost interest in.

--------------------------------------------------------------------------------------[9]----

As a user,

I want to be able to view only unique tweets, i.e. no retweets

So that,

So that I do not need to read the same thing twice.

---------------------------------------------------------------------------------------------

As a user,

I want to be able to view the distribution of tweets amongst the different dimensions

So that,

So that I can understand the reputation distributions and possible effect of a tweet within those reputation distributions.

---------------------------------------------------------------------------------------------

**Non-Functional Requirements**

1. All operations on the collection are done through the REST service.
2. REST service to record logs of all GET and POST requests.
3. REST to give endpoint for downloading JSON of n amount of tweets with class attribute removed.