**High Level Project Goals**

Part 1) Identify a Tweet as Medical or Non-Medical

To be continued…

**Current set of Ideas**

1. **Medical API’s** - Look at some of the available Medical related web APIs that are out there. These would provide a wealth or terms, conditions, medications, etc that could be used to classify data as Medical vs Non-Medical.  
     
   <http://www.nlm.nih.gov/api/>  
   [http://blog.programmableweb.com/2012/05/16/72-medical-apis-avvo-national-library-of-medicine-and-nhs...](http://blog.programmableweb.com/2012/05/16/72-medical-apis-avvo-national-library-of-medicine-and-nhs/)

2) **Some Naïve concepts -** At the simplest level we can have some key word search (Example cancer, medicine, hospital etc). Probably use a list of all medicine names, disease names etc.

3) **Medical Hashtags** - Medical hashtags' dictionary could potentially be used to ease the problem of discovering medical tweets. @ <http://www.symplur.com/healthcare-hashtags/>

4) **The Data Science Approach**

An approach to create a "dictionary of medical terms in tweets" would be to pick some obvious words ("doctor", "hospital", "nurse", "wound", "sick"...) as start, then check what are the most popular nontrivial words that co-occur in tweets with those, and then add them to the library and iterate until the tweets start being "clearly" nonmedical.

**Essentially :-** Do some statistics to find out how much each of these terms predicts whether a tweet is medical or not in a training set, and then in a test set. Co-occurrence algorithm similar to usage in Bio-informatics )

**Additional Challenge** - Co-occurrence of different terms and correlation in their relative positions to isolate medical-related tweets.

Some Other resources which may be handy :

TAGS @ <http://mashe.hawksey.info/2013/02/twitter-archive-tagsv5/> - A mash to directly download tweets related to specific search criteria to a google spread sheet.