Geo-location and text wrangling with Twitter data

Pretend I have just given you access to an enormous amount of written text – in digital form - from people all over the world where they discuss their everyday comings-and-goings, desires and aspirations, medical complaints, and reactions and opinions about public affairs and events. Let's also stipulate that this is public data, volunteered by people for the whole world to see (though, truth-be-told, many of them may have no idea just how visible their content is). Along with this data is additional information - let's call it *metadata* - that tells you when this information was shared and, maybe, the person's interactions with other people in the data set, their location, and various other items of information including their gender, age, marital status, and religion. I can make this data sound even more interesting by telling you that the text can be in any language and, to a large extent, informal, full of slang, misspellings, alternate spellings, idiomatic expressions, and code switching and mixing between languages. There's also another, unfortunate wrinkle to this data in that people are not always who or where they say they are, and may not actually be people at all; just the traces of automated processes pretending to be people. There is also a lot of this data, coming at you at the rate of hundreds of terabytes per day.

I've given you this data. Now what? How would you even begin to work with it? More importantly, how would you put it to use? Investigate how ideas spread geographically? Look for disease outbreaks? Gauge the effects of - or even detect - a natural or man-made disaster? Look at the differences in how men and women express themselves? Determine who is the most influential and trusted person? Identify people who aren't really who they say they are, or are just robots? Maybe even try to assess public opinion? The list of things you want to do with this data is, hopefully, a long one. And, as if you didn't already know it, this data exists and you don't need me to give it to you or show you where it is.

User generated content on the World Wide Web - not only limited to text, but also including photos and videos - presents great opportunities and challenges to social scientists, computer scientists, software developers, and engineers. Not surprisingly, there has been a flood of published work in the past decade addressing all of the questions thrown around in the previous paragraph.

User generated content is not a new phenomenon. USENET thread discussions go back to 1980 and on line forums were a staple of the Web from its beginning. Blogs have been around since the late 1990s and continue to be popular. The last ten years, however, has seen the mass adoption by people of social networking sites such as Twitter and Facebook, content sharing sites such as YouTube, Flickr, and Pinterest, and new-generation threaded discussion platforms such as reddit. Hundreds of millions of people across the world are contributing to these sites every day, many of them via mobile devices. It is safe to say that researchers interested in the different facets of human behavior have never seen anything like this before: massive amounts of publicly available "folk data" from all around the world available in near real time.

Working with this data is a challenge for many of the reasons I mentioned above: the informality of the language used, multiple languages, not to mention – or understate - its volume and velocity. An understanding of the particular platform the data is coming from is helpful: how the provided API works (if one is provided), the social norms of the site (e.g, whether people are expected to use their real names), and any site specific conventions adopted by users (e.g., hashtags on Twitter). Twitter has been one of the most popular sites for researchers working in this field. This may be due to the ease with which large collections of *tweets* can be collected, the brevity of the content, and the well-defined user relationships on the site: friend/follower connections and links derived from retweets and mentions. Twitter also provides a good, well documented API. For those with deep pockets, there are also commercial sources for tweets that will provide access to historic data, a consistent sample of the stream, or the entire stream of ≈400 million tweets per-day.

My talk will focus on the work I have done with Twitter data, focusing on two tasks that are required for downstream analysis: geolocating users, and normalizing tweets into a canonical form that can be used for natural language processing, social network analysis, and other tasks. The Twitter API allows for geographic queries that promise to provide tweets from a specific location: the results, in addition to the tweets, also provide various forms of geographic metadata for each user. I will discuss the common errors found in these queries results and how to address them, and will then discuss how to match the metadata against a gazetteer to give more detail about where a user may be tweeting from. Next I will address some of the noisiness of Twitter text. In particular, I will discuss the normalization of alternate spellings found in (English) tweets (yes, *aii* is short for *alright*), miscellaneous Unicode symbols such as ♫, emoticons, and hashtags. Finally, examples of the work that motivated this preprocessing will be presented, focusing on the reactions of Twitter user in Nigeria to major public issues. I hope this talk will give the listener appreciation of the issues involved with working with these data and how one might begin to tap their potential in helping us understand what’s going on with people, especially those whose voices, till now, were mostly out-of-reach.