**One pager**

**Goals**

1. Provide **an initial exposure** and **whet appetites** for the area of analysing text data
   1. *Desired audience*: early undergrads in humanities (linguistics, English), social sciences and marketing, and also in the early stages of data-science education
   2. Facilitate easy, fun experiences that lead to Aha!-movements and a broadening of horizons
2. Provide **genuinely useful analytic tools** that require minimal learning, are very quick and easy to use, and hide all the “ugliness”

Consequence: **Initially focus on areas where** …

* **the types of problem** you want to solve can be **easily understood by a layperson**, of interest to a large number of people and
* **resultant displays** are reasonably **easy to read** and **interpret**

*Later, to also have the system writes and stores the R code that does what it just did*

**Address**

* What are these people mainly talking about and how?
* How are they feeling about the issue/experience they are talking about? (“sentiment”)
* How do documents/blocks within a document/contributions by different people differ in terms of
  + what they are talking about and how
  + their sentiments?

**Don’t address:**

* Automatically extracting features from text fields to be fed in as predictor variables for predictive/classification models – *too technical*
* corpus linguistics (discovering language structure from text data) –*too technical and too small a community that this is of interest to*
* Things that are insufficiently generic, or too difficult, to be reasonably catered for by the software
  + Types of questions that the learner/user needs to write own regular expressions to answer – *too technical*
  + Things that the learner/user can only do by writing non-standard wrangling code – *too technical*

**“TOOLS CAN HELP”**

* Can often get useful insights into what subjects the documents/people are talking about and how by looking at the words, or combinations of words, they are using a lot
  + and in making comparisons by looking at words that are often used in one document/ by one person but infrequently in another
* We often get some useful insights into how people are feeling (sentiments) by looking at the frequencies of words in the document that match pre-compiled lists of words that are deemed to convey positive or negative emotions

**BIGGEST LESSONS**

* Lots of words and symbols people often use when they write/talk are not useful for understanding what the subject matter is
  + We try to strip these out before analysing the text
    - Danger from stripping out too much: We strip out all the words & symbols that belong to some precompiled list of “useless words”. This can inadvertently lead us astray by stripping out some words that would actually be important, in the current context, for understanding what the subject matter is
* Sentiment analyses look at frequencies of the use of words deemed to convey positive or negative emotions
  + Danger: The precompiled lists of “positive” or “negative”-emotion words can lead us astray in 2 ways …
    - (Words in the lists that shouldn’t be) some words that usually convey positive/negative emotions may not in the particular situation under investigation
    - (Words not on the lists that should be) some words that do not usually convey positive/negative emotions might do so in the situation under investigation