Assignment &
The Way Forward

Assignment 1 - Twitter Tokenisation (A[1])

- What should you consider addressing ?
  - All the tokenisation challenges and the design considerations in them.
  - And ... More
- The text in twitter is far from the perfect world of News Corpora.
- Here are some of the concerns:
  - Emoticons : Standardisation [:) or :-)]
  - Mentions (@Somebody)
  - HashTags (#Something)
  - Ellipsis and excessive punctuation (!!!) (Lots of it)
  - URLS/Emails
  - Unicode Glyphs

What do you need to do ?

- Look at the data !
  - Find out where the problem is harder than the vanilla tokenisation and figure out what you need to do.
- Develop a tokeniser
- Prepare a report on the challenges of this task and how your algorithm attempts to face them.
- In many instances, it might come down to design choices. It is okay. Just defend your choices.

The next P&P : Part 1

- You have two choices : Choice 1
  - Take the text from P&P-1. Ideally it should be a continuous text if not, add a few more tokenised sentences to it. Now, for each word in that text, count the number of times it occurs.
    - PS : You can keep a threshold count say (count = 3 or 5 depending upon your corpus)
  - Now plot the words vs frequency in an ascending order (sort them before and then plot)
- Choice 2: Use your tokeniser.
  - Use the twitter data or some big enough corpus (see gutenberg.org) and tokenise
  - Just as above count how many times a word occurs
    - You can use this linux command sequence : cat <filename> | tr ‘ ‘ ‘\n’ | sort | uniq -c | sort -k1nr
  - Same as above plot the words-vs-freq in ascending order (above command gives you that)
- How does the plot look like ? and Why ?

The next P&P : Part 2

- Look at this corpus :
  - Once upon a time, there were a little old woman and a little old man who lived in a little cottage near the river . The little old woman and the little old man were hungry, so the little old woman decided to bake a gingerbread man .
  - The little old woman made a big batch of gingerbread dough, then rolled it flat and cut it in the shape of a gingerbread man . She gave him raisins for eyes , a cinnamon drop for a mouth , and chocolate chips for buttons . Then she put the gingerbread man in the oven to bake.
- If I say :
  - The gingerbread <X> what should be the X ?
  - The little old <Y> what should be the Y ?
- How are we able to do this ?. Can you formalise it ?
- In general, given a word w_i can we predict w_{i+1}