TABA Batch 8 - Individual Homework 1 (Regex and basic Text-An)

On LMS there are two csv files containing 3000+ tweets each from @IBMResearch and @IBMWatson.

Your task is to:

1. Clean up the @IBMResearch tweets corpus as far as possible of nonsensical patterns using any tools at your disposal - including Regex.

2. Index and text-analyze the cleaned corpus for some hint of content and meaning. Use standard display aids for help and interpretation.

3. In particular, find the top 5 most used [unigram] words which aren't stopwords, the top 5 bigrams, the top 5 hastags (i.e., '#htags') and the top 5 twitter handles (e.g., '@RaviKumar').

4. Build a dataframe with columns 'tweet_num', 'top_words', 'top_bigrams', 'top_hashtags' and 'top_handles' and populate it with the above data.

5. *Functionize* your work in steps 1-4. That is, write a [set of] directly callable function[s] that will do steps 1-4 for any other input dataset-as-argument.

6. Now invoke your function and apply it to the corpus of tweets from @IBMWatson.

7. Compare the results from the two corpora. Speculate and write a few lines on the major similarities and differences between the two based on your analysis.

Deliverable: A Markdown document as an HTML file.

Deadline: Midnight of 1-May-2017.

Any Qs etc, ask Preethi or Aashish. If unresolved, write to me.

Sudhir