Eval IL Pack Components and Structure

Directory Structure

Notes:

1. "LANG" is a variable to be replaced by the 3-letter iso code for the incident language.
2. Each “set” directory will be encrypted, with decryption keys provided as specified in the evaluation plan.
3. Only monolingual_text/ directory will be under data/ in setE delivered to performers. NIST will also receive annotation/ and translation/ subdirectories

```
README.txt
set0/
   README.txt
   data/
      monolingual_text/
      translation/
         found/
            eng/
               ltf/
               psm/
            sentence_alignment/
            LANG/
               ltf/
               psm/
   docs/
      categoryII/
      categoryI_dictionary/
   dtds/
   tools/
      encoding/
      ltf2txt/
      twitter-processing/
set1/
   README.txt
   data/
      monolingual_text/
   docs/
set2/
   README.txt
   data/
      monolingual_text/
   docs/
setS/
   README.txt
   data/
      monolingual_text/
   docs/
```

setE/
  README.txt
  data/
    monolingual_text/
    annotation/
      entity/
      situation_frame/
    translation/
      eng/
        ltf/
        psm/
      LANG/
        ltf/
        psm/
  docs/

Items in red will be sent only to NIST for evaluation languages.

Data Volumes

Set0 (all monolingual and parallel text from pre-incident epoch):
Monolingual text – 225Kw (45% NW, 33% DF/WL, 22% SN)
Parallel text – 300Kw (33% each NW, DF/WL, SN; can substitute 300Kw comparable for 100Kw parallel)
Parallel dictionary – 10K lemmas
Other resources ("Category II") – 5 of 8 resource types: parallel IL - > non-English dictionary, monolingual IL dictionary, monolingual IL grammar book, parallel IL - > English grammar book, monolingual IL primer, monolingual IL gazetteer, parallel IL - > English gazetteer, English gazetteer for incident region

Note: Minimum target to be exceeded by 500% target for **one** of monolingual text, parallel text, or parallel dictionary

Set1 (incident date and later):
Monolingual text – 1/3 of leftover after set E is met

Set2 (incident date and later):
Monolingual text – 2/3 of leftover after set E is met

SetS (incident date and later):
English text, some incident-relevant – approximately 50Kw, genre balance will vary based on availability

SetE (incident date and later):
Monolingual text – 200Kw (approximately 50% NW, at least 25% DF/WL, up to 25% SN)
Reference translations and annotations for Simple Named Entity and Situation Frames will be provided to NIST for (subsets of) this data.

Documentation

The following items should be included in all eval IL docs/ directories:

- categoryI_dictionary/ – directory containing (pointer to) dictionary SET 0 ONLY
- categoryII/ – directory containing (pointers to) all category II resources SET 0 ONLY

- source_codes.txt – 4 columns: genre, source code, source name, (base) url ALL SETS
- twitter_info.tab – 4 columns: file name, "network_id" (Tweet ID), md5sum, author_id ALL SETS
- urls.tab – list of all source docs with document urls ALL SETS