Ex: More Tips

Wednesday, 14 November 2018 08:47

What follows is a set of tips for a minimal attempt at the assignment. There are far more ambitious things you can try using the various tools and techniques we've covered. However the following is a basic approach.

Tagging the seminar information

1. Your first steps are in learning to:

    a) read in a document (or an email),

    b) do some processing,

    c) write out a new version of the email with tags in a different directory.

    See https://birmingham.instructure.com/courses/31164/pages/corpora-and-regex

2. Use the tokeniser & a POS tagger to tag the text.

    We've done both at the start of term. Do not train your POS tagger on the emails. Train it on regular text (as we did in the lab class).

    See https://birmingham.instructure.com/courses/27273/pages/pos-tagging

3. Use Regular Expressions

    a) use regex to grab information from email headers and then match this information in the text

    b) use regex to tag times, dates etc. These are very regular. You can learn how to do this by studying the training data. You'll need to have several regex for different formatting styles for time and date.

4. Person names are harder. However you can try the following:

a) We've given you three long lists of people names as gazetteers. Use them!

https://birmingham.instructure.com/courses/27273/files/folder/Data

b) Regex for job titles Dr.|Mr.|Ms|Mrs|Prof. (NN1 NN2) probably indicates that NN1 and NN2 are a person name. Similarly the POS tag pattern NN1 NN1 Verb probably indicates a name.

c) Distinguishing who is actually the seminar speaker might be hard. Look at patterns such as

Dr Smith will present etc.

There's a small list of verbs which seem relevant here - and you can quickly get this list from the training data. In addition you can eliminate who isn't the speaker by looking at email header information.

5. This part of the exercise doesn't ask for topic identification. However it makes sense to think about this as part of the annotation. Then building an ontology of topics becomes much easier.

6. The object of the assignment is evaluation. We don't care if you don't get a high accuracy score. We'll be very impressed if you do but to get a good mark for this assignment you just need to implement something sensible and then evaluation it in terms of precision, recall, and f measure. You'll need to write some code to calculate this. This involves comparing the original data with annotations with the corpus your system produces.

Ontology construction

1. look at the topic header in most of the emails. There is often a subject term e.g. "Computing". There might be other words. Consider what resources could be used. At a minimum - collect words in all topics in the email corpus and then order them by frequency. The highest frequency terms are probably a good place to start as terms for the ontology.

2. Once you have a list of terms then you can manually construct the ontology tree. For instance:

science -----> physics -----> nuclear
-----> chemistry -----> organic

-----> inorganic

etc. etc.

3. Then your task is simply to tag each email with a specific ontology tag from the above tree.

(A more advance solution might involve wordnet, word2vec, wikification, relation extraction.)