Ex

Monday, 12 November 2018 09:42

- In the Data folder you'll find two files:
    - seminars_training.zip contains 300 hand tagged seminar announcements downloaded from mailing lists.
    - seminars_untagged.zip contains the same 300 emails without tags.

The assignment is in two parts.

1. **Entity Tagging (worth 10% of module mark)**

- First write code which takes untagged seminar announcements and automatically labels as follows:

    <0.14.7.88.10.56.03.Kai-Fu.Lee@SPEECH2.CS.CMU.EDU.0>

    Type: cmu.cs.proj.speech

    Topic: Talk

    Dates: 15-Jul-88

    Time: <stime>11:00 AM</stime>

    PostedBy: Kai-Fu.Lee on 14-Jul-88 at 10:56 from SPEECH2.CS.CMU.EDU

    Abstract:

    <paragraph><sentence><speaker>Paul Chou</speaker> will be given a vision talk Friday at <stime>11:00AM</stime> in <location>4605 Wean</sentence></location>.

    <sentence>The title of his talk is</sentence>:

    Probabilistic Information Fusion for Multi-Modal Image Segmentation

    <sentence>This talk may be of interest to speech researchers in two ways: (1) Paul

is interested in joining our speech group, (2) his work on combining

knowledge sources using Markov Random Fields may be applicable to speech

recognition</sentence>.</paragraph>

- New data will be released at the start of Week 6. You must submit 2 pieces of work - the actual code of your system and an evaluation of your code on the new data (as a written report). You will be marked on the intelligent use of NLP techniques and not on the actual success of your system.

2. Ontology Construction (worth 10% of module mark)

- The seminar announcements are from various academic departments. Using the automatically annotated data you've created in part 1, you need to automatically create an classification of the announcements. For example, you need to produce a list of all announcements (or their document ID) related to chemistry.

- This classification needs to be structured. For example, both artificial intelligence and HCI are subjects in Computer Science. Your code should be tested on the 2nd set of released data (which is released in week 8). You must submit your code plus an an evaluation of your code on the new data. You will be marked on the intelligent use of NLP techniques and not on the actual success of your system.