|  |  |  |
| --- | --- | --- |
| **Task** | **Person** | **Date and Time** |
| Define project idea and scope | Dirk, Huda | 11/08 11:00 – 14:00 |
| Create training data (emails) | Dirk, Huda | 11/08 14:00 – 14:30 |
| Load training data in the code | Dirk | 11/08 14:30 – 15:00 |
| Define features | Huda | 11/08 14:30 – 15:00 |
| Create chunker | Dirk, Huda | 11/08 15:00 – 18:00 |
| Create features | Dirk, Huda | 11/10 14:00 – 16:00 |
| Write and comment code | Dirk, Huda | 11/10 16:00 – 18:00 |
| Write and comment code | Dirk | 11/12 13:00 – 14:30 |
| Improve code | Dirk | 11/12 22:30 – 00:00 |
| Write up results and write report | Huda | 11/13 00:00 – 01:00 |

We choose suggestion 1 of the Information Extraction assignment. Based on emails with invitations of events and conferences or talks we want to extract information about the location where the event will take place.

We trained our classifier on a set of hand-annotated emails, which were also tagged for part of speech.

For example: There is an event on Friday in South Hall 202. The NE classifier should find ‘South Hall 202’.

Features that we think are important:

* Words have capitals except for stopwords
* Words are in all caps
* There are numbers included
* Most words are nouns
* There are brackets ()
* Most words won’t be in the dictionary
* Most words are names NP
* The locations often include hall, school or house

- Prefaced by the word, “Location”

We train the classifier by saving the email invitations in a training file and other similar messages in a test file. All files are automatically tagged by the NLTK POS-tagger and are all assigned the IOB tag O by default. We then manually assigned the correct IOB tag to the locations mentioned in the emails. The tags used are B-Loc and I-Loc. Based on earlier projects on Named Entity recognition we choose for this structure. We then move the files to the nltk\_data folder into the conll2000 corpus to read the files in the script through the conll2000 object.

Source: <https://groups.google.com/forum/#!topic/nltk-users/jg33BbMW4mQ>

The classifier is trained by the features mentioned above. We originally planned on using a Maxent classifier, since that does not assume independence of the features. We wanted to consider this classifier because we believed our features were not independent. However, we decided to switch to Naïve Bayes because we had difficulty using the MaxEnt classifier with the combination of versions of Scipy and NLTK.

Evaluation:

#Number of test emails:

Precision: (true positive)/(true positive + false positive) =

Recall: (true positive)/(true positive + false negative) =