Assignment: Extracting-Structure

June 4, 2019

1 This tutorial is Largely based on the paper Context-aware Argumentative Relation Mining

Huy V. Nguyen Diane J. Litman

However the implementation is not the exact procedure stated in the paper, rather covers the overall intention. The main intention being somehow give the contextual information to the classifier model by extracting topics using LDA.

The implementation is divided into 6 parts :

```
Part 1 : Importing and structuring the Dataset
Part 2 : Window Context Extraction ( TODO )
Part 3 : LDA topic Extraction ( TODO )
Part 4 : Creating and Adding the features ( TODO )
Part 5 : Applying Classification Models ( TODO )
Part 6 : Hyperparameter tuning ( additional )
```

2 Part 1.

2.1 Importing and structuring the Dataset of 90 Persuasive ESSAYs

Importing the actual essays: The code for importing the dataset is already implemented. Once it is executed, you should see the following output.

```
In [9]: essay_dict['essay34'] # displaying sentences of essay number 34

Out[9]: ['Study at school or get a job?',
'Many people believe that children should study at school to have
more knowledge that prepare better for their future.',
'Others, however, think that these children may disrupt their
school work and should be allowed to leave school early to find a
job.',
'Personally, I tend to agree with the point of view that student
have to be forced to study at school.',
'First of all, schools offer to students a good environment with
experienced professors and high quality programs for studying.',
'It creates the best conditions for students education and can
force them to focus on their school work instead of wasting their
```

```
time to do useless things.',
'Second of all, schools provide lots of academic knowledge to
students.',
'Students may learn professional skills, expand their
understandings and gain experiences.',
'Therefore, they have more opprotunities to find a job and to be
successful in the future.',
'For example, as we know, employer always prefer to hire an
employee of high degree who have professional skills.',
'Nevertheless, it is not unreasonable that some people think that
children should interrupt their school work and get a job.',
'Whether children can learn a lot at school, there are many
subjects that will be of little value to them in the future.',
'Furthermore, children can learn social skills when they have a
job.',
'They can get more experiences that can not be obtained at
school.',
"Working helps children be more independent and teach them to
esteem and manage the money that they've earned.",
'Overall, I believe that students should study at school.',
"Even though there are some advantages of leaving school to find a
job, studying at school is always the best choice for children's
future.",
'There are many ways that can train children to learn independent
and social skills instead of getting a job.']
```

Importing the Annotations: The code for importing the annotations is also implemented. After execution, the annotations of each ADU(source and target) should look like below. This dataset contains all permutated combinations of the relation between each source and target ADUs, however we are only interested in 'attacks/supports' relation. Thus we remove intentionally the 'no relation' rows.

```
In [12]: dataset.head() # first 5 rows of the dataset
```

```
Out[12]:   src_id                                                src src_type src_strt  \
         0     T1  competition can effectively promote the develo...    Claim       78
         1     T1  competition can effectively promote the develo...    Claim       78
         2     T1  competition can effectively promote the develo...    Claim       78
         3     T1  competition can effectively promote the develo...    Claim       78
         4     T1  competition can effectively promote the develo...    Claim       78

           src_end tgt_id                                              tgt  \
         0     140     T2    we should attach more importance to cooperation
         1     140     T3  In order to survive in the competition, compan...
         2     140     T4  through cooperation, children can learn about ...
         3     140     T5  What we acquired from team work is not only ho...
         4     140     T6  During the process of cooperation, children ca...
```

```
     tgt_type tgt_strt tgt_end     relation    essay
0  MajorClaim      503     550      attacks  essay01
1     Premise      142     283  no relation  essay01
2       Claim      591     714  no relation  essay01
3     Premise      716     851  no relation  essay01
4     Premise      853    1086  no relation  essay01
```

```
In [13]: dataset[dataset['relation'] != 'no relation']['relation'].value_counts()
```

```
Out[13]: supports    1312
         attacks      161
         Name: relation, dtype: int64
```

3 Part 2.

3.1 Context Window Extraction

Getting then neighbouring sentences for each source and target pairs: In this section we try to find n window sentences of the source and target ADUs from the essay corpus. These $2n$ (n previous and n next) sentences are then added to the dataset created previously. Once the n window sentences corresponding to each source and target ADU are extracted successfully and added to the dataset, it should look something like below.

Here only first 5 rows and 2-window sentences are shown for source and target sentences

```
In [150]: neighbours[['src_next_sent1','src_next_sent2','src_prev_sent1','src_prev_sent2',
                      'tgt_next_sent1','tgt_next_sent2','tgt_prev_sent1','tgt_prev_sent2']].head(
```

```
Out[150]:                                          src_next_sent1  \
          0  In order to survive in the competition, compan...
          1  However, when we discuss the issue of competit...
          2  What we acquired from team work is not only ho...
          3  During the process of cooperation, children ca...
          4  All of these skills help them to get on well w...

                                             src_next_sent2  \
          0  However, when we discuss the issue of competit...
          1  From this point of view, I firmly believe that...
          2  During the process of cooperation, children ca...
          3  All of these skills help them to get on well w...
          4  On the other hand, the significance of competi...

                                             src_prev_sent1  \
          0  Should students be taught to compete or to coo...
          1  It is always said that competition can effecti...
          2  From this point of view, I firmly believe that...
          3  First of all, through cooperation, children ca...
          4  What we acquired from team work is not only ho...
```

```
                                        src_prev_sent2  \
0  Consequently, no matter from the view of indiv...
1  Should students be taught to compete or to coo...
2  However, when we discuss the issue of competit...
3  From this point of view, I firmly believe that...
4  First of all, through cooperation, children ca...

                                        tgt_next_sent1  \
0  First of all, through cooperation, children ca...
1  In order to survive in the competition, compan...
2  First of all, through cooperation, children ca...
3  What we acquired from team work is not only ho...
4  What we acquired from team work is not only ho...

                                        tgt_next_sent2  \
0  What we acquired from team work is not only ho...
1  However, when we discuss the issue of competit...
2  What we acquired from team work is not only ho...
3  During the process of cooperation, children ca...
4  During the process of cooperation, children ca...

                                        tgt_prev_sent1  \
0  However, when we discuss the issue of competit...
1  Should students be taught to compete or to coo...
2  However, when we discuss the issue of competit...
3  From this point of view, I firmly believe that...
4  From this point of view, I firmly believe that...

                                        tgt_prev_sent2
0  In order to survive in the competition, compan...
1  Consequently, no matter from the view of indiv...
2  In order to survive in the competition, compan...
3  However, when we discuss the issue of competit...
4  However, when we discuss the issue of competit...
```

4 Part 3.

4.1 LDA Topic Extraction

Loading the extra Essay corpus: The code for loading the other essay corpus is already implemented. It contains extra essays including the previously loaded essays. The intention here is to extract only the heading/first sentence of each of these essays. Once the headings of each essays are extracted, the data would look like below.

```
In [81]: data[:5]
```

```
Out[81]: ['Should students be taught to compete or to cooperate?\n',
          'More people are migrating to other countries than ever before\n',
```

```
 'International tourism is now more common than ever before\n',
 'International tourism is now more common than ever before\n',
 'Living and studying overseas\n']
```

Now a set of text preprocessing needs to be implemented. Some of which are already complete, while some others are left for implementation. Mainly the below tasks need to performed on the extracted headings of the essays:

Tasks:

- remove newline characters and single quotes.
- remove stop words.
- lemmatize each word/token as noun, verb and adjective.

Once the preprocess steps are completed, the headings should look like this.

In [151]:
```
print(preprocess_pipeline(data)[60:90])
```

```
[['distance_learn', 'cannot', 'bring', 'benefit', 'traditional', 'college', 'offer'],
['protect', 'animal'],
['lesson', 'teacher', 'versus', 'others', 'source'],
['every', 'student', 'either', 'male_female', 'give', 'equal', 'opportunity'],
['convincing'],
['improve', 'medical', 'care'],
['successful', 'people', 'something', 'new', 'take_risk'],
['university_education', 'available', 'good', 'student'],
['improve', 'facility', 'best_way'],
['high_school', 'must', 'responsible', 'future', 'well', 'plan', 'curriculum'],
['importance', 'game', 'adults', 'compare', 'child'],
['opinion', 'regard', 'world', 'culture'],
['opportunity', 'receive', 'education', 'university'],
['email', 'text', 'message', 'threat', 'write', 'language'],
['read', 'fiction', 'pleasant'],
['various', 'local', 'language', 'become', 'extinct'],
['believe', 'elderly', 'people', 'use', 'live', 'good', 'world'],
['high_school_student', 'taught', 'manage', 'money'],
['live', 'roommate', 'better', 'living', 'alone'],
['smoking', 'permit', 'restaurant'],
['grow', 'violence', 'film', 'affect', 'youngster', 'negative', 'way'],
['gossip', 'base', 'information'],
['old', 'building', 'preserve'],
['popularity', 'mobile_phone', 'young_people'],
['use', 'animal', 'benefit', 'human', 'being'],
['people', 'believe', 'fix', 'punishment', 'crime', 'type'],
['leisure', 'activity', 'spend', 'free', 'time', 'outdoors', 'indoors'],
['cheap', 'air', 'travel', 'encourage'],
['convenient', 'life', 'city'],
```

```
['salary', 'increase', 'promotion', 'new', 'position', 'way']]
```

After preprocessing, a Document Term Matrix is created. The code for this already implemented. Also a mapping of each word to ID is created. Both of these are then fed to a LDA model to assign topics to the words. Lastly, the number of topics is varied as a parameter to check the Coherence Score of the LDA model.

```
In [90]: # View
         print(corpus[9])
```

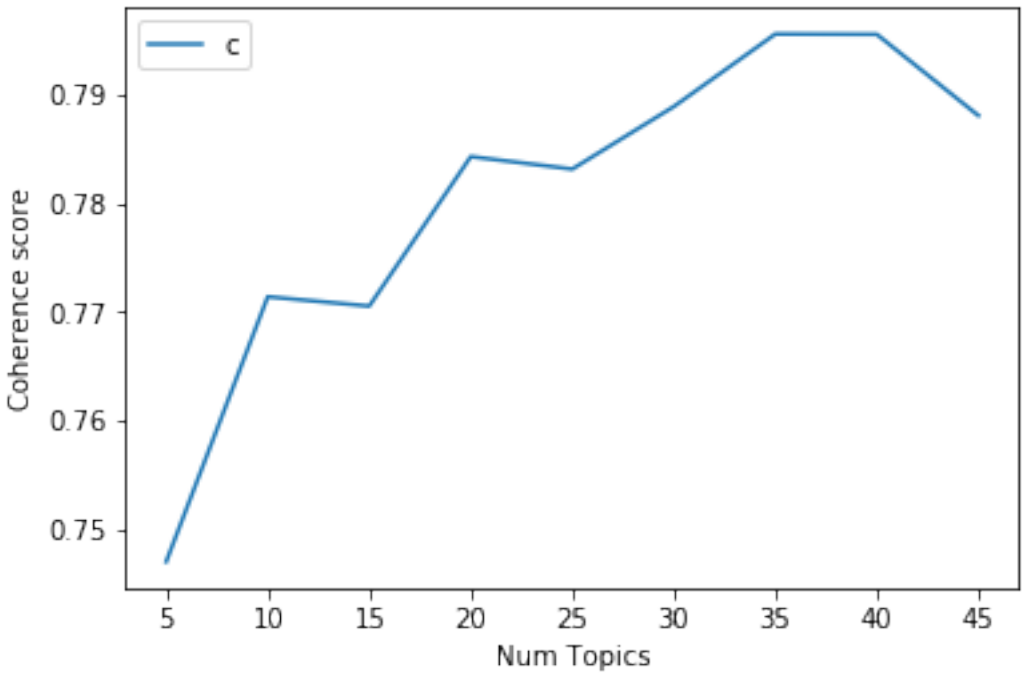
```
[(25, 1), (26, 1), (27, 1)]
```

The below graph shows that for the available corpus, 35-40 topics yields the best coherence score. Students can try out different values of topics and check the score.

```
In [97]: # Compute Coherence Score

         print('\nCoherence Score: ', coherence_lda)
```

```
Coherence Score:  0.7924420473678659
```



5 Part 4.

5.1 Creating and adding Features:

In this section we mainly focus on transforming and creating the actual features from the dataset that we have built so far.

The topics which were extracted in previous step contains words which are closely related/ belong to similar context. Below is an example of topic 13 and its top 5 words.

```
In [99]: topic_words[13][:5] # top 5 words of topic 13

Out[99]: ['pay', 'high', 'student', 'university_education', 'jury']
```

The code to transform the features are partly implemented and partly left for exercise. The main tasks include:

Tasks:

- create feature based on the overlap of source/target ADU tokens and topic tokens
- create feature based on word counts of source/target ADU segments.

Once all the features have been extracted and transformed, the dataset should look something like this.

```
In [154]: X[['src_type_Claim','src_type_Premise',
              'tgt_type_Claim','tgt_type_MajorClaim','tgt_type_Premise',
              'abs_diff_strt','abs_diff_end','word_count_src','word_count_tgt']].head() # first 5
```

Out[154]:

	src_type_Claim	src_type_Premise	tgt_type_Claim	tgt_type_MajorClaim	tgt_type_Premise	abs_diff_strt	abs_diff_end	word_count_src	word_count_tgt
0	1	0	0	1	0	425	410	62	47
1	0	1	1	0	0	64	143	141	62
2	1	0	0	1	0	88	164	123	47
3	0	1	1	0	0	125	137	135	123
4	0	1	1	0	0	262	372	233	123

6 Part 5.

6.1 Applying Classification models:

In this section we mainly call the classification models and train it to predict the results.

However first the dataset has to be divided into training and test dataset. This is left as a task.
Tasks:

- divide the dataset into train set and test set

Apply the best Classification model

```
In [144]: model = SVC(C=5,class_weight='balanced')

          model.fit(X_train,Y_train)
          Y_pred = model.predict(X_test)
```

Calculating the precision macro, recall macro, f1 macro and accuracy of the model

```
In [145]: p_macro, r_macro, f_macro, support_macro = precision_recall_fscore_support
                                                      (y_true=Y_test,
                                                      y_pred=Y_pred,
                                                      labels=[0,1],
                                                      average='macro')

          print('Accuracy:',round(accuracy_score(Y_test, Y_pred),2),
                '\nKappa:',round(cohen_kappa_score(Y_test,Y_pred),2),
                '\nMacro Precision:',round(p_macro,2),
                '\nMacro Recall:', round(r_macro,2),
                '\nMacro F1:',round(f_macro,2,),
                '\nF1:',round(f1_score(Y_test, Y_pred),2,)
                )
```

```
Accuracy: 0.84
Kappa: 0.11
Macro Precision: 0.57
Macro Recall: 0.55
Macro F1: 0.55
F1: 0.91
```

7 Part 6.

7.1 Hyperparameter tuning:

This part is additional. It is mainly to select the best hyperparameters for the classification model. The steps can just be executed in order to get the best classifier.

In [146]:
```
param_test1 = {
    'C': [c for c in range(1,120,2)]
}

clf = SVC(class_weight='balanced')#GradientBoostingClassifier(random_state = 42)#Rando

gsearch1 = GridSearchCV(
    n_jobs=-1,
    estimator=clf,
    param_grid=param_test1,
    scoring= scorers,
    verbose= True,
    iid=True,
    refit='f1_score',
    cv=KFold(n_splits=5))

gsearch1.fit(X_train, Y_train)

gsearch1.best_params_, gsearch1.best_score_
```

```
Fitting 5 folds for each of 60 candidates, totalling 300 fits


[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  42 tasks      | elapsed:   11.6s
[Parallel(n_jobs=-1)]: Done 192 tasks      | elapsed:   41.7s
[Parallel(n_jobs=-1)]: Done 300 out of 300 | elapsed:  1.1min finished
```

Out[146]:
```
({'C': 5}, 0.5473313503201346)
```