Instructions:

- You can use either Keras/Tensorflow to work on this case.
- No sharing of work. You can work in your teams of two only. Note: since we have an odd number, the person who won't have a team member will work on any one task
- You are expected to submit a report that summarizes the key steps in your implementation as a flow chart and submit fully functional code.
- Deadline: 11/03/2017 4.59 PM. Late submissions lose 10% points per day.
- Each team will have 15 minutes to present the 2 subtasks + 5 min Q&A on 11/03/2017

Preparation:

LSTM Review:

Review class notes and video posted on Blackboard:

Keras:

Review Keras implementation examples here:
https://github.com/fchollet/keras/tree/master/examples

Read the following articles for descriptions

- https://machinelearningmastery.com/predict-sentiment-movie-reviews-using-deep-learning/ for explanation
- https://machinelearningmastery.com/sequence-classification-lstm-recurrent-neural-networks-python-keras/

Theory:

Skim through https://www.cs.uic.edu/~liub/FBS/chapter-1-and-chapter-2.pdf for an intro on Sentiment analysis

Tensorflow:

Another perspective on Sentiment analysis with Tensorflow:

- Watch: https://www.oreilly.com/learning/perform-sentiment-analysis-with-lstms-using-tensorflow
- Code: https://github.com/adeshpande3/LSTM-Sentiment-Analysis

Midterm:

Task: Your task in teams of 2 is to build a Sentiment analysis engine for the Semeval 2017 task: http://alt.qcri.org/semeval2017/task5/. The data is available on the site.

Review this paper for details: http://nlp.arizona.edu/SemEval-2017/pdf/SemEval089.pdf . Approaches that were used in Semeval 2017 are here: http://nlp.arizona.edu/SemEval-2017/

This task has two subtasks:

1. Sentiment analysis of Twitter data


```
{
  "source": "twitter",
  "cashtag": "$HOT",
  "sentiment score": "0.405",
  "id": "719547552874512384",
  "spans": [
    "Airplane And Hospitality Industries Set Their Sights On #Cuba"
  ]
},
{
  "source": "stocktwits",
  "cashtag": "$BBRY",
  "sentiment score": "0.296",
  "id": "18346099",
  "spans": [
    "nice bounce"
  ]
},
```


2. Sentiment analysis of News headlines


```
{
  "id": 3,
  "company": "IMI",
  "title": "IMI posts drop in first-quarter organic revenue; warns on full year",
  "sentiment": -0.344
},
```


Tasks:

1. You should use Neural network algorithms for this task. (You could use any algorithm MLP, CNN, RNN).
2. You can choose to use the same or different approaches for each subtask.
3. You should document your experiments, parameters, assumptions and tests you conducted for full credit
4. You should create a function to compute the evaluation metric discussed here: http://alt.qcri.org/semeval2017/task5/index.php?id=evaluation

Deliverables:

a. Source code in Keras/Tensorflow in ipyb notebooks through github (with all results embedded in the notebooks)
b. Run the prediction for the full training data in the json format

   (See:

   https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-1/src/beadeb1fd0f9b8093e4828a198a92e651a4e10c6/Microblog_Trainingdata.json?at=master&fileviewer=file-view-default

   &

   https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-2/src/2fb645e839b7fb9923d2402d8ee817242360993f/Headline_Trainingdata.json?at=master&fileviewer=file-view-default)

c. Compute and generate outputs in csv format (Task1_train.csv, Task2_train.csv)
   SubTask 1

id	spans	Source	Cashtag	Sentiment score	Predicted Sentiment score

   SubTask 2

id	Company	Title	Sentiment score	Predicted Sentiment score

d. Compute the Cosine scores for both subtasks
e. Write a report with clear explanations and analysis of the models you build, network design, evaluation criteria, performance metrics and how your model performed.
f. At 2 pm EST on 3rd November, you will receive 2 test files in the format listed below.

   https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-1/src/beadeb1fd0f9b8093e4828a198a92e651a4e10c6/Microblogs_Testdata.json?at=master&fileviewer=file-view-default

   and

https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-2/src/2fb645e839b7fb9923d2402d8ee817242360993f/Headlines_Testdata.json?at=master&fileviewer=file-view-default

g. You should compute tables similar to step d and submit the outputs in the following format. (Task1_test.csv, Task2_test.csv)

SubTask 1

id	spans	Source	Cashtag	Predicted Sentiment score

SubTask 2

id	Company	Title	Predicted Sentiment score