Data Science
Coding Assignment #2
Headline Attractiveness Predictor

Headline Attractiveness

- In this era, how to attract people‘s attention becomes an important issue. It is an essential issue for general advertisers.

Outline

- Data Annotation
  - Description / Data Format / Submission Format
  - Grading Policy
  - Important Information
- Attractiveness Predictor
  - Description / Data Format / Submission Format
  - Grading Policy
  - Important Information

2-1 Data Annotation (10%)

Deadline: 2020/10/16 23:59 (Friday)

Attractiveness

- Is the title of the article written in a way that will raise curiosity and attract people to read the main content of the article?
- The attractiveness will be scored in 5-point Likert scale:
    1. Very Low : the headline can not provide enough information nor literally attract attention. (seldom)
    2. Low : the headline is dull and flat.
    3. Moderate : the headline is embellished with rhetoric to increase variety.
    4. High : the headline is embellished and constructed in special way to arouse the interest.
    5. Very High : the headline is well-refined to intrigue the curiosity. (seldom)

Attractiveness - Example

Headline	Score
Newcastle 3-2 Everton: Papiss Cisse, Ayoze Perez and Jack Colback goals end Magpies' run of three-straight league defeats	2
NBC Cameraman with Ebola lands in the US from Liberia to begin life-saving treatment for virus	3
'Men have gone berserk': fury in India as five tribesmen are arrested for gang rape of Swiss tourist, 39, 'attacked in her tent as her husband was tied to a tree'	4
Being hairy can be good for you, man or a woman	5

- Note:
  - Do **NOT** concentrate on the topic of headline (e.g., sports, politics).
  - When you are uncertain about an annotation, recall the scoring is based on **“How much the headline raises curiosity and attract people?”**.

Data Format

- The data to be annotated are placed in Google drive. Please download the file according to your student ID.
- The file is a **CSV file** with the following 3 columns:
  - Headline ID
  - Headline
  - Category
- There are **80 headlines** to be annotated for each of you.

	A	B	C
1	34fbd58ecdc572cf4248eae6af1f402e3ae7dba2	Conjoined twins faith and hope pass away at 19 days old: tributes pour in for brave little girls with two hea	news
2	4c7a64acde811e789490e5a17f159b1b29043a1c	Oldest man in America dies aged 110 after revealing oatmeal breakfast, daily exercise and early nights wer	news
3	129eb389d03dc9c5d152902c71a5dc541263d18b	Eight children killed in playground cluster bomb attack launched from Syrian warplanes	news
4	0a54d0e7fe6400dcf971d7ded7e8f111b67ea7a6	Sunderland make surprise loan bid for Barcelona Wonderkid Alen Halilovic	football
5	009bd9569450d2d5cc12fedc2a6e9b23fdb5f396	Wife of fraudster who stole millions from property firm says she knew nothing about her 'Cad' husband's c	news

Submission Format

- Please submit the annotated file to e3.
- The submitted file should be a **CSV file** with the following 4 columns:
  - Headline ID
  - Headline
  - Category
  - Label

	A	B	C	D
1	34fbd58ecdc572cf4248eae6af1f402e3ae7dba2	Conjoined twins faith and hope pass away at 19 days old: tributes pour in for brave little girls with two hea	news	1
2	4c7a64acde811e789490e5a17f159b1b29043a1c	Oldest man in America dies aged 110 after revealing oatmeal breakfast, daily exercise and early nights wer	news	2
3	129eb389d03dc9c5d152902c71a5dc541263d18b	Eight children killed in playground cluster bomb attack launched from Syrian warplanes	news	3
4	0a54d0e7fe6400dcf971d7ded7e8f111b67ea7a6	Sunderland make surprise loan bid for Barcelona Wonderkid Alen Halilovic	football	4
5	009bd9569450d2d5cc12fedc2a6e9b23fdb5f396	Wife of fraudster who stole millions from property firm says she knew nothing about her 'Cad' husband's c	news	5

Grading Policy

- Each headline will be annotated by three people, and an annotation is regraded as ‘disagree’ if :
  - The annotation is different with the other two annotations, and the other two annotations are the same. EX: 2/3/3
  - All annotations are different, then the ones that are different with the judgement from TAs. EX: 2/3/4 (students), 3 (TAs).
- **Disagree Ratio = # of disagree annotation / # of total annotation**
  - $[0, \frac{1}{3}]$: 10 points
  - $[\frac{1}{3}, \frac{1}{2}]$: 5 points
  - $[\frac{1}{2}, 1]$: 0 point

Important Information

- We recommend you to separately annotate the data for **3~4 days**. According to our experience, the quality can be better ensured.
- The annotation file (data.csv) is encoded using UTF-8. It can be opened with pure text editor, or using Excel for easy reading.
- Using Excel to open UTF-8 file (Window and Mac) : https://excel.officetuts.net/en/examples/how-to-import-csv-file-that-uses-utf-8-encoding (NOTE: our data don't have header, so uncheck "My data has header" in step 6)
- The above process is only required for the first time you open the file. After modifying and storing, the encoding will be changed depending on your OS.

Important Information

- Deadline: 2020/10/16 23:59 (Friday)
- Google drive link: https://drive.google.com/drive/folders/1CT4p_TO2YY5QGI_owNX-LKkcx5FicjK3?usp=sharing
- If you have any question (e.g. incomplete data, ambiguous data), please post on the e3 forum. TAs will response as soon as possible.
- TAs:
  - Yi-Syuan Chen: yschen.eed09g@nctu.edu.tw
  - Yun-Zhu Song : yunzhusong.eed07g@nctu.edu.tw

2-2 Attractiveness Predictor (90%)

Deadline: 2020/11/10 (Tuesday) 23:59

Kaggle Competition:

https://www.kaggle.com/t/de7aa21921c641cc93c1256a4bbb0e7d

Task Description

- In this Kaggle competition, you need to predict the attractiveness of a headline based on the **content** or the **category**. There are expected 2267 data, while 2040 for training and 227 for testing (50% public and 50% private).

Data Format

- Please download the training/testing dataset from Kaggle.
- The training/testing dataset are **CSV files** with the following 4 columns:
  - ID
  - Headline
  - Category
  - Label : real value in [1,5]

	A	B	C	D
1	ID	Headline	Category	Label
2	1	Conjoined twins faith and hope pass away at 19 days old: tributes pour in for brave little girls with two hea	news	2
3	2	Oldest man in America dies aged 110 after revealing oatmeal breakfast, daily exercise and early nights wer	news	4
4	3	Eight children killed in playground cluster bomb attack launched from Syrian warplanes	news	4
5	4	Sunderland make surprise loan bid for Barcelona Wonderkid Alen Halilovic	football	3
6	5	Wife of fraudster who stole millions from property firm says she knew nothing about her 'Cad' husband's c	news	5

Submission Format

- Please submit the testing results to Kaggle. You are allowed to submit the testing results **5 times per day**.
- The submitted file should be a **CSV file** with the following 2 columns:
  - ID : the index of the original order
  - Label : real value in $[1,5]$

	A	B
1	ID	Label
2	1	1.6
3	2	3.8
4	3	4.2
5	4	2.7
6	5	2.5

Grading Policy

- There are two phases for testing, i.e., public and private.
- You can only see the result of public testing data when the competition going. After the competition, the private testing result will be revealed, and the grade of this part is based on the private testing result.
- The evaluation metric is MSE (Mean Square Error)

$$L_{mse} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_i - y_i)^2$$

Grading Policy

- Top 10% : 100 points
- Top 25% : 90 points
- Top 50% : 80 points
- Top 75% : 75 points
- Others : 70 points
- Bellow baseline (**public testing loss > 0.4796**) : 0 point

Important Information

- Your team name on Kaggle should be your student id (i.e., 0750123).
- Submit your **zipped source code** {student_id}.zip to e3. After unzip the file, it should appear a folder {student_id} :
  - {student_id}
    - {student_id}.sh : run this script to regenerate your final submission result
    - requirements.txt : refer to HW1
    - Other files
- Never use others' code or submission file. This homework should be done individually (or you will get 0 point).

Important Information

- **You are only allowed to use static word embedding (or you may not get any score).**
- Static word embedding:
  - Skip-Gram & CBOW (a.k.a Word2Vec)
  - Glove
  - fastText
  - Poincaré Embeddings to learn hierarchical representation
- Dynamic word embedding:
  - ELMO (Embeddings from Language Models)
  - ULMFiT (Universal Language Model Fine-tuning)
  - BERT (Bidirectional Encoder Representations from Transformers)
  - GPT & GPT-2 (Generative Pre-Training)
  - Transformer XL (meaning extra long)
  - XLNet (Generalized Autoregressive Pre-training)
  - ENRIE (Enhanced Representation through kNowledge IntEgration)

Important Information

- **You are only allowed to use general modules in deep learning frameworks to build model, which means you can not use packages from paper publications or other organizations (ex: HuggingFace).**
- Example (you can use):
  - Pytorch : torch.nn.LSTM
  - Tensorflow / Keras : tf.keras.layers.LSTM
  - …
- Example (you can **not** use):
  - transformers.BertModel
  - transformers.BertSelfAttention
  - …
- If you are not sure whether the modules you used are valid, please post on e3 forum.

Important Information

- Deadline: 2020/11/10 (Tuesday) 23:59
- Kaggle competition link: https://www.kaggle.com/t/de7aa21921c641cc93c1256a4bbb0e7d
- If you have any question, please post on the e3 forum. TAs will response as soon as possible.
- TAs:
  - Yi-Syuan Chen: yschen.eed09g@nctu.edu.tw
  - Yun-Zhu Song : yunzhusong.eed07g@nctu.edu.tw