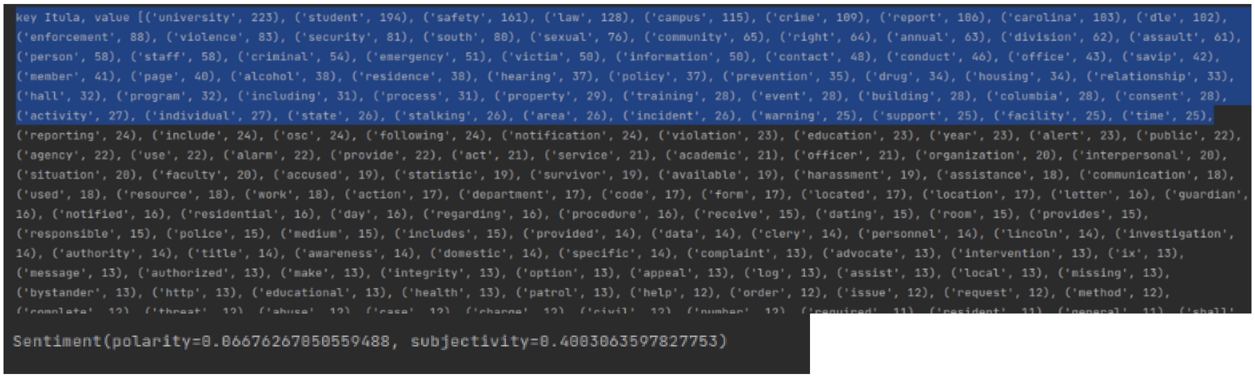
Question1:

- From my perspective, I prefer the LLM-based method on these three tasks under the support of the GPU core, which is unignorable. The key differences between TF-IDF and BERT are as follows: TF-IDF does not take into account the semantic meaning or context of the words whereas BERT does. For sentiment analysis. As opposed to directional models (TF_IDF), which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). For text summarization and topic analysis, both methods achieved a comparable result. However, LLM-based methods such as BERT are able to capture the semantic meaning of words in the context of other words around them. Approaches such as TF-IDF are still iterations of the bag-of-words approach which lays emphasis on the frequency of particular words and is unable to capture the meaning of words in the context of other words surrounding them such as that in a sentence or in a paragraph. As an example, in TF-IDF different words but with the same meaning are treated as separate words, because TF-IDF is unable to capture the relation between words.
- It is obvious that LLM-based methods are expensive. It requires more computation because of its size, which comes at a cost. And also, it is slow to train because it is big and there are a lot of weights to update. It is designed to be input into other systems, and because of that, it has to be fine-tuned for downstream tasks (sentiment mining, topic analysis, etc.), which can be fussy.

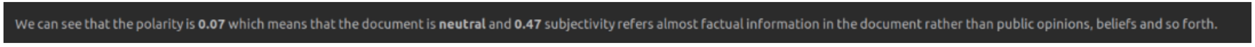
Question2:

- The university is safe. According to the results of text summarization and sentiment analysis, the top 30 words from the entire report shown are relatively towards positive attitude.



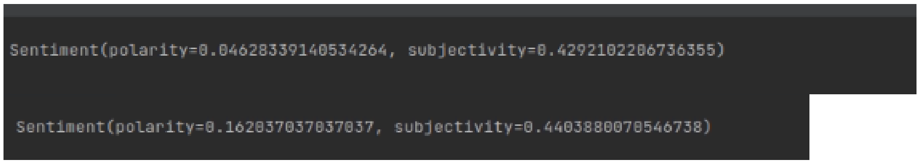
```
key Itula, value [('university', 223), ('student', 194), ('safety', 161), ('law', 128), ('campus', 115), ('crime', 109), ('report', 106), ('carolina', 103), ('dle', 102),
('enforcement', 88), ('violence', 83), ('security', 81), ('south', 80), ('sexual', 76), ('community', 65), ('right', 64), ('annual', 63), ('division', 62), ('assault', 61),
('person', 58), ('staff', 58), ('criminal', 54), ('emergency', 51), ('victim', 50), ('information', 50), ('contact', 48), ('conduct', 46), ('office', 43), ('savip', 42),
('member', 41), ('page', 40), ('alcohol', 38), ('residence', 38), ('hearing', 37), ('policy', 37), ('prevention', 35), ('drug', 34), ('housing', 34), ('relationship', 33),
('hall', 32), ('program', 32), ('including', 31), ('process', 31), ('property', 29), ('training', 28), ('event', 28), ('building', 28), ('columbia', 28), ('consent', 28),
('activity', 27), ('individual', 27), ('state', 26), ('stalking', 26), ('area', 26), ('incident', 26), ('warning', 25), ('support', 25), ('facility', 25), ('time', 25),
('reporting', 24), ('include', 24), ('osc', 24), ('following', 24), ('notification', 24), ('violation', 23), ('education', 23), ('year', 23), ('alert', 23), ('public', 22),
('agency', 22), ('use', 22), ('alarm', 22), ('provide', 22), ('act', 21), ('service', 21), ('academic', 21), ('officer', 21), ('organization', 20), ('interpersonal', 20),
('situation', 20), ('faculty', 20), ('accused', 19), ('statistic', 19), ('survivor', 19), ('available', 19), ('harassment', 19), ('assistance', 18), ('communication', 18),
('used', 18), ('resource', 18), ('work', 18), ('action', 17), ('department', 17), ('code', 17), ('form', 17), ('located', 17), ('location', 17), ('letter', 16), ('guardian',
16), ('notified', 16), ('residential', 16), ('day', 16), ('regarding', 16), ('procedure', 16), ('receive', 15), ('dating', 15), ('room', 15), ('provides', 15),
('responsible', 15), ('police', 15), ('medium', 15), ('includes', 15), ('provided', 14), ('data', 14), ('clery', 14), ('personnel', 14), ('lincoln', 14), ('investigation',
14), ('authority', 14), ('title', 14), ('awareness', 14), ('domestic', 14), ('specific', 14), ('complaint', 13), ('advocate', 13), ('intervention', 13), ('ix', 13),
('message', 13), ('authorized', 13), ('make', 13), ('integrity', 13), ('option', 13), ('appeal', 13), ('log', 13), ('assist', 13), ('local', 13), ('missing', 13),
('bystander', 13), ('http', 13), ('educational', 13), ('health', 13), ('patrol', 13), ('help', 12), ('order', 12), ('issue', 12), ('request', 12), ('method', 12),
('complete', 12), ('threat', 12), ('abuse', 12), ('case', 12), ('charge', 12), ('civil', 12), ('number', 12), ('required', 11), ('resident', 11), ('general', 11), ('shall',

Sentiment(polarity=0.06676267050559488, subjectivity=0.4003063597827753)
```



We can see that the polarity is **0.07** which means that the document is **neutral** and **0.47** subjectivity refers almost factual information in the document rather than public opinions, beliefs and so forth.

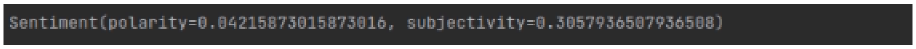
- The rights are balanced. I typed in two groups of rights separately in csv file as input, then got the sentiment score. It can tell that the both groups of policies are polarity, which means they are neutral.



```
Sentiment(polarity=0.04628339140534264, subjectivity=0.4292102206736355)

 Sentiment(polarity=0.162037037037037, subjectivity=0.4403880070546738)
```

- It is better to anonymously report a crime than openly. For this time, I input the paragraph of CONFIDENTIAL REPORTING OF CRIME AND OTHER SERIOUS INCIDENTS for sentiment analysis and text summarization. The results show this paragraph has relatively higher score of subjectivity which means almost factual information in the document rather than public opinions. Thus, I certain the 10-top words which indicate the confidential of openly report a crime.



```
Sentiment(polarity=0.04215873015873016, subjectivity=0.3057936507936508)
```