Assignment 4: Text Data - Group Submission - Returned

Title	Assignment 4: Text Data
Groups	Group B"'
Students	Nikita Kapoor (nikitakapoor), Mona Malik (monamalik27), Sannath Vemula (sannathreddyv)
Submitted Date	Dec 2, 2019 3:15 PM
Grade	**3.40 (max 4.00)**

Instructions

Objective
In this last assignment, you are provided with some raw text data and asked to illustrate some exciting insights hidden within it. In contrast to the previous assignments, you are given a few extra constraints and expected to model the data in a specific way. You have just over two weeks (until next Sunday). Also, this time you will demo your final visualizations to the whole class the Monday one day after the deadline. Still there is no pre-defined "correct solution"; you will demonstrate mastery of text data modelling in the context of doing Data Science.

Data
In the lead-up to the 2016 US presidential election, a Russian "troll factory," the *Internet Research Agency* (IRA), is alleged to have deliberately sought to sow political discontent in the US with inflammatory social media content. Twitter reported to the US congress a list of thousands of Twitter handles associated to the IRA. You are provided with all the tweets produced by these twitter accounts since 2012: https://github.com/fivethirtyeight/russian-troll-tweets. In addition to the raw text content, several dimensions of meta data are provided. You can read more about the origins of the dataset in the blog post describing its release.

Please note the warning in the README that tweet text contains links and these may still lead to active sites with unsavory content.

Implementation
It is expected that you will use Python, as in the previous assignments. Moreover, you will likely want to use the NLTK (natural language toolkit). For data modelling, the rubric assumes that you will use word embeddings (from, e.g., word2vec), but you are welcome to demonstrate the same objectives with respect to another text model (e.g., Brown clusters).

Some of your visualizations should be devoted to word embeddings and bias. These do not need to be tied to your overall story. For the embeddings, you will want to use a tool to generate them from your corpus after some appropriate pre-processing. You will also want to find some "pre-trained" embeddings online with which to compare them.

Post-submission Demo
In class on 25 Nov, each group will have 5 minutes (enforced) to demo their final visualizations to

the entire class, using a different presenter than the previous assignment. Please send a pdf with at most 4 pages by email to me prior to class so that we do not need to connect laptops to the projector. This is not tied to the evaluation; the purpose is simply to share results with each other as inspiration for the final term project.

Submission
You should submit your implementation in raw python or python notebook and a short technical report (10 pages + 1 appendix) in pdf format. The markers will primarily grade the submission based on the report, but may validate that the implementation works as described. The report will describe:

- the insights revealed by the visualization/application
- the preprocessing (cleaning/filtering) done on the text to arrive at the visualizations
- how this submission meets the requirements set out in the rubric (i.e., a justif ied self-evaluation)
- other details that you consider relevant

Grading
The entire group will receive a grade based on how well the submission adheres to the rubric below. Note that your report provides the opportunity to persuade the markers, but that they will ultimately grade according to the rubric.

Assignment 4 Rubric

Component	Weight	4	3	2	1
Story telling	40 %	Visualization is very informative and tells a story about the data; visualizations are clear and easy to interpret without aide.	Visualization is informative, albeit possibly with some support from text	Visualization is complete; information is presented but lacks complexity or depth.	Very difficult to extract any insights from the visualization, or no visualization at all
Data Modelling	20 %	**Both** an advantage and a disadvantage of using pre-trained rather than corpus-specific embeddings is directly visualized	**Either** an advantage or a disadvantage of using pre-trained rather than corpus-specific embeddings is directly visualized	Embeddings are used at some point in the assignment	No apparent use of concepts in text data modelling
Reproducibility	20 %	The visualizations can be **exactly** reproduced from the description in an appendix of the report	The visualization can be **mostly** reproduced from the description in the report, though some minor differences may exist in the final rendering	The visualization can be reconstructed from the report, but requires some **informed guesses**	Not at all clear how the visualizations were constructed

Connection to Research	20 %	**Both** allocation and representation bias are directly visualized	**Either** allocation or representation bias is directly visualized	Some attempt is made to demonstrate bias in the dataset, but this is not clearly depicted in any visualizations	Does not engage with research papers presented in class

Tips
1) Try to load the data into a usable structure early (first few days). You may find it more difficult than you expect to model the raw datasets as workable data structures.

2) Ascend quickly to a working visualization and add complexity later; i.e., first build a minimal viable product (MVP). You will be exposed to new research and new ideas in class as you work on the assignment, so you want to be *agile* with your development patterns. At the same time, even in groups, it may take longer than you expect to go from raw data to a working visualization; so, you don't want to leave this until the last few days before the demo or you could end up with nothing to show at all!

3) Use libraries prolifically.

4) Consider addressing each rubric component individually in your tech report.

5) Reflect closely on the feedback for Assignments 1-3. This is meant to guide you towards a higher grade on subsequent assignments.

Additional resources for assignment

No attachments yet

Submitted Attachments

- Assignment -4 Final Report-2.pdf (2 MB; Dec 2, 2019 3:15 pm)
- Text-Data(Code).zip (385 KB; Nov 28, 2019 12:44 am)
- Assignment -4 Final Report.pdf (2 MB; Nov 28, 2019 12:43 am)

Additional instructor's comments about your submission

Overall
4/3/3/3 = 3.4

Insights
4
+ Overall, lots of interesting insights explored investigating a common theme around the different handle categories. Some tips below to improve:
+ Section 2.1 very good/unique; tip: would be easier to read if the color scale was consistent throughout the section
- Section 2.2, why do the tags *BlackLivesMatter*, *MAGA*, *top*, *FAKENEWS*, *tcot*, *amb* appear twice in the word cloud?

- Visualization of popular account categories over time would probably be much easier to read as a stacked bar chart
- At times the analysis seems to generalize too much, i.e., seems to extract general points about entire election from just this small subset of tweets supposedly from IRA

Modelling
3
+ An advantage of pre-trained model directly shown in terms of mitigating bias
+ Advantage of corpus-specific model not very clear from description in section: some mention of being able to retrieve matches with higher score, but not clear enough what this means

Reproducibility
3
+ Many useful preprocessing steps described
- Some preprocessing difficult to reproduce from the report alone; e.g., it is unclear here what qualifies as punctuation (does *won't* include punctuation, because if you remove the apostrophe it becomes a different word?) or as a stop word (is some list from some particular version of some particular library used?).

Research
3
+ Interesting example of representation bias visualized on corpus-specific model
- No discussion of allocation bias
= Makes use of word-scale graphics from viz paper, but not relevant to the rubric on *this particular* assignment