Exam text analysis

2020-11-27

About this exam

- The exam consists of three questions. You can earn 100 points in total. Each question states how many points you can earn with it.
- The deadline for the exam is 4 December 2020, 17.00.
- You are expected to work on this exam alone. We will be strict on submitting non-original work.
- Using code from the manual is allowed.
- Using some functions from the other sources (e.g. from StackOverflow) is also fine with us, but be sure to add a short comment with the source and the reason you need that function.
- You'll both be graded on your code as well as on the quality of your interpretations and your ability to reflect on the used methods. For the first two questions, we expect you to use a specific technique; for the third one, you are free to choose one.
- We don't expect you to perform *statistical* analyses we haven't discussed in this course to interpret the results. If you're unsure whether a result would be statistically significant, just include that in your answer.
- Each question has a guideline on how much you should write. There is no minimum number of words, and if you need less than the guideline to answer the question, that is no problem.
- The questions must be submitted in 3 separate files on Blackboard. Make sure your Jupyter Notebooks are well-ordered and you do not print the complete tokenized corpora. Submitting messy Jupyter Notebooks can result in the deduction of points.

Question 1

`30 points` Compare the use of clickbait titles between democrats and republicans in `framing.p`. How many times do democrats refer to an article with a clickbait title and how many times do republicans do? Inspect the titles in the dataset that were classified as clickbait and try to explain the results.

(Hint: consult the manual to see how to classify the title of the articles as clickbait/non-clickbait!)

Your answer must consist of the following:

- The complete code to answer the question with a short comment for every step (ca. 2 sentences per step)
- An answer to the question + explanation (ca. 200 words)

Question 2

`30 points` Choose two genres from the song dataset from exercise 3.1 to examine and compare the gender bias in songs of both genres. Explain why the two genres you choose are relevant to compare in this context, and formulate a hypothesis.

Train two word embeddings models (one for each genre), and use the lists of female and male words uploaded to BB (Assignments / exam 1 / question 2) for your analysis. Compare the biases between the two genres you choose using the method by Wevers. Interpret the results and relate them to your hypothesis.

For your reference, the columns in `word_cats.p` represent the following categories:

- affect: Affect
- posemo: Positive emotions
- negemo: Negative emotions
- social: Social
- family: Family
- cogproc: Cognitive Processes
- percept: Perceptual Processes
- body: Body
- work: Word
- leisure: Leisure
- money: Money
- relig: Religion
- occupation: Occupation

Your answer must consist of the following:

- A statement on the relevance of your comparison and the hypothesis (ca. 150 words)
- The complete code to answer the question with a short comment for every step (max 2 sentences per step)
- Interpretation and conclusion (ca. 200 words)

Question 3

`40 points` A television producer has approached you with the question whether they should release the new season of their show all at once, like Netflix does, or once a week. As their market research has shown that both release strategies will result in more or less the same ratings, they want to know which release strategy will **engage** their audiences more; which release strategy will result in more (**valuable**) discussions.

Try to formulate a good operationalization of this question using the methods we discussed in the last three weeks, and argue why this operationalization would be suitable to formulate a substantiated advice for the television producer.

Then implement your operationalization using `discussions.p` (where the column 'type' indicates whether a show is released all in once [value 'netflix'] or linearly [value 'linear']). Try to formulate a substantiated advice for the television producer. If your method doesn't produce meaningful results, try to formulate suggestions on how to improve the method you proposed instead.

Note: you will **not** be graded on the extent to which your proposed method actually produces valuable results, but on your thought process and argumentation. Don't try to fine-tune your method until it spits out something interesting.

Your answer must consist of the following:

- An operationalization of the question (ca. 350 words)
- The complete code to answer the question with a short comment for every step (max. 2 sentences per step)
- Interpretation and conclusion (ca. 200 words)