Exam text analysis

2021-12-03

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 10 December 2020, 17.00. If you submit your exam late, you lose 10 points for every day you submit late (with the first day starting at 17.01). If there's a valid reason you are unable to submit the exam on time, contact us *before* the deadline.
- You are expected to work on this exam alone. We will be strict on submitting non-original work. This applies to the textual answers as well as to the code.
- Using code from the manual is allowed.
- Using some functions from the other sources (e.g. from StackOverflow) is also fine with us, but then you have to add a short comment with the source (url) 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 general 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 .ipynb files on Blackboard. Make sure your Jupyter Notebooks are well-ordered, your textual answers are written in markdown, and you do not print the complete tokenized corpora (so make sure to check the file size of your notebooks before submitting). Submitting messy notebooks can result in a deduction of 5 points per question.

Question 1

`30 points` Train a topic model on the `description` column in `framing.p`. Describe and explain your pre-processing steps and parameters.

Choose two news media from the column `netloc` (e.g. vox.com and breitbart.com) that are relevant to compare and explain your choice. Compare the topic proportions between those two news media. Try to explain your results, also given what you know about how the data set is constructed.

Your answer must consist of the following:

- Explanation of preprocessing steps + parameters (150 words)
- Motivation of the two news media (50 words)
- The complete code to answer the question with a short comment for every step (ca. 2 sentences per step)
- Interpretation and discussion (ca. 200 words)

Question 2

`30 points` Choose two shows from `discussions.p` and compare the gender bias in the discussions on both shows. Explain why the two shows you choose are relevant to compare in this context and formulate a hypothesis.

Train two word embeddings models (one for each show). Compile a list of male and female related words that you deem relevant for your corpus and compare the gender biases between the discussions on the two shows you choose, using the method of the paper by Wevers. Interpret the results, relate them to your hypothesis, and discuss whether your results say something about the shows themselves or about the *discussions* on these shows.

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 discussion (ca. 200 words)

Question 3

`40 points` A record producer has approached you with the question of whether there are any distinguishable features of a pop song. In other words: what sets a pop song (lyrically) apart from other genres?

Try to formulate a good operationalization of this question (how are you going to quantify this, what method do you use, what steps do you need to take, what time period are you focusing on) using the methods we discussed in the last three weeks (you are free to choose one!), and argue why this operationalization would be suitable to formulate an answer to the question of the record producer.

Implement your operationalization using the dataset from exercise 3.1 and formulate an answer to the question of the record producer.

After that, consider the following new lyrics provided by our hypothetical producer (copy-pastable lyrics here):

```
The sky breaks open and the rain falls down
Oh, and the pain is blinding
But we carry on...
Seems like the end of everything
When the one you love
Turns their back on you
And the whole world falls down on you...
It's the end of the world
It's the end of the world
Well, I'm holding on...
But the world keeps dragging me down...
It's the end of the world
It's the end of the world
Well, I'm holding on...
But the world keeps dragging me down...
I got my heart on lockdown
And my eyes on the lookout
But I just know that I'm
Never gonna win this
They can keep the lights on
They can keep the music loud
I don't need anything
When I got my music
And I'm holding on...
The sky breaks open and the rain falls down
And the pain is blinding
```

```
But we carry on
(They tell you lies)
I'm holding on...
(They tell you lies)
```

Based on (the results) of your method, to what extent could you say whether this is a ”classic” pop song or not? Explain your answer (in a data-driven way) and try to refer both to your results and to the lyrics itself.

Your answer must consist of the following:

- An operationalization of the general question of the record producer (ca. 250 words)
- The complete code to answer the question(s) with a short comment for every step (max. 2 sentences per step)
- Interpretation and conclusion (ca. 200 words)
- Answer the question/explanation about the new lyrics (ca. 100 words)