Project 1_text_2021

(group project)

Sentiment classification-Scikit-learn

You are asked to train and test a few models, and produce what you think the best settings are for the data we're using.

Please submit any raw results, figures, or code you come up with at the end. There is .txt files that contains a corpus of reviews. Each review is on one line and is headed by two meaningful tags:

1. A tag that specifies one of **six topics (labels)**: books, camera, dvd, health, music, software.
    - The labels we are interested in are the six denoting the topics: books, camera, dvd, health, music, software
2. A tag which indicates the sentiment expressed by the review, in terms of a **positive or negative value**: **pos, neg**.
3. A third column contains the **id**, and the rest is the review's text. The text has already been tokenised

You will be using both the sentiment as well as the topic labels. Use default settings for your model and report results. Try to change the value of the C parameter and see if you can get better results validation or on a development set, just use the same settings as above). Experiment with changing the values and different models, and report what you observe.

❓ The questions that you need to find answers for them, are as follows:

1. Which model do you select for the purpose of this study (Naive bayes, SVM, ???)? Why?
2. What is the best model for your experiment?
3. What are your default settings?
  3.1. You can use default settings: what is the result of it?
4. From the whole dataset select reviews belonging to just two classes — you can choose any two classes (e.g., music and health, or dvd and books or whatever you like; you can even experiment with different pairs).
5. Check if the distribution of sentiment across classes is evenly distributed, if yes, do you think it's matter which two classes you're picking
  5.1. Run the model, do you get better scores when the gold labels are the sentiment labels or the topic labels?
6. Based on your observations and results on data or on your development set, make a final decision, and report it in the document. You are expected to experiment with adding features, too. Possibly n-grams might help? Or part-of-speech information?
7. You can also come up with your own interesting questions to answer with this dataset (It has an extra mark for you 😉).

****Please compile the results of your analysis into a report (containing any results, graphs or explanations you come up with) and submit any code you used to analyze the data.**

What you have to hand in, as a team:

- Your code: You have to assume that we will run your code like this:
  DSHproject1_text.py / .ipynb
- Research report (based on the template that we provided to you at the beginning of the course) that describes what you've done and also includes answers/discussion to all questions in this document. Please, make sure you submit a pdf file.
- As this is a group work, please include a section (same as the latex report template) at the end where you clarify who did what.