Project COMP 262: Natural Language processing and recommender systems

Introduction

Throughout this two-phase project assignment, each team needs to construct "a sentiment analysis model for products based on customers' textual reviews," using both a Lexicon approach and a machine learning approach.

First phase involves uploading the data, cleaning it up, pre-processing the data in order to create a textual representation, and finally, building and testing the Lexicon classifier. In the second phase, the team needs to construct the same procedure using a Machine learning approach and compare the results of each approach. Lastly, a study of how to utilize the same review data to construct a recommender system is required.

The project would be governed by a set of deliverables per phase and there are certain check points with the professor, as illustrated in the project timetable key-milestones section.

deliverables will be evaluated based on rubric illustrated in the Rubric section.

A project plan should be built by the team and updated on a weekly basis, in addition, a simple log of all team meetings should be maintained. Both should be submitted with final project documentation and code as appendices to the project report.

At the end of each phase, the team needs to present their work to the class.

Grading is both at the team level and at the individual level.

Data sets

We will use the Amazon product review datasets available at:http://jmcauley.ucsd.edu/data/amazon/ we will use the small review subsets referenced as the k-core.

Each team will tackle one dataset, as follows:

Team #1: Beauty

Team #2: Automotive

Team #3: Musical Instruments

Team #4: Office products

Team #5: Digital Music

Please reference the publishers of these datasets, in your report.

Deliverables:

Phase #1

1. Dataset data exploration: List the main finding of the dataset. Be thorough and creative. For example, look at:
    a. Counts, averages
    b. Distribution of the number of reviews across products
    c. Distribution of the number of reviews per product
    d. Distribution of reviews per user
2. Text basic pre-processing:
    a. Randomly select 500-1000 reviews from your dataset and perform steps b through d.
    b. Label your data based on the value of "rating of the product" i.e. as follows:
        i. Ratings 4,5: Positive
        ii. Rating 3: Neutral
        iii. Ratings 1,2: Negative
    c. Chose the appropriate columns for your sentiment analyzer. (Give this some thought)
    d. Split the data into 70% for training and 30% for testing,—Use stratified splitting based on the rating value field.
3. Text representation: Represent your text using one of the approaches explained in module #2. Justify why you chose that approach.
4. Modeling (Sentiment Analysis) Lexicon approach:
    a. Build two sentiment analysis models using 70% of the data. Use one of the following Lexicons packages to build your models:
        i. Valence Aware Dictionary and Sentiment Reasoner (VADR) you can find out more information here: https://github.com/cjhutto/vaderSentiment
        ii. TextBlob you can find out more information here: https://textblob.readthedocs.io/en/dev/quickstart.html
        iii. SENTIWORDNET you can find more information here: http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf
5. Testing: Test out the model using the 30% test data note the accuracy, precision, recall and F1 score.
6. Presentation: Check project presentation requirements.
7. Project report: Check project report requirements/ phase #1
8. Submit a documented code with reference to any external dataset.

Phase #2

9. Modeling (Sentiment Analysis) Machine Learning approach:
    a. Build two sentiment analysis models using 70% of the data. Choose two of the following Machine Learning algorithms to build your models:
        i. Logistic Regression

ii. SVM
iii. Naïve Bayes
iv. Gradient Boosting

10. Testing: Test out the two models using the 30% test data note the accuracy, precision, recall and F1 score.
11. Compare the test results of the Lexicon model versus the two machine learning models.
12. Review the attached paper "Recommender systems based on user reviews: the state of the art", can also be accessed at the centennial library. Examining the options presented in the paper carryout the following:
    a. Explain how you can enhance the rating values of your data using the review data.
    b. Provide diagrams and pseudo-code: implementation is not required.

Timetable – key milestones

Milestone	Week #
Project teams assembled, and datasets assigned	3
Check point # 1 "Data exploration & pre-processing" progress	5
Check point # 2 "Text representation results"	6
Presentation & submission phase #1	8
Check point #3 progress on modelling	12
Presentation & submission phase #2	14

Peer-evaluation

With every phase submission, each team member should fill in the peer evaluation form and submit it to the assessment box named "Peer evaluation Phase X", where X is 1 or 2. This form is confidential, and only the professor will access it. In summary, this form is to express what each team member has worked on and how the team member views the contribution of the rest of the team members. If all team members have contributed equally, then give all a rate of 100%, if a team member did not contribute then give a 0%, finally, if a team member contributed but not to the level of the team agreement, then a score between 1% to 99%.

Project Report requirements:

1. Cover page
2. Table of contents
3. Detailed results of dataset exploration & conclusions
4. Dataset pre-processing steps with explanation and justification of choices.
5. Text representation model with explanation and justification.
6. **Models**; per model clarify:
    a. Assumptions/Heuristics/algorithms used
    b. Explain each model, how it works
    c. List any external datasets
7. Testing results summary.
8. Future work: Suggested recommender design, refer to deliverable # 12
9. Final conclusion.
10. Assumptions.
11. References.
12. Appendix 1: Project plan.
13. Appendix 2: Meeting register, simple table showing date and time of each meeting, who attended, subjects discussed and assignments.

Note: phase #2 deliverables are appended to the phase #1 report (i.e. Only one report for the whole project).

Presentations requirements:

1. All team members need to participate.
2. Present working code.
3. Present power point summarizing key points related to the project.

Rubric

Evaluation criteria	**Not acceptable**	**Below Average**	**Average**	**Competent**	**Excellent**
	0% - 24%	**25%-49%**	**50-69%**	**70%-83%**	**84%-100%**
Dataset data exploration Phase #1	Data exploration completely missing or what is submitted is below 30% with no relationship analysis.	Only 50%-60% of dataset attributes have been explored or exploration not complete on # of missing values, only a few relationships are captured, minimum visualizations.	Only 60%-70% of dataset attributes have been explored or exploration not complete on # of missing values not all relationships are captured.	Most dataset attributes columns have been explored and a complete description of each attribute value meaning has been reported in addition to exploring some relationships between attributes and presented a few visualizations.	All dataset attributes columns have been explored and a complete description of each attribute value/meaning/distribution has been reported in addition to exploring all relationships between attributes supported by a complete set of visualizations.
Text basic pre-processing Phase #1	Data not pre-processed No comments explaining code.	Some major errors in the data model. Issues with sampling labelling. Outliers not addressed. Normalization not implemented as needed. Minor comments are implemented.	Some errors in the data model. Issues with sampling labelling. Outliers not addressed. Normalization not implemented as needed. Some code is correctly commented.	Correct sampling, labeling and splitting of data. Data outliers are cleaned up as needed, normalization/standardization is implemented as needed. Appropriate text pre-processing is implemented Selection and build of the data model not justified. selected attributes. Majority of code is correctly commented.	Correct sampling, labeling and splitting of data. Data outliers are cleaned up as needed, normalization/standardization is implemented as needed. Appropriate text pre-processing is implemented. Logical selection/merging and justification of selected attributes. All code is correctly commented.
Text representation Phase #1	Missed to represent the text completely.	Shows some thinking and reasoning but text representation not suitable for the nature data/task.	Text representation model can work but not the best for the nature of the data/task.	Suitable text representation without justification clearly explained	Suitable text representation with justification clearly explained.
Modelling Phase #1 Phase #2	Majority of Models are not implemented.	Some models are implemented with errors.	Majority of models are implemented but not with optimal hyperparameters.	All models are implemented correctly but not with optimal hyperparameters.	All models are implemented correctly.
Testing Phase #1 Phase #2	No model evaluations conducted	Some metrics are generated for each model, with no	Some metrics are generated for each model, with minimum	All metrics are generated for each model and a comprehensive	All metrics are generated for each model and a comprehensive comparison

		comparisons/conclusions presented.	comparisons presented with partial conclusions.	comparison presented with partial conclusions.	presented with clear conclusions.
Project report Phase #1 Phase #2	Writing lacks logical organization. It shows no coherence, and ideas lack unity. Missing most conclusions or assumptions or references Serious errors. No transitions. Format is very messy.	Writing lacks logical organization. It shows some coherence but ideas lack unity. Serious errors. Missing many conclusions or assumptions or references Format needs attention, some major errors.	Writing is coherent and logically organized. Some points remain misplaced. Missing many conclusions or assumptions or references Format is neat but has some assembly errors.	Writing is coherent and logically organized, with transitions used between ideas and paragraphs to create coherence. The overall unity of ideas is present. Missing some conclusions or assumptions or references. Format is neat and correctly assembled.	Writing shows a high degree of attention to logic and reasoning of all points. Unity clearly leads the reader to the conclusion. Covers all deliverable results. Covers all assumptions and conclusions. Includes references. Format is neat and correctly assembled with a professional look.
Presentations Phase #1 Phase #2	Very weak, no mention of the code changes. Execution of code not demonstrated. Some team members do not participate.	Some parts of the code changes are presented. Execution of code partially demonstrated. Some team members do not participate.	All code presented but without explaining why. Some parts of the code demonstrated is not working and have errors. Some team members do not participate.	A comprehensive view of all code demonstrated presented with an explanation, exceeding the time limit. Working code demonstrated. All team members participated but without equal participation. Some team members are not confident of their input.	A comprehensive view of all code demonstrated in working condition with explanation, within the time limit. All team members participate equally and are confident in their responses.