CSCE 771: Computer Processing of Natural Languages

Prof. Biplav Srivastava, Fall 2024

Quiz 3 / Instructions

- This is a in-class exercise. Some programming is needed.
- Code/ data has to be submitted in a directory of your GitHub as mentioned below. For answers, they should be in Github in sub-folder "Quiz3". (They can also be uploaded to BB).
- All work is due by the end of class. Extra time is allowed till end of day with 20% penalty.
- Total points = 50 + 50 = 100 points
- Obtained =

Student Name:

Rasoul Nikmehr

In this in-class Quiz, we will analyze a NLP dataset in class given by the instructor and also your own dataset.

a) For the instructor dataset, consider QA for SC. It is available from:
https://docs.google.com/spreadsheets/d/1IZZoW_xud546eknMBqXAfvdHNLO3UTRs/edit?usp=sharing&ouid=108698138157665592408&rtpof=true&sd=true
For SC, source urls are:
* Official - https://scvotes.gov/voters/voter-faq/
* Secondary - https://www.vote411.org/south-carolina
b) For yours, it is the state you had selected.

Q1: Prepare the datasets as per guidance discussed in last class [25 + 25 = 50 points]

- Format in .json; name file as "**xy**_qa.json", where xy is the two-character US state acronym
- Fixed attributes in .json
  - state: xy
  - num_questions: **a**, where a is the number of questions
  - num_answers: **b**, where b is the number of answers
  - contributor: student name
- questions: List of Q/As with attributes for each it:
  - q // question
  - a // answer
  - s // source url from where the information is taken
  - t // time when the information is taken – UTC format
- Store it in your github repo; put in sub-dir like "project/data"
- Instructor will keep it in common place inside course github repo and share.

1. for dataset (a) [25 points]
2. for dataset (b) [25 points]

Q2. Comparison the QA datasets [10 + 10 + 10 + 10 + 10 = 50 points]
Select two criteria to compare text. It can be raw statistics, English quality (typos, sentiments, ..) or other criteria.

1. What are the two criteria you chose and how will you measure them: [5 + 5 = 10 points]

Criteria 1: Lexical Diversity. Measuring the ratio of unique words to the total number of words in the text.
Criteria 2: Response Length. It measures the number of the characters or words in the answer.

2. Compare the questions (Q) of the two states based on the two criteria [10 points]
SC Lexical Diversity (Questions): 0.9692450502107214, IL Lexical Diversity (Questions): 0.8982305333490367
SC Average Question Length: 12.264705882352942 words, IL Average Question Length: 29.854368932038835 words

3. Compare the answers (A) of the two states based on the two criteria [10 points]
SC Lexical Diversity (Answers): 0.7521632085464837, IL Lexical Diversity (Answers): 1.0
SC Average Answer Length: 76.94117647058823 words, IL Average Answer Length: 2.3300970873786406 words

4. Compare each answer with respect to its correspond question on the two criteria [10 points]
SC Avg Lexical Difference (Answer - Question): -0.21708184166423752, IL Avg Lexical Difference: 0.10176946665096316
SC Avg Length Difference (Answer - Question): 64.67647058823529 words, IL Avg Length Difference: -27.524271844660195 words

5. Which state has been QA dataset based on the two criteria: yours or instructor provided (SC) and why? : [5 + 5 = 10 points]
Based on the two criteria, SC has better QA dataset. Baseed on the Lexical Diversity analysis, SC questions have more lexical diversity(0.97) compared to Illinois.
Illinois has a lexical diversity in answers but the answers are short.(2.33 words on average).
Based on Response Length,SC answers are longer and provides more detailed.But Illinois answers are short.
Also, SC has longer length difference between answers and questions.
SC provides a better QA dataset.