ARK Undergraduate/Masters Demonstration of Research Interest

Philosophy. Undergraduate and masters students have been an important part of my research group since I started it. I encourage my Ph.D. students and postdocs to mentor junior scholars and involve them fully in our research, and we take a lot of pride in the careers of our former undergraduate/masters mentees. As interest in NLP has increased over the years, I've introduced this task as a point of entry to research with my group. By completing the task, you will demonstrate your interest in working with us and let us see a little bit about how you approach problems. If you find the task fun, that's probably a good sign that you'll enjoy working on research with us.

Diversity leads to stronger science, and I actively seek, welcome, and encourage people with diverse backgrounds, experiences, and identities to apply. Scholars who self-identify as: women, people of color, non-binary or genderqueer, transgender, people who have lived in poverty, people with disabilities, immigrants, religious minorities, and lesbian, gay, bisexual, and queer are strongly encouraged to apply. I also believe that strong science doesn't happen in a rush. In the early stages of research, a person needs time to explore and learn. This is what we emphasize in undergrad/masters research projects in my group: exploration and learning. Publication is a secondary goal, something we focus on only after we believe we have a discovery worth sharing with the larger community.

1 Preliminaries

In a document named `aboutme.pdf`, please:

- Tell me a little bit (a paragraph or so) about why you're interested in working on research.
- If there's a specific problem you want to work on with my group, feel free to describe it in (no more than a half page or so). This is optional!
- List relevant coursework you've completed (e.g., CSE 447, CSE 446, courses in Linguistics, and anything else you think might be relevant—there are no wrong answers) along with the quarter, instructor, and grade you earned.
- Tell me the program you're enrolled in, the university and campus, and when you expect to graduate.
- Optionally, you may include a one-page resume as part of this file.

2 Exercise

No part of this task is a "standard" NLP problem that you would learn about in an NLP class; there is no known right answer, and there are many ways to tackle each part. It's not expected that you'll spend more than about one day on the whole problem. This problem is possibly *very hard*, and it is not expected that you will manage to solve it perfectly well. The purpose of this exercise is to see how you approach the problem and how well you execute the solution. Your written answer is more important than any quantitative measure of performance. **Please work alone on your solution; if you discuss it with someone else, that's okay, but you must acknowledge their help in your writeup.**

There are three parts. **Please read all instructions carefully and format files exactly as described here (even minor deviations make it harder for me to evaluate your solution).** The data you need is in `https://nasmith.github.io/files/challenge-data.tgz`.

2.1 Is it English?

Each instance in the provided training set is a pair of strings; one is a naturally occurring English sentence, found in the wild. The other is a corruption of that sentence. At training time, you are told which is which, and at test time, your system must guess. The training set is provided in `train.txt`. Each line contains the English string and its corruption, separated by a tab character. The test set is in `test.rand.txt`, which is formatted the same way except that the original and corrupted strings are presented in random order. You may use any additional resources or tools to build your classifier, but **you may not use any additional data (or derivatives of other data)**. That includes pretrained models or word vectors of any kind.

Your solution should be a plaintext file called `part1.txt`. It should have one line per test instance. Line i should contain a label, either the character `A` (indicating that the string on line i *before* the tab character is English) or the character `B` (indicating that the string on line i *after* the tab character is English), followed by a newline. Each pair in the test set has a correct answer; your goal is to get as many of these instances right as you can. Some pairs may be nearly impossible. Note that if your submission is formatted incorrectly, you will not get any of them right, because an automatic script will be used to assess your accuracy. For reference, one recent time I used this exercise, the median score attained was 83%.

2.2 Ruin English

Now, you are tasked with *creating* a dataset like the one above. Use the original strings from the first tab-separated column of `train.txt` as your original strings. Your solution should be a plaintext file called `part2.txt`. It should be formatted just like the training set in part 1, but with *your* corruptions in the second tab-separated column. You may use any additional resources or tools, **but not data or derivatives**. Your goal is to produce sentences that will challenge a system like the one you built in part 1. (You might want to use your own part 1 system as a check.) Note that if any line of your solution is identical to its corresponding line in the input, you will have failed.

2.3 Write in English

Finally, describe your methods for the first two parts, and your thinking behind your choice. Please write succinctly and clearly. If you used any code written by others, or discussed ideas with others, acknowledge them. Your report should be about one page total and should be submitted as a pdf named `part3.pdf`.

What to Submit

In a gzipped tarball that has your name in the filename, submit: (1) `aboutme.pdf`; (2) `part1.txt` and `part2.txt` as described above; (3) your source code for parts 1 and 2; and (4) `part3.pdf`, your one-page writeup. Submit by emailing a URL for your tarball to Noah Smith (subject: "Research Interest Demonstration").

Deadlines. Always 11:59pm Pacific time, and

Sept. 7	for autumn quarter;
Dec. 7	for winter quarter;
March 7	for spring quarter;
June 7	for summer.