ARK Undergraduate Research Challenge Problem (February/March 2018)

Submit by emailing a URL for your tarball to Noah Smith. Deadline: March 19, 2018, 11:59pm.

This problem is offered as a way to demonstrate your interest in working with my group on research. No part of it 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 can get your submission disqualified.** The data you need is in `https://homes.cs.washington.edu/~nasmith/temp/spring2018.tgz`.

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 `spring2018.train.txt`. Each line contains the English string and its corruption, separated by a tab character. The test set is in `spring2018.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.

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, the last time I used this exercise, the median score attained was 83%.

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 `spring2018.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.

Your goal is to produce sentences that will make it challenging for a system like the one you built in part 1 to perform well. (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.

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`.

If there's a specific problem you want to work on with my group, feel free to describe it in a second page.

What to Submit

In a gzipped tarball that has your name in the filename, submit:

- `part1.txt` and `part2.txt` as described above;
- your source code for parts 1 and 2; and
- `part3.pdf`, your one-page writeup.