Programming assignment 3: Language Modeling

CSE 156: Statistical NLP: Spring 2022

University of California, San Diego

Released: April 18, 2022

Due: April 25, 2022

In this programming assignment, you will build probabilistic language models. Your task is to analyze texts from different domains using your language models.

1 Preliminaries

1.1 Data and Code

We provide three corpora to conduct the evaluation (`Brown, Gutenberg, and Reuters`), summarized below, you are encouraged to examine them.

- **Brown Corpus:** The objective of this corpus is to be the standard corpus for present day (1979 -) American English [1].
- **Gutenberg Corpus:** Contains a selection of text from public domain works by authors including Jane Austen and William Shakespeare [2].
- **Reuters Corpus:** A collection of financial news articles that appeared on the Reuters newswire in 1987 [3].

Some initial source code is provided. The interface and a simple implementation of a language model is available in *lm.py*, which you can extend to implement your models. In *generator.py*, we provide a sentence sampler for a language model. The file *data.py* contains the main function, that reads in all the train, test, and dev files from the archive, trains all the unigram models, and computes the perplexity of all the models on the different corpora. The README file provides a little bit more detail. The code has been tested on **Python 3**. Feel free to ignore any part of the code that you do not find it useful.

2 Unigram Language Model Analysis (4.5 points)

The language model provided in the starter code is a unigram one — it is a terrible language model as it completely ignores context. Nevertheless, you can still analyze its behaviour.

2.1 Analysis on In-Domain Text (2 points)

The starter code trains a unigram language model for each of the domains. Analyze them as follows:

- **Empirical Evaluation:** Provide empirical analysis of the performance of the unigram model, such as perplexity as amount of training data is varied (the different domains have different training data sizes you can note their performance, it is not required to retrain the models on different data splits, although you are welcome to do so).

[1] `http://www.hit.uib.no/icame/brown/bcm.html`

[2] `http://gutenberg.net/`

[3] `https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection`

2.2 Analysis on Out-of-Domain Text (2.5 points)

In this part, you will analyze performance of the unigram models on text from a domain different from the one it was trained on. For example, you will be analyzing how a model trained on the Brown corpus performs on the Gutenberg text.

- **Empirical Evaluation:** Inspect the perplexity of all three of the models on all three domains (a 3× 3 matrix, as computed in *data.py*). Do you see a trend? If so, explain what this trend means.

3 Implement a Context-aware Language Model (8)

Implement a language model that considers context. You are free to pick any type of language model covered in class. We recommend implementing an n-gram language model, it should at least use the previous two words, i.e. a trigram model. Use appropriate smoothing to ensure your language model outputs a non-zero and valid probability distribution for out-of-vocabulary words as well. In your write up, define and describe the language model in detail (saying "trigram+laplace smoothing" is not sufficient). Include any implementation details you think are important (for example, if you implemented your own sampler, or an efficient smoothing strategy). Also describe what the hyper-parameters of your model are and how you set them (you should use the dev split of the data if you are going to tune it).

- Compute the perplexity of the test set for each of the three domains (the provided code does this for you), and **compare it to the unigram model**. If it is easy to include a baseline version of your model, for example leaving out some features or using only bigrams, you may do so, but this is not required.
- **Show examples of sampled sentences to highlight what your models represent for each domain**. It might be quite illustrative to start with the same prefix, and show the different sentences each of them results in. You may also hand-select, or construct, sentences for each domain, and show how usage of certain words/phrases is scored by all of your models (function *lm.logprob_sentence()* might be useful for this).
- Perform Out-of-Domain Text Analysis (Empirical): Include the perplexity of all three of your models on all three domains (a 3× 3 matrix, as computed in *data.py*). Compare these to the unigram models, and your baselines if any, and discuss the results (e.g. if unigram outperforms one of your models, why might that happen?). Include additional graphs,plots, tables to support your analysis.
- Perform Out-of-Domain Text Analysis (Qualitative): Provide an analysis of the above results. Why do you think certain models/domains generalize better to other domains? What might it say about the language used in the domains and their similarity? Provide graphs, tables, charts, examples, or other summary evidence to support any claims you make.

4 Adaptation (Bonus: 2 points)

Suppose you trained a model on corpus A and wish to test it on the test set for corpus B. Design an approach for using a small fraction of corpus B's training data to adapt the model trained on corpus A. How does it influence performance (for example, does it outperform the initial model trained just on corpus A?) How close can you get to performance when training on corpus B's full training set? Note: *if you didn't manage to complete Section 3.3, you can still perform this section with the unigram language model*

5 Submission Instructions

Submit on Gradescope (more details on Gradescope to be provided).

- **Code:** You will submit your code together with a neatly written README file to instruct how to run your code with different settings. We assume that you always follow good practice of coding (commenting, structuring), and these factors are not central to your grade.
- **Report:** As noted above, your writeup should be *four pages long, or less, in pdf* (reasonable font sizes). Part of the training we aim to give you in this class includes practice with technical writing. Organize your report as neatly as possible, and articulate your thoughts as clearly as possible. We prefer quality over quantity. Do not flood the report with tangential information such as low-level documentation of your code that belongs in code comments or the README. Similarly, when discussing the experimental results, do not copy and paste the entire system output directly to the report. Instead, create tables and figures to organize the experimental results.

6 Report Details

Make sure your report includes the following items:

6.1 Unigram Language Model Analysis

a. Find trends in both sections 2.1 and 2.2, and explain what the trends mean.

6.2 Context-aware LM: Implementation

a. Clear description of model; and smoothing method

6.3 Context-aware LM: Analysis of In-Domain Text

a. Report your perplexity values; compare with unigram model and discuss your observations

b. Show performance w.r.t hyperparameters

c. Show examples of sampled sentences

6.4 Context-aware LM: Analysis of Out-of-Domain Text

a. Report your perplexity values; compare with unigram model and discuss your observations

b. Compare performance on different corpora and discuss your observations

6.5 Adaptation

a. Discuss your approach, show relevant comparisons

7 Acknowledgements

Adapted with changes from similar assignments by Yejin Choi, Noah Smith and Sameer Singh.