CS4740: Intro to NLP
Project1: Language Modeling and Word Embeddings

Part 1 (optional due date for feedback) **: Sunday, Sept 16, 11:59PM**
Part 2: Due **Monday, Sept 24, 11:59PM**

1 Overall Goal

In this project we will build an **n-gram-based language model** for **text classification** and also investigate a different representation for words using **word embeddings**.

The project is divided into parts. In Part 1, we will get started with the steps towards building the language model using a dataset of Obama's and Trump's speeches. In Part 2, you will (a) use this language model to classify speeches and (b) intrinsically evaluate word embeddings using a word analogy task and also extrinsically evaluate them on speech classification.

Logistics: You should work in groups of 2 or 3 students. Students in the same group will get the same grade. Thus, you should make sure that everyone in your group contributes to the project. Also, **remember to form groups on BOTH CMS and Gradescope**. You can use Piazza for finding your project partners.

Advice: The report is important! The report is where you get to show that you understand not only *what* you are doing but also *why* and *how* you are doing it. So be clear, organized and concise; avoid vagueness and excess verbiage.

2 Unsmoothed n-grams

To start, you will write a program that computes **unsmoothed unigram and bigram probabilities** from an arbitrary corpus. You can use any programming language(s) you like.

Assume that you are given raw (ascii) text as input (see 2.1). You may need to do tokenization, based on the design decisions you make. You may use existing tools just for the purpose of preprocessing (e.g. word tokenization) but you must write the code for gathering n-gram counts and computing n-gram probabilities yourself. For example, consider the simple corpus consisting of the sole sentence:

the students liked the assignment

Part of what your program would compute for a unigram and bigram model, for example, would be the following:

$$
\begin{aligned}
&P(the) = 0.4 \\
&P(liked) = 0.2 \\
&P(the|liked) = 1.0 \\
&P(students|the) = 0.5
\end{aligned}
$$

2.1 Dataset

You are given (via Piazza) a corpus of Obama's and Trump's speeches. In the DATASET folder you will find folders including training corpora for Obama's speeches and Trump's speeches, respectively. **You should consider Obama's speeches and Trump's speeches as separate corpora and generate a language model for each.**

2.2 Preprocessing

The `obama.txt` and `trump.txt` files included in each of the subfolders in the DATASET folder are already tokenized and hence it should be straightforward to obtain the tokens by using space as the delimiter. Feel free to do any other preprocessing that you might think is important for this corpus. **Do not forget to describe and explain your pre-processing choices in your report.**

3 Random sentence generation

Next, you will write code to generate random sentences based on your unigram and bigram language model(s). Develop a random sentence generator for each `training` corpus (i.e. for the training corpus associated with each speaker). It should produce sentences from scratch as described in class. Also, experiment with seeding — start from an incomplete sentence of your choice and use your language model to complete it (instead of generating from scratch). **Include examples of the generated sentences from each language model in your report. Analyze the generated sentences and compare the sentences generated with unigram vs. the bigram models.**

4 Smoothing and unknown words

For the speech classification task, you will need to implement smoothing and and a method to handle unknown words in the test data. Teams can choose any method(s) that they want for each. **The report should make clear what methods were selected**, providing a description for any non-standard approach, e.g., an approach that was not covered in class or in the readings.

5 Perplexity

Implement code to compute the perplexity of a “development set.” (“Development set” is just another way to refer to the validation set — part of a dataset that is distinct from the training portion and the test portion.) **Compute and report the perplexity of each of the language models (one trained on Obama’s speeches and one on Trump’s) on the `development corpora`.** Compute perplexity as follows:

$$
\begin{aligned}
PP &= \left(\prod_{i}^{N} \frac{1}{P(W_i|W_{i-1}, \ldots, W_{i-n+1})} \right)^{\frac{1}{N}} \\
&= \exp \frac{1}{N} \sum_{i}^{N} - \log P(W_i|W_{i-1}, \ldots, W_{i-n+1})
\end{aligned}
$$

where N is the total number of tokens in the test corpus and P($W_i|W_{i-1}, \ldots, W_{i-n+1}$) is the n-gram probability of your model. Under the second definition above, perplexity is a function of the average (per-word) log probability: use this to avoid numerical computation errors.

If you experimented with more than one type of smoothing or unknown word handling, it would be a good idea to report, compare and discuss the results of experiments with each.

> The work from Sections 2 until 5 constitute **PART ONE** of the project. See Section 10 for a summary of what to submit for this part of the project. Early submission of Part One is **optional**; it is for feedback purposes only and will not be given a grade. **No late submissions of Part One will be accepted.**

6 Speech classification

In this part of the project, you will use language models to automatically predict the speaker of paragraphs from an unseen portion of the corpus.

Hint. There are potentially many ways to use language models as predictors. The phrasing in the very first sentence in this section gives away one approach that is probabilistically meaningful. You may explore other ways as well; regardless, **make sure to clearly explain your method, and why you chose it, in the report**. We don’t recommend that you try fancy or advanced ideas without first implementing and evaluating the simple, straightforward one we intended. (Good life advice in general!) As in the Perplexity section, **if you experimented with more than one classification approach, it would be a good idea to report, compare and discuss the results of experiments with each.**

Methodology suggestion. You are given plenty of labeled speech excerpts of both Obama and Trump. To adjust your models for best performance (e.g. tuning smoothing parameters, or choosing between design options) you should use the provided development sets to evaluate the performance of your model. The folder structure indicates the `train` and `test` sets as well as the classification (speaker) labels (e.g. all the excerpts in `train/obama.txt` are Obama's speeches). You should use only the training data and development data for developing your model. **Submit your predictions for the excerpts in the test data on Kaggle.** See Section 6.1 below for the format. Keep in mind that the evaluation metric used on Kaggle is accuracy — the number of correct predictions divided by the number of excerpts in the test set.

6.1 Kaggle submission format

Apart from your report, you will submit a **.csv** file to Kaggle. There should be one prediction on each line for each speech inside the test folder. Each line should contain an Id (the line number of the speech in the test set) and the category predicted by your algorithm for the speech, separated by a comma. The prediction values are integers using the following encoding: obama: 0; trump: 1. So your output file should look like:

```
Id,Prediction
0,0
1,0
2,1
...
```

7 Evaluating word embeddings

In this part of the assignment, you will try to evaluate the quality of pre-trained word embeddings (Section 7.2). Word embeddings are often described as a distributed meaning representation for words. Most methods for deriving word embeddings/vectors are based on the general idea that words that occur in similar contexts are similar in meaning and should have similar vector representations. Word embeddings are trained on large collections of text without any specific NLP task in mind.

One method proposed to evaluate the quality of word embeddings is the **word analogy task** described in class (e.g., *man* is to *woman* as *king* is to *queen*). In particular, given a pair of words ⟨ a, b ⟩ and another word c, a word analogy system should find the word d such that c and d are related in the same way that a is related to b. For the purpose of this assignment, we need not identify the type of relationship.

Mikolov et al. [2] proposed vector operations on word embeddings for this analogy task. Let $\mathbf{v}_a$ be the vector representation for a, $\mathbf{v}_b$ for b. Then, in order to find the word d such that the analogy holds, we expect

$$\mathbf{v}_b - \mathbf{v}_a \approx \mathbf{v}_d - \mathbf{v}_c \tag{1}$$

$$d = \arg\max_{d \in V \setminus \{a,b,c\}} \cos(\mathbf{v}_d, \mathbf{v}_b - \mathbf{v}_a + \mathbf{v}_c) \tag{2}$$

Thus, you can compute cosine similarity with respect to all the words in the vocabulary (V) and then report the word with the highest cosine similarity.

7.1 Dataset

You have been given a subset of the data for Mikolov's analogy task in the file `analogy_test.txt`, which includes four types of semantic relations (family, adjective-to-adverb, comparative, plural) and another four types of syntactic relations (city-in-state, currency, capital-world, nationality-adjective). In the test file, each line is a analogy question, for example:

Athens Greece Oslo Norway

Thus, your cosine similarity metric will take the first three words as input and determine the fourth word. And your answer will be considered correct if the word you predict matches the fourth word in the analogy.

7.2 Pre-trained word embeddings

Pre-trained word embeddings are word embeddings that have already been trained on a large corpus. For this assignment, you can use any pretrained embeddings you can find. Remember to cite the paper and mention the choice of the word embeddings in your report.

Some of the pretrained embeddings that are popular are :

- Word2vec [2] `https://code.google.com/archive/p/word2vec/`
- Glove [3] `https://nlp.stanford.edu/projects/glove/`
- Dependency-based embeddings [1] `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`

7.3 Your task

- Pick any **two** sets of pre-trained word embeddings (see 7.2) and then compare their performance using the analogy test dataset provided with the assignment. **The report should include a table of the numeric results as well as a discussion of the results.**
- Design two new types of analogy tasks with three examples of each type. You will create your own test questions and report the performance of the choice of word embeddings on those tasks. **The report should describe the analogy tasks, include the three examples of each, and include a table and discussion of the results.**

- One of the known problems with word embeddings is the way they handle antonyms — both the word and its antonyms tend to have similar word embeddings. Verify this by finding the top 10 most similar words (using the cosine metric) for a few verbs such as *enter* or *increase*. **Report on and explain your findings.**

8 Speech classification with word embeddings

This part of the assignment asks you to use pre-trained word embeddings (of your choice [1]) to perform speech classification of the same Obama and Trump speech excerpts used earlier in this project. Given a paragraph, you want to determine its speaker (Obama or Trump). For this purpose, you will use the word embeddings to compute a vector representation for the speech and compare it to a similarly computed vector representation for the entirety of Obama's (training set) speech paragraphs vs. Trump's (training set) speech paragraphs. **If the excerpt representation is closer (via the cosine similarity metric) to Obama's vs. Trump's speeches, then return 'Obama' as the predicted speaker; return 'Trump' otherwise.**

To compute the speech excerpt and speech representations, you will experiment with a very simple model that **simply averages the word embeddings of all the words in a given speech or set of speeches**.

More formally, if the speech (R) is

oh boy. we love nevada.

then you use the pre-trained word embeddings as vector representations of its words: $\mathbf{v}_{oh}$, $\mathbf{v}_{boy}$, $\mathbf{v}_{.}$, $\mathbf{v}_{we}$, $\mathbf{v}_{love}$, $\mathbf{v}_{nevada}$. Next, we compute the representation (x_R) of the speech R, by averaging the word vector representations:

$$x_R = (\mathbf{v}_{oh} + \mathbf{v}_{boy} + \mathbf{v}_{.} + \mathbf{v}_{we} + \mathbf{v}_{love} + \mathbf{v}_{nevada})/6. \tag{3}$$

To represent the **set** of Obama's (or Trump's) speeches, you can use the above approach to obtain a vector representation for each speech, and then take the average of the speech-level vectors.

For submission, **run your model on the files in the test folder and submit your predictions on Kaggle**.

8.1 A machine-learning variation

THIS OPTION ONLY MAKES SENSE FOR TEAMS WITH PRIOR MACHINE LEARNING EXPERIENCE. As an alternative to the above, you can use the word embedding representation of the speeches (from the training as well as the test set) as **features** for any classifier of your choice.

If you were to use a linear classifier (you can use any existing implementation) then you can learn the weights (w) associated with each feature such that the classifier

[1] You should consider using ELMo [4] `https://allennlp.org/elmo` as one of the alternatives for obtaining pretrained word embeddings as they are known to model context better.

learns to predict the correct label (y).

$$y = \mathbf{w}.\mathbf{x} \tag{4}$$

Like the language-model-based classifier from Section 6, train your model on the examples in the train folder and tune your model using the development set.

For submission, **run your model on the files in the test folder and submit your predictions on Kaggle**.

9 Grading rubric

Part One:

- Unigram and bigram probability table; random sentence generator (10%)
- Smoothing, unknown word handling, and perplexity calculation (10%)
- Implementation of perplexity (10%)

Part Two:

- Speech classification with language models (10%)
- Evaluating word embeddings (15%)
- Speech classification with word embeddings (10%)
- Experiment design and methodology; report quality (30%)
- Kaggle submission (5%)

Note: We encourage you to submit Part One by the first deadline to obtain feedback before the final deadline thus giving you a chance to improve your models and report. However, it is optional. You are required to submit everything (both Part One and Part Two) in the assignment by the final deadline.

Performance concerns: While you shouldn't focus on optimization, nothing in this assignment should need to be slow. If you find your implementations taking really long amounts of time, something is probably wrong with your design. You may lose points for impractically slow, brute-force style solutions, where applicable.

10 What to submit

You must create your Group on both CMS and Gradescope. Note that, at least as of last year, resubmissions to Gradescope REMOVED the Group structure. So be sure to re-create your Group with every resubmission to Gradescope.

For Part One (via CMS and Gradescope): Submit a zip file with your code and the portion of the report associated with Part One (the latter as a pdf file).

For Part Two (via CMS and Gradescope): Check the submission format for each of the tasks as explained above. Submit your (1) report (a .pdf) and (2) code on CMS as a .zip or .tar file. Also submit (3) the csv files produced for Section 6 and Section 8 on Kaggle. We will have separate Kaggle challenges for the two submissions. We will notify about the submission link later via Piazza.

In order to be graded as a group, **you must form groups with your partner(s) on Kaggle.** You will have a limited number of Kaggle submissions per group. You can only submit up to 5 predictions daily.

The links for the Kaggle competition are:

- **For Section 6**
  `https://www.kaggle.com/c/cs4740-speech-lm`
- **For Section 8**
  `https://www.kaggle.com/c/cs4740-speech-emb`

11 The report

With the Part Two submission, you should submit a short document (no more than 6 pages) that describes your work. Your report should contain **a section for each of the Sections 2-8 above** as well as **a short *workflow* section** that explains how you divided up the work for the project among group members. **Important:** the role of this document is to show us that you **understand** and **are able to communicate** what you did, and how and why you did it. **Come see us if you're not sure how to answer either of these questions!**

You might think this means you must write a lot to explain something, it may be a bad sign. Describe the general approach, the data structures used (where relevant). Use examples, identify all design choices and justify them (e.g., how did you deal with unknown words? how did you perform smoothing?). Wherever possible, **quantify your performance** with evaluation metrics (e.g., report classification performance on the development set). If you tried any extensions, perform measurable experiments to show whether or not your extensions had the desired effect? If something doesn't work or performs surprisingly, try to look deeper and explain why.

Code. You may include snippets of your code in the report, if you think it makes things clearer. Include **only** relevant portions of the code (such as a few lines from a function that accomplish the task that you are describing a nearby paragraph, to help us understand your implementation). Including boilerplate code or tedious, unnecessary blocks (e.g., reading files) only makes your report cluttered and hard to follow, so avoid it.

References

[1] Omer Levy and Yoav Goldberg. Dependencybased word embeddings. In *In ACL*, 2014.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[3] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[4] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.