COMP-551 Final Project

T1-10: Distributed Representations of Sentences and Documents

Team: Stochastic Gradient Descent into Madness

Daniel Cimento
daniel.cimento@mail.mcgill.ca
260679318

Adam Edery
adam.edery@mail.mcgill.ca
260691043

Howie Zhao
haoyi.zhao@mail.mcgill.ca
260631984

I. Objective

For this project, we were tasked with recreating the baselines reported in the paper "Distributed Representations of Sentences and Documents" (Le & Mikolov, 2016). Our objective was two-fold: first we needed to achieve the same basline performances reported, then if possible, to tune those baselines further, to achieve more competitive performance. Though we anticipated that we would be able to beat the reported baselines, we did not expect to outperform the main model of the paper, the Paragraph Vector model.

II. Methodology

Our methodology for this project was divided into several distinct phases. First, we applied some pre-requisite processing to the data set. Then, we implemented the simpler linear models that were reported in the paper. After this, we applied further tuning to these models, to optimize the performance wherever possible. We then explored alternative linear models, to see if any could outperform those chosen by the paper's authors. Finally, we explored the performance of the other, more complex models like word vector averaging and neural networks, though these were not intended to be the focus of our efforts.

A. Data Pre-processing

The pre-processing and feature selection composed a non-negligible part of our methodology, so we will cover the steps taken to process the raw data sets.

1) Manual Data Cleaning: In the first place, the sentence files provided in the Stanford sentiment analysis database were encoded in Latin-1. Since our code would benefit from having a consistent encoding standard throughout the development process, we needed to handle these files with the UTF-8 encoding standard. This resulted in parts of the file being inconsistently encoded, with mojibake appearing occasionally throughout our dataset, which proved to be troublesome for Python's default IO implementation. Consequently, the first step of our data pre-processing involved cleaning the sentence file by replacing all Latin-1 encoded characters with their corresponding unicode equivalent.

Furthermore, the Stanford sentiment analysis database is formatted with the intention of being used for syntax parsing models, as the sentences are structured in a tree-like format, and their corresponding sentiments are stored across several files. For our application, we wanted to handle the full text of each sentence, so we wrote the code necessary to convert the tree-like structure into simple "sentence-sentiment" pairs.

2) Vectorization: Once we had our data loaded in the proper format, we needed to extract the features from each sentence to create vectors for our models. The simplest representation would be bag-of-words, so we started with that, using sklearn's built in CountVectorizer.

However, we felt that just using bag-of-words wouldn't accurately capture the semantic similarity between inflected words (e.g. "film" and "films"), so we replaced CountVectorizer's built-in tokenizer with a tokenizer that tagged the parts of speech for each word, then retrieved that word's lemma from the WordNet database. To do so, we took advantage of several features built into the nltk package. We also used a CountVectorizer that included both unigrams and bigrams (for our Bigram Naive Bayes implementation), which used the same tokenization scheme.

Similar to the paper in question, we labelled each sentence in the Stanford sentiment analysis dataset based on its sentiment score (using the labels Very Negative, Negative, Neutral, Positive, and Very Positive). We likewise created a set with coarser label classifications (with labels Negative and Positive), ignoring all neutral data points for our coarse-grain performance evaluations.

The IMDB dataset was vectorized in a similar way, but since it used a binary label approach, we didn't need to make any modifications to the labels.

B. Simple Linear Models

For the simple linear model methodology, there is little to discuss that hasn't already been covered by the paper in question, but we will summarize the main points of each.

1) Naive Bayes:

2) Linear SVM:

3) Bigram Naive Bayes:

C. Improving Linear Models

1) Naive Bayes:

2) Linear SVM:

D. Alternative Simple Models

1) *Random Forests:*

2) *Logistic Regression:*

3) *K-Nearest Neighbors:*

E. Complex Models

1) *Word Vector Averaging:*

2) *Neural Networks:*

III. Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi nec orci sed turpis placerat dapibus ac non justo. Phasellus volutpat sem felis. Nam id tincidunt massa. Quisque diam ante, mattis et molestie porttitor, rhoncus a dolor. Aliquam sed diam porta turpis dapibus ultricies. Nam sit amet vulputate felis. Vestibulum commodo placerat neque, et feugiat odio. Mauris tristique malesuada leo, nec finibus orci varius non. Sed quis mattis velit. Morbi rhoncus ante vel sodales ullamcorper. Duis feugiat accumsan metus id pellentesque. Nulla pellentesque, leo eget vehicula malesuada, turpis diam egestas neque, non vulputate purus nibh lobortis odio. Nunc maximus aliquam nisl vitae molestie. In hac habitasse platea dictumst. Fusce ut luctus diam. Aliquam aliquam laoreet felis, sed dictum nulla facilisis et.

In interdum suscipit mattis. Sed porta imperdiet purus, at imperdiet leo condimentum eget. Pellentesque at velit mauris. Phasellus venenatis arcu eget lorem efficitur, ut pulvinar purus tincidunt. Aliquam vel nibh tempor, blandit lacus vel, ullamcorper sem. Nam porttitor, ipsum tempus varius elementum, quam neque ultricies nisi, sed commodo tortor eros in nibh. Proin aliquet placerat erat in varius. Proin semper odio ut eros condimentum vestibulum. Mauris id porttitor ipsum. Sed tincidunt enim enim, sed dapibus urna aliquam sed. Sed a ex elementum, hendrerit eros ac, euismod dui. Etiam ligula ligula, scelerisque vitae lectus et, ultrices semper erat.

Vivamus id dictum nulla. Praesent consectetur pellentesque dui quis scelerisque. Nunc sed augue vehicula, scelerisque felis ut, viverra nulla. Curabitur aliquet ipsum non tincidunt congue. In pretium nec nisl et euismod. Sed vitae euismod lorem, vitae auctor est. Mauris varius orci ac augue semper, sed sollicitudin mi porttitor.

Nullam efficitur ante in bibendum rutrum. Suspendisse congue pretium metus vitae scelerisque. Maecenas vel justo finibus, bibendum nibh non, egestas metus. Vivamus sed mollis purus. Mauris interdum a ligula vitae tristique. Nam vel varius metus, a tincidunt elit. Curabitur convallis augue lacus, eget dapibus risus pretium non. Proin ut consectetur quam, ut lobortis nibh. Cras dui elit, suscipit euismod laoreet et, interdum id nulla. Cras varius augue eget enim semper, sed tempus enim sagittis.

Phasellus mattis maximus dui, id vestibulum mi finibus ut. Maecenas commodo fringilla magna quis dignissim. Praesent eget felis vulputate, eleifend dolor eu, finibus elit. Aliquam rhoncus mi ut tellus laoreet sodales. Sed sed luctus purus. Praesent hendrerit tortor sit amet nibh dignissim, sed gravida magna dictum. Nam mollis fermentum dui placerat auctor. Pellentesque vel quam tellus. Mauris gravida, leo eu commodo euismod, velit ante rutrum sem, sed sollicitudin mi tellus vel diam. In mattis malesuada sagittis. Vivamus vel fringilla tortor. Nulla ac sem in turpis rutrum sodales. Nullam accumsan dolor ac auctor pellentesque. Cras suscipit, eros at lobortis commodo, erat diam sollicitudin augue, interdum ullamcorper turpis dolor sit amet magna.

IV. Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi nec orci sed turpis placerat dapibus ac non justo. Phasellus volutpat sem felis. Nam id tincidunt massa. Quisque diam ante, mattis et molestie porttitor, rhoncus a dolor. Aliquam sed diam porta turpis dapibus ultricies. Nam sit amet vulputate felis. Vestibulum commodo placerat neque, et feugiat odio. Mauris tristique malesuada leo, nec finibus orci varius non. Sed quis mattis velit. Morbi rhoncus ante vel sodales ullamcorper. Duis feugiat accumsan metus id pellentesque. Nulla pellentesque, leo eget vehicula malesuada, turpis diam egestas neque, non vulputate purus nibh lobortis odio. Nunc maximus aliquam nisl vitae molestie. In hac habitasse platea dictumst. Fusce ut luctus diam. Aliquam aliquam laoreet felis, sed dictum nulla facilisis et.

In interdum suscipit mattis. Sed porta imperdiet purus, at imperdiet leo condimentum eget. Pellentesque at velit mauris. Phasellus venenatis arcu eget lorem efficitur, ut pulvinar purus tincidunt. Aliquam vel nibh tempor, blandit lacus vel, ullamcorper sem. Nam porttitor, ipsum tempus varius elementum, quam neque ultricies nisi, sed commodo tortor eros in nibh. Proin aliquet placerat erat in varius. Proin semper odio ut eros condimentum vestibulum. Mauris id porttitor ipsum. Sed tincidunt enim enim, sed dapibus urna aliquam sed. Sed a ex elementum, hendrerit eros ac, euismod dui. Etiam ligula ligula, scelerisque vitae lectus et, ultrices semper erat.

Vivamus id dictum nulla. Praesent consectetur pellentesque dui quis scelerisque. Nunc sed augue vehicula, scelerisque felis ut, viverra nulla. Curabitur aliquet ipsum non tincidunt congue. In pretium nec nisl et euismod. Sed vitae euismod lorem, vitae auctor est. Mauris varius orci ac augue semper, sed sollicitudin mi porttitor.

Nullam efficitur ante in bibendum rutrum. Suspendisse congue pretium metus vitae scelerisque. Maecenas vel justo finibus, bibendum nibh non, egestas metus. Vivamus sed mollis purus. Mauris interdum a ligula vitae tristique. Nam vel varius metus, a tincidunt elit. Curabitur convallis augue lacus, eget dapibus risus pretium non. Proin ut consectetur quam, ut lobortis nibh. Cras dui elit, suscipit euismod laoreet et, interdum id nulla. Cras varius augue eget enim semper, sed tempus enim sagittis.

Phasellus mattis maximus dui, id vestibulum mi finibus ut. Maecenas commodo fringilla magna quis dignissim. Praesent eget felis vulputate, eleifend dolor eu, finibus elit. Aliquam rhoncus mi ut tellus laoreet sodales. Sed sed luctus purus. Praesent hendrerit tortor sit amet nibh dignissim, sed gravida magna dictum. Nam mollis fermentum dui placerat auctor. Pellentesque vel quam tellus. Mauris gravida, leo eu commodo euismod, velit ante rutrum sem, sed sollicitudin mi tellus vel diam. In mattis malesuada sagittis. Vivamus vel fringilla tortor. Nulla ac sem in turpis rutrum sodales. Nullam accumsan dolor ac auctor pellentesque. Cras suscipit, eros at lobortis commodo, erat diam sollicitudin augue, interdum ullamcorper turpis dolor sit amet magna.

V. Conclusion

VI. Statement of Contributions

VII. References