Word Vectorization

REPORT

Ashmit Chamoli

Introduction

Word vectorization refers to the task of generating meaningful numerical representations for words. These are also known as distributional semantic algorithms. In this repository, we explore 2 word vectorization techniques:

- SVD
- Word2Vec (skip-gram with negative sampling)

Dataset

We use the News Classification Dataset which contains labeled news snippets. First, we obtain the word embeddings using both SVD and Word2Vec techniques. Then we use these embeddings to train an LSTM classifier to classify the news snippets.

Experiments

We train 6 embedding models, 3 each for SVD and Word2Vec, varying context size from 2 to 4 with an embedding size of 300 and k=3 for Word2Vec.
We use the same classifier for all the 6 set of embeddings with the following hyperparameters:

Hidden Size	Num Layers	Bidirectional	Hidden Layers	Activation
256	3	True	[128, 64]	tanh

Analysis

We observe that Word2Vec embeddings give better performance than SVD embeddings.

LM 1: Tokenization + 3-gram LM + Good-Turing Smoothing

Pride and Prejudice

Train	90651.12
Test	12722.59

The high perplexity values on the test set indicate that the model is quite confused about it's prediction. The higher value of perplexity in the train set is because Good Turing smoothing assigns a very high probability to unseen n-grams which in turns results in probability of seen but low frequency n-grams to be extremely low. In the test set however, we see a lot of unseen n-grams for which the good turing model returns a very high probability.

Ulyess

Train	215313.48
Test	16327.17

Here similar trend is followed, except that the train perplexity is much higher than we see in Pride and Prejudice. This is because of a much larger vocabulary set and a larger dataset. This results in the probability of unseen n-grams to be even higher than was the case in Pride and Prejudice dataset.

LM 2: Tokenization + 3-gram LM + Linear Interpolation

Pride and Prejudice

Train	27.70
Test	813.61

Ulyess

Train	97.05
Test	2463.43

The perplexity in the for the train sets is quite low, indicating that the model is quite sure of it's prediction. On the test sets however, the overall perplexity is much lower than we see in Good Turing, meaning that the model is performing better in this case.

The perplexity scores are always higher for the Ulyess dataset as compared to the Pride and Prejudice dataset. This might be because of the richer vocabulary in the Ulyess dataset.

The best performance we achieve is by LM2 on the Pride and Prejudice dataset.