CIS 545 - Final Project

Peter Kong

December 9, 2018

1 Introduction

1.1 Problem Domain

1.2 Problem Description

2 Data Acquisition & Preparation

3 Exploration

4 Feature Engineering

5 Implementation

6 Analysis

6.1 Error Analysis

We tried one variation of a sparse vector representation, which expanded the model to the top 1000 (instead of top 500) words.

We experimented primarily with multiple clustering algorithms, trying the `AgglomerativeClustering`, `Birch`, and `SpectralClustering` algorithms.

We compare the impact of our model and clustering choices by evaluating their performance over the dev set (Table **??**):

Vector Space Model	KMeans	Spectral	Agglomerative	Birch
500vec	0.3662	0.2876	0.3748	0.3649
1000vec	0.3661	0.2739	0.3727	0.3652

Table 1: Paired F-Score on the dev set by different vector space models and clustering algorithms.

7 Dense Vector Representations

As with the sparse vectors, we experimented with multiple clustering algorithms. We used an SVD transformation with 200 components. Varying the number of components in either direction did not produce meaningfully different results.

We compare the impact of our model and clustering choices by evaluating their performance over the dev set (Table **??**):

Dense Model	KMeans	Spectral	Agglomerative	Birch
SVD	0.3298	0.3165	0.3569	0.3076

Table 2: Paired F-Score on the dev set by different dense vector space models and clustering algorithms.

8 Comparison

Overall, our sparse model performs slightly better. One explanation for this is that, for this data set, the singular value decomposition generalizes poorly by discounting the lower order dimensions. In other words, there is significant signal in the lower order dimensions of this data set that is lost through SVD.

In general, we can compare dense and sparse models by looking at instances where one does well and the other fails.

Examples of target words where the dense model scores high and the sparse model scores low are... (describe what they have in common, hypothesize why this might be the case)

simple.a 0.2051 sparse
simple.a 0.3030 dense
miss.v 0.3500 sparse
miss.v 0.4158 dense

Different parts of speech, but fairly common words. Since fewer cells for these words will be zero, their influence on the dense vectors may be proportionally greater.

Examples of target words where the sparse model scores high and the dense model scores low are... (describe what they have in common, hypothesize why this might be the case)

expect.v 0.4924 sparse
expect.v 0.3679 dense
begin.v 0.3273 sparse
begin.v 0.2831 dense

Both verbs; both quite common. Difficult to hypothesize why. We separated precision and recall out of the fscore, but there was lots of variance in both metrics and no clear pattern.