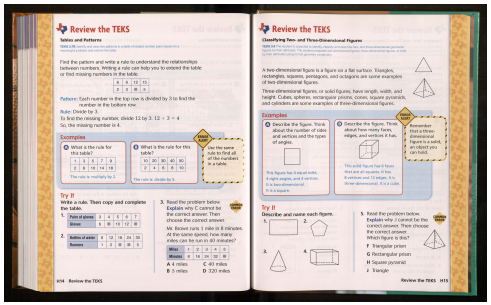
Developing Advanced Techniques for Textual Analysis: Word Embeddings as Messaging Indicators

Anjali Adukia[1], Alex Eble[2], Noah McLean[1], Jeffrey Tharsen[1]
[1] University of Chicago [2] Columbia University

Background

The goal of our project is to use new methods in text analysis to understand the level of and variation in hidden messages about gender and race contained in text. Our focus will be on school materials, e.g., textbooks. Below is an example of our "source material":



Word Embedding Models

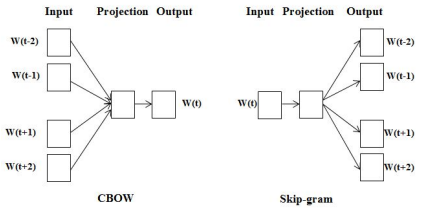


Fig: CBOW and Skip-gram neural network architecture

After testing the two most-used word embedding neural networks, we decided to implement Skip-gram in our pipeline. Although CBOW is often faster, the benefits of Skip-gram: better representation of rare words and better performance on small datasets, far outweigh the benefits of CBOW models. Using pre-trained embeddings as the backbone of our system further decreased our need for CBOW's speed.

Pipeline

Through our pipeline (below), we have separated the text analysis aspect of our project into 7 distinct phases for each book. Our work over the past several months has focused on improving the transitions between each phase. Once this process is complete for each book in the pilot (1000+), we can begin to search for patterns or relationships in the messaging of each book.

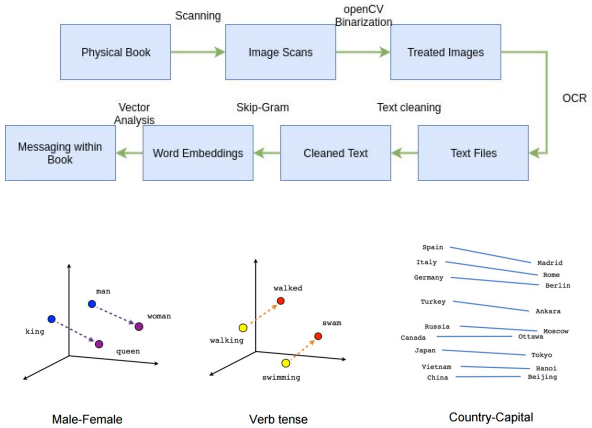


Fig: 3 examples of word embedding vector analysis usage
Source: Ruizendaal, 2017

Results

Preliminary results depict some evidence of implicit gender messaging, although future findings may vary largely across book types, books themselves, or authors.

```
Male-neutral top word similarities
        me, 1 occurences
        them, 1 occurences
        it, 2 occurences
        itself, 1 occurences
        themselves, 1 occurences
        my, 1 occurences
        myself, 1 occurences
------------------------------------------------------------
Female-neutral top word similarities
        myself, 1 occurences
        my, 1 occurences
        it, 2 occurences
```

Fig: most similar non-gendered specification words to male and female words compared

Future work

Combined embedding analysis

We currently store 65 scanned books on Midway, which translates to roughly 1.73 terabytes of data. These statistics do not include scanned books currently backed up on encrypted drives or physical copies of books we plan to scan but have not yet received. We will:

- Establish accuracy through analogy comparison
- Conduct cosine similarity analysis on our vector data
- Identify possible messaging through word relations

Text isolation

The construction of accurate word embeddings requires full access to the text data. However, small errors in the OCR stage cause words to be misidentified, which ripples outwards into our results. One of our attempts to improve the OCR's performance will include text detection and isolation withing the scanned images using openCV.

Challenges: Identifying Racial Messaging

So far, our preliminary results are exclusive to implicit or explicit gender messaging. Our forays into racial messaging analysis have been especially difficult.

Garg et al. (2018) approached identifying implicit racial messages through word embeddings (one of the most thorough investigations of this topic) has some serious flaws we must address, including:

- Unable to differentiate black and white subjects
- Only able to analyze subjects with surnames
- Requires an extremely accurate and large dataset

Acknowledgements

Research Computing Center, University of Chicago.

Garg, N., L. Schiebinger, D. Jurafsky, & J. Zou. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS.
Suleiman, D., A. Awajan, & N. Al-Madi. (2017). Deep Learning Based Technique for Plagiarism Detection in Arabic Texts.
Ruizendaal, R. (2017). Deep Learning #4: Why You Need to Start Using Embedded Layers.