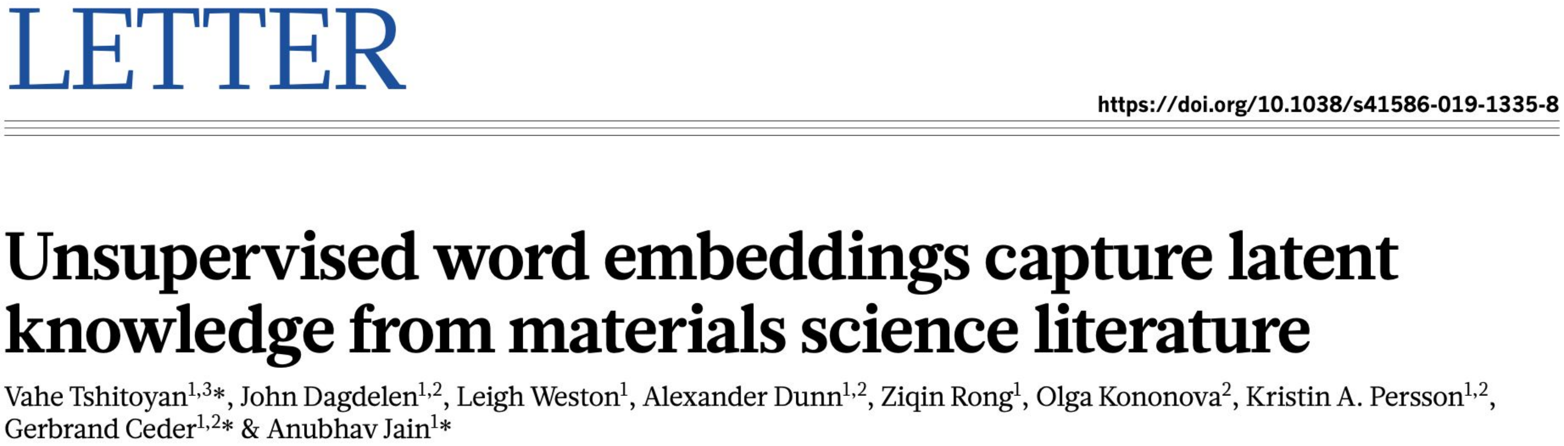
Final Project: Word Embeddings and NLP in Materials Science

Aden Weiser

How can we use knowledge from natural language processing to supplement materials discovery?



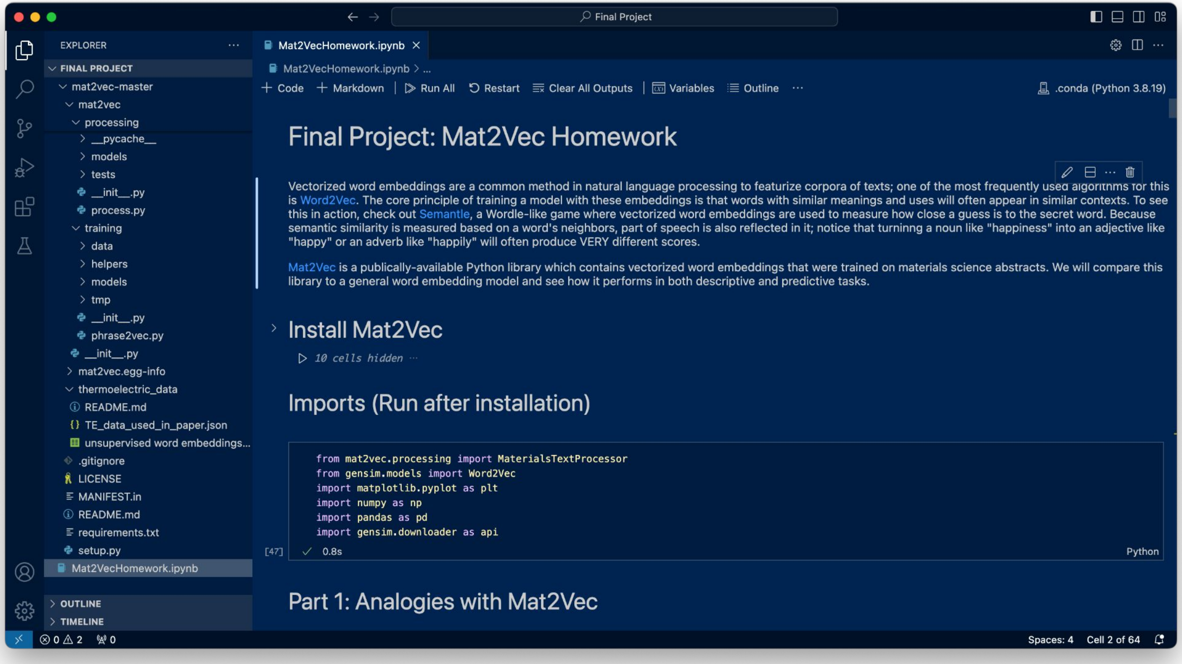
LETTER

https://doi.org/10.1038/s41586-019-1335-8

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan[1,3]*, John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2]* & Anubhav Jain[1]*

Deliverable: Homework-style Jupyter Notebook



Final Project: Mat2Vec Homework

Vectorized word embeddings are a common method in natural language processing to featurize corpora of texts; one of the most frequently used algorithms for this is Word2Vec. The core principle of training a model with these embeddings is that words with similar meanings and uses will often appear in similar contexts. To see this in action, check out Semantle, a Wordle-like game where vectorized word embeddings are used to measure how close a guess is to the secret word. Because semantic similarity is measured based on a word's neighbors, part of speech is also reflected in it; notice that turninng a noun like "happiness" into an adjective like "happy" or an adverb like "happily" will often produce VERY different scores.

Mat2Vec is a publically-available Python library which contains vectorized word embeddings that were trained on materials science abstracts. We will compare this library to a general word embedding model and see how it performs in both descriptive and predictive tasks.

Install Mat2Vec

Imports (Run after installation)

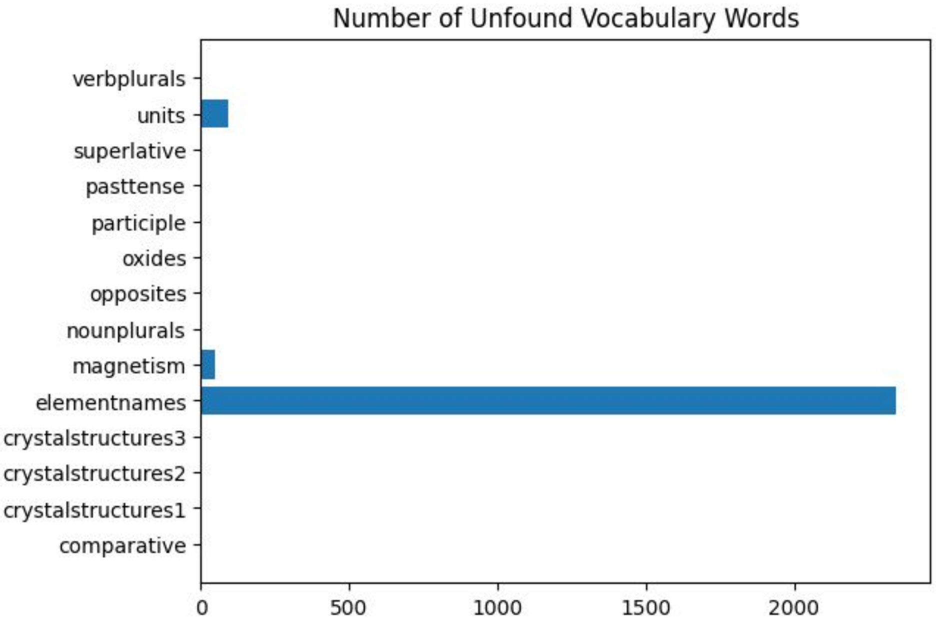
```python
from mat2vec.processing import MaterialsTextProcessor
from gensim.models import Word2Vec
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import gensim.downloader as api
```

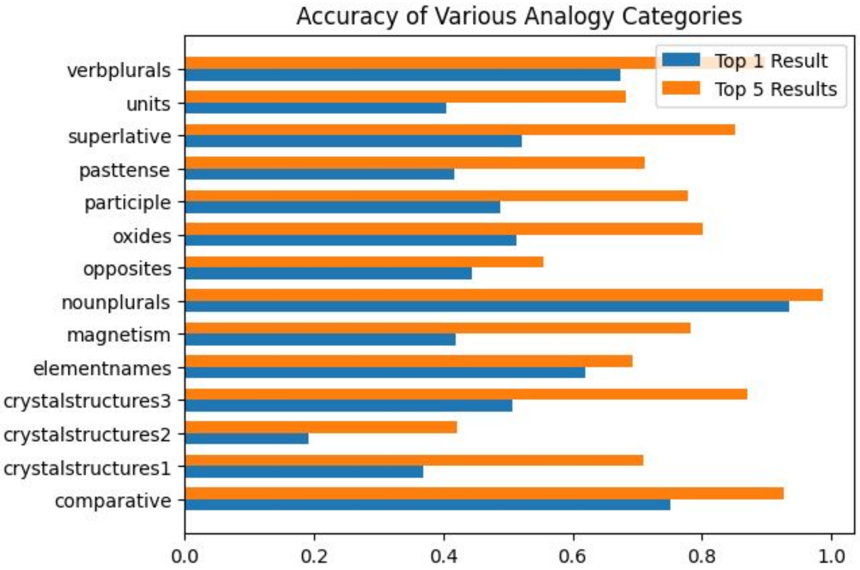
Part 1: Analogies with Mat2Vec

Key Learning Outcomes

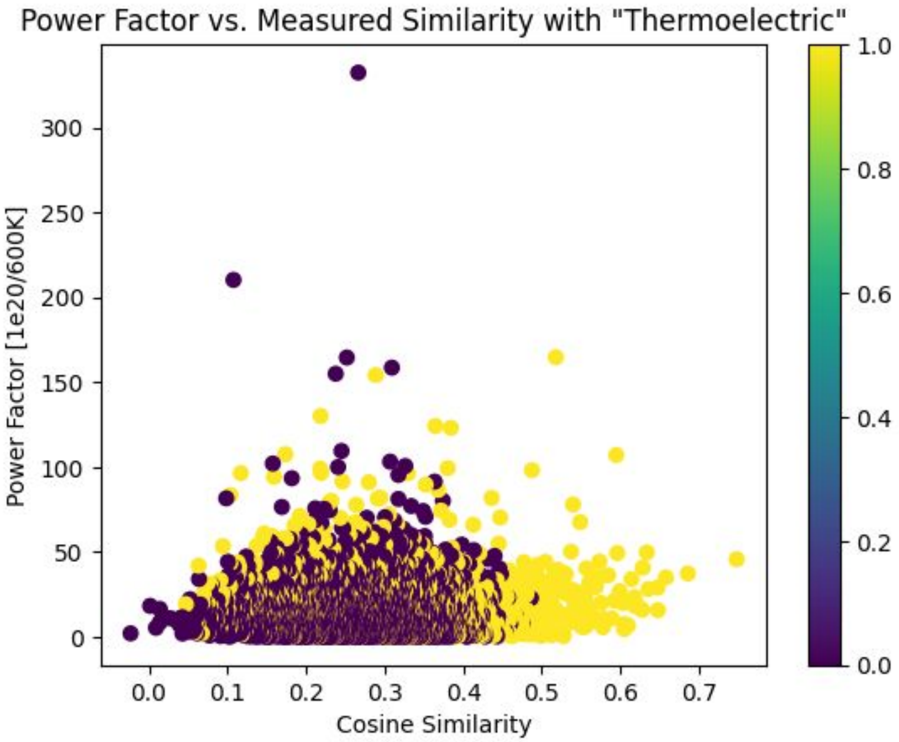
- Understand basic terminology about natural language processing (corpora, tokenization, embeddings)
- Learn how training models with different data sets results in different performance on the same task
- Explore the extraction of “latent knowledge” in scientific abstracts

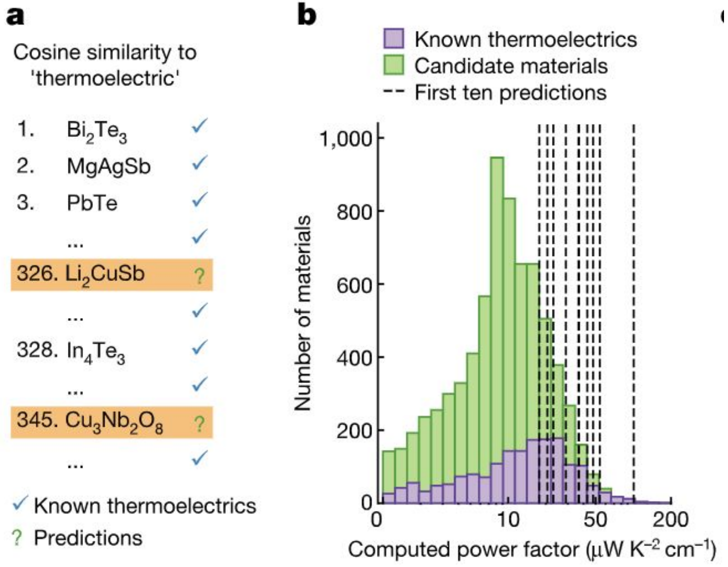
Task 1: Analogies with Mat2Vec only





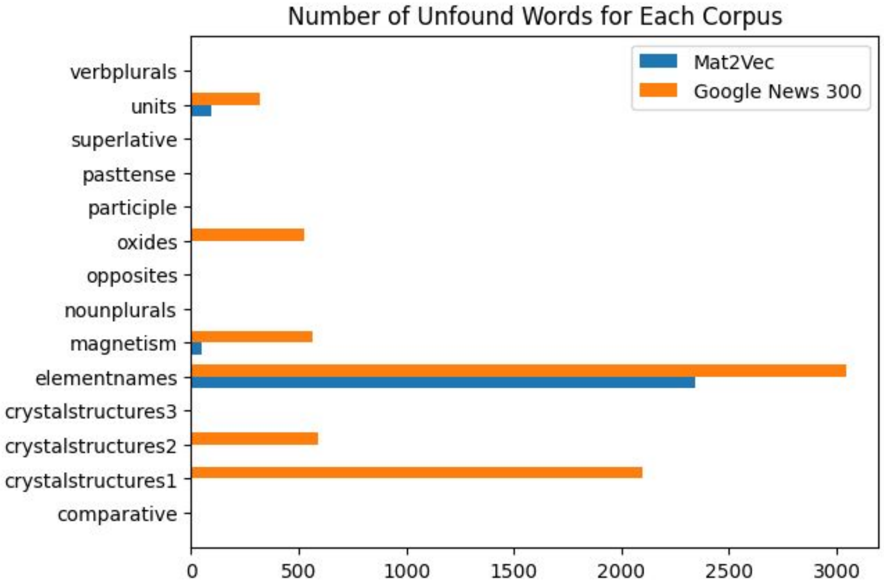
Task 2: Materials Prediction with Mat2Vec

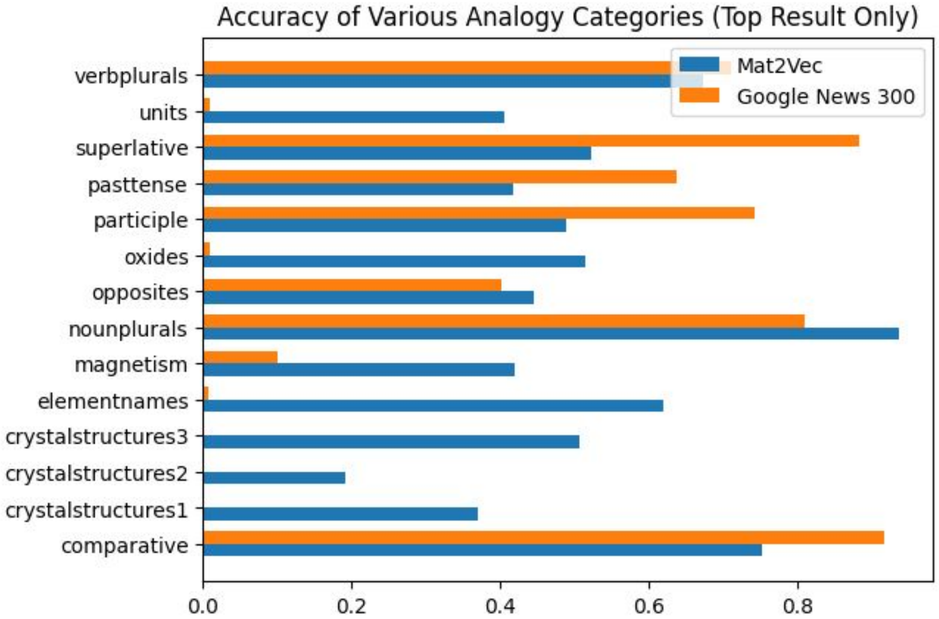




Tshitoyan et al.

Task 3: Comparison with Google News 300 Embeddings





Conclusions

- Training a neural network with domain knowledge results in high performance within a domain, training a model with general knowledge results in better grammatical intuition
- Linguistic similarity is another way to featurize materials data to extract meaningful information
- Annotate your data!!!

Future Work

- Expanding/refining materials prediction task
- Integration of multiple datasets
- Google Colab integration