Vector representation with a finance corpus

Student: Ritesh Bachhar, University of Rhode Island
riteshbachhar@uri.edu

Mentor: Murat Aydogdu, Rhode Island College
maydogdu@ric.edu

Date: August 10, 2022

CAREERS CYBERTEAM

Vector representation with a finance corpus

- Timeframe
  - May 15th, 2022
  - August, 2022

Vector representation with a finance corpus

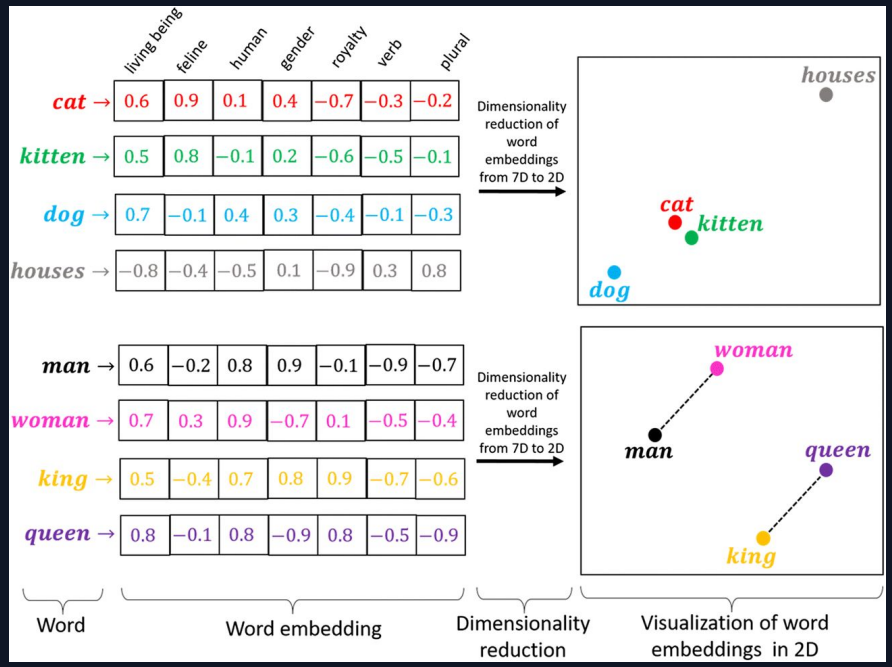
- Goals
  - Vector representation & NLP
  - Wikipedia Articles [1]
  - 10-K documents of public traded companies[2]
  - GloVe[3] implementation of word vector
  - Setup workflow in HPC

1 https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
2 https://www.sec.gov/Archives/edgar/
3 https://nlp.stanford.edu/projects/glove/

CAREERS CYBERTEAM

GloVe: Global Vector for word representation

- Word embedding
- Co-occurrence matrix
- Matrix Factorization Method (SVD ect.)
- Prediction based approach (Skip-Gram, CBOW)
- GloVe combines



https://medium.com/@hari4om/word-embedding-d816f643140

GloVe: Global Vector for word representation

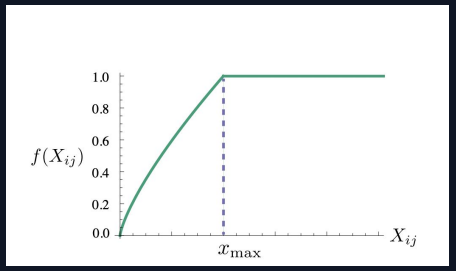
- X_{ij}: word-word co-occurrence matrix
- Global statistics to predict the probability of word j appearing in the context of word i

Model:

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}) .$$

Cost function:

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2 ,$$



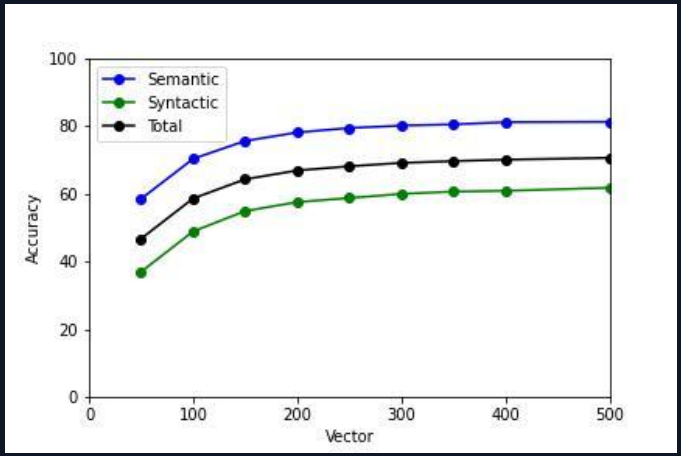
CAREERS CYBERTEAM

GloVe on Wikipedia Articles

- ~ 5.8 mil article in english wikipedia
- Step 1: Download Wiki dump file (XML)
- Step 2: Gensim package to convert xml to json; much easier to parse
- Step 3: Batch of 10,000 article in a single json file
- Step 4: Process each batch using Spacy; json to txt file
- Step 5: Concatenate all article in single file

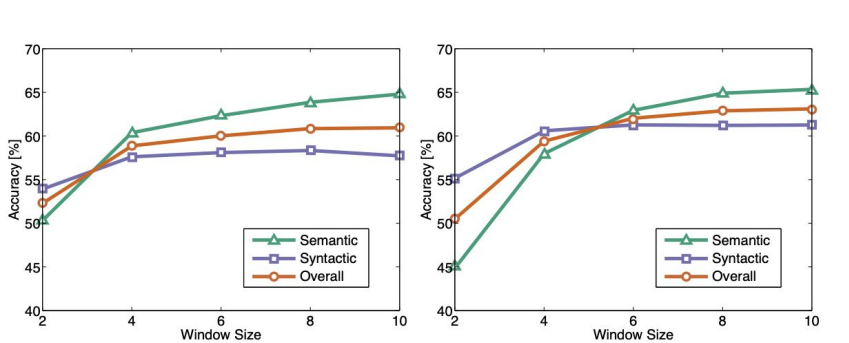
GloVe on Wikipedia Articles

- Result: Word analogy tasks (Semantic and syntactic)
- “a” is to “b” as “c” is to ___?
- king - man + woman = queen
- Dancing - dance + play = playing



CAREERS CYBERTEAM

GloVe on Wikipedia Articles



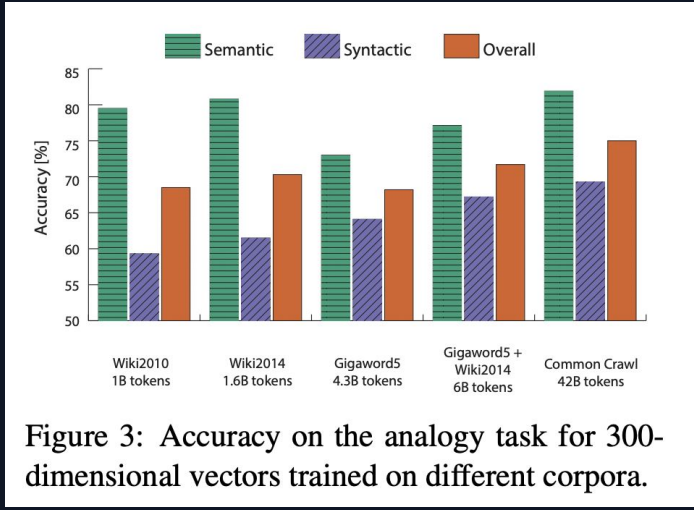
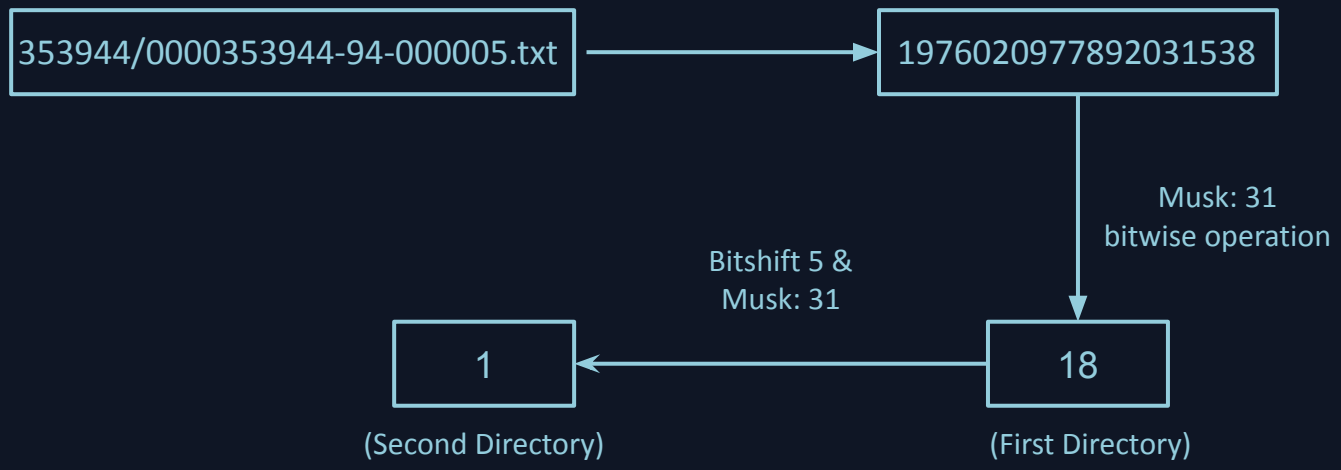


Figure 3: Accuracy on the analogy task for 300-dimensional vectors trained on different corpora.

CAREERS CYBERTEAM

10K files

- Total files: ~200k
- Problem: Listing files, existence of file
- Hashed directory
- Two level directory structure



CAREERS CYBERTEAM

Future work

- Processing of finance of document
- Identification of finance specific words
- Comparison between domain specific vectors

CAREERS CYBERTEAM

Thank You!