Interpreting Word Embeddings with Eigenvector Analysis

Jamin Shin Andrea Madotto Pascale Fung

Team number 33:
Elizaveta Saygina
Ilya Beniya
Ivan Naidenov
Ksenia Makarova

Plan

- Introduction
- Methodology
- Distribution of Eigenvector Elements
- Inverse Participation Ratio
- Column Space Analysis
- Conclusion

Introduction

- ‘What is the meaning of high and low values in the columns of W?’
- ‘How can we interpret the dimensions of word vectors?’

Methodology

- Positive Pointwise Mutual Information Matrix:

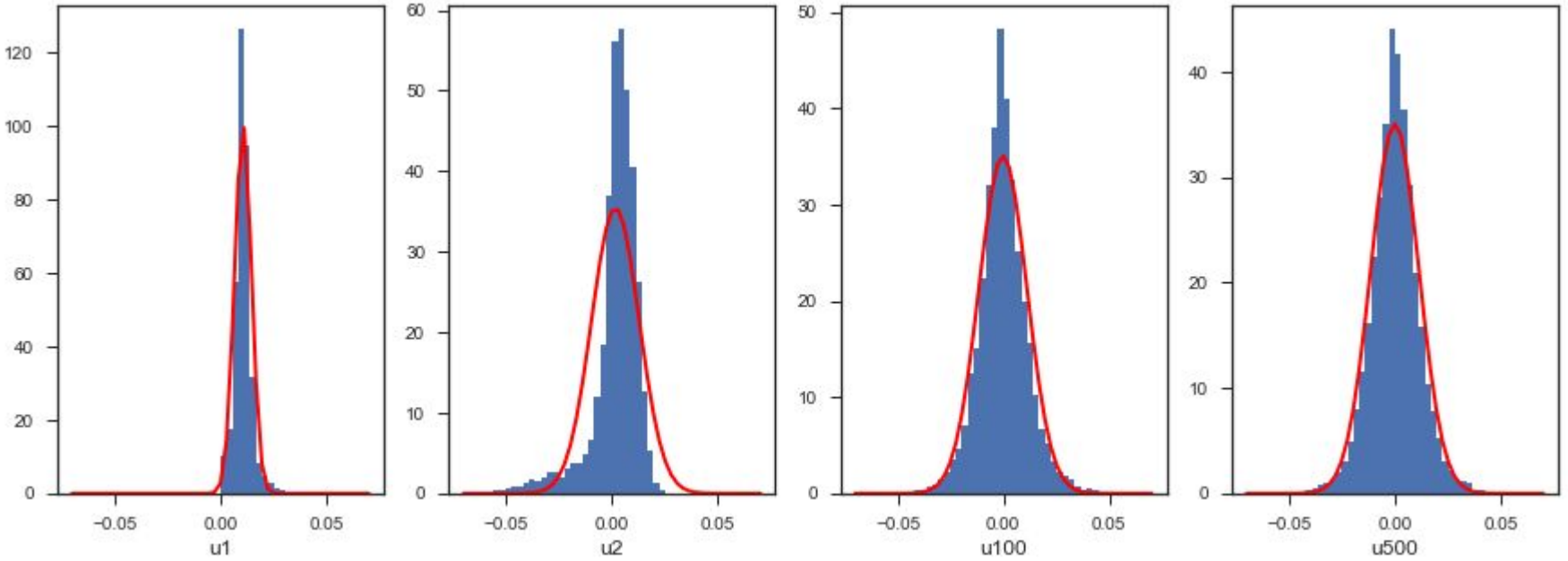
$$PMI(w,c) = \log \frac{\hat{P}(w,c)}{\hat{P}(w)\hat{P}(c)} = \log \frac{\#(w,c)|D|}{\#(w)\dot{\#}(c)} \qquad PPMI(w,c) = \max(PMI(w,c),0)$$

- Truncated Singular Value Decomposition
- Skip-Gram with Negative Sampling:

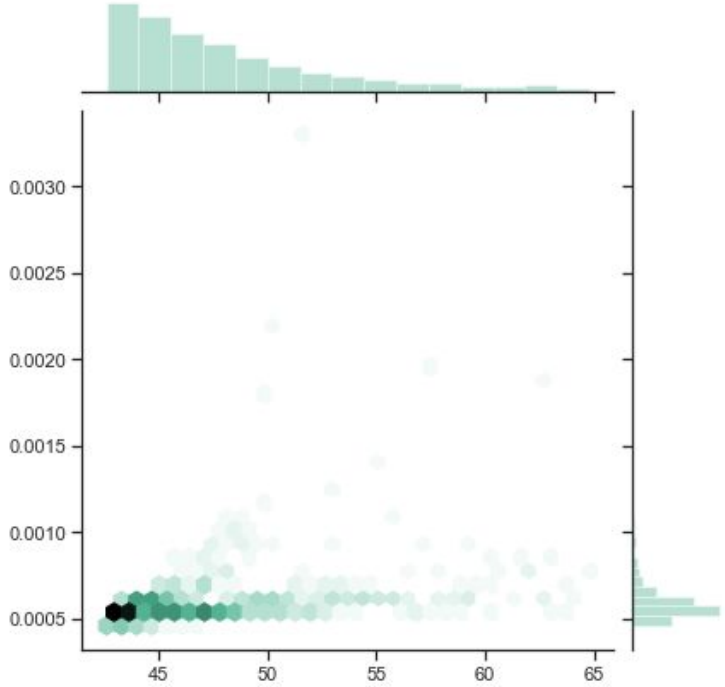
$$P(\mathbf{C}_j|\mathbf{W}_i) = Softmax(\mathbf{W}_i \cdot \mathbf{C}_j), \qquad \text{where } Softmax(\mathbf{W}_i \cdot \mathbf{C}_j) = \frac{e^{\mathbf{W}_i \cdot \mathbf{C}_j}}{\sum_k e^{\mathbf{W}_i \cdot \mathbf{C}_k}}$$

- Eigenvector Analysis Methods

Distribution of Eigenvector Elements



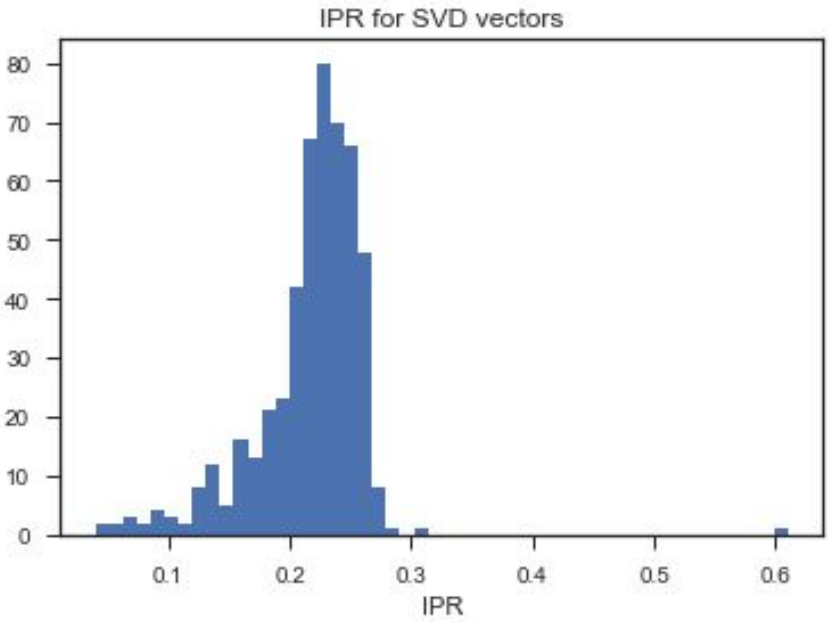
Inverse Participation Ratio

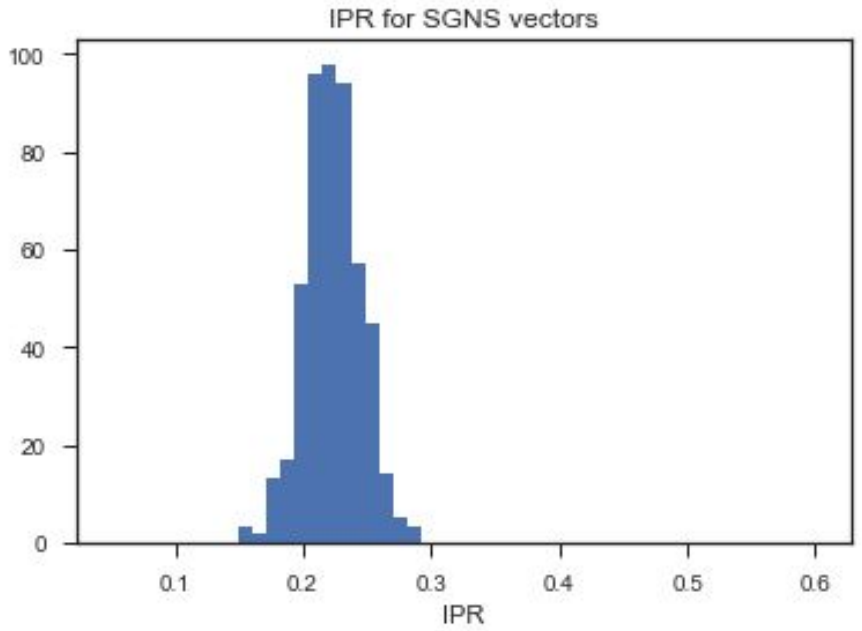


IPR for SVD

$$I^k \triangleq \sum_{i=1}^{|V|} [\mathbf{u}_i^k]^4$$

Inverse Participation Ratio





Column Space Analysis

For SVD vectors

For column 5

nashville
oregon
atlanta
nebraska
downtown
pennsylvania
philadelphia
kansas
indiana
michigan
oklahoma
seasonal
tennessee
missouri
airlines
municipal
county
kentucky
ohio
airport

For column 216

kills
eddie
edu
svg
shaw
greene
dorothy
fischer
thompson
dale
conducting
jerry
gregory
anderson
barbara
eugene
linda
helen
danny
laura

For SGNS vectors

may
living
able
decided
do
returned
shall
won
moved
said
will
can
could
must
let
should
isbn
does
did
would

confederate
allied
broke
signed
troops
throne
flag
emperor
republic
agreement
years
soviet
secretary
lieutenant
office
treaty
command
king
minister
army

allowed
owned
purchased
required
released
granted
expected
intended
sold
founded
operated
capita
able
available
median
awarded
modified
designed
considered
used

Conclusion

- Analyzed the eigenvectors, or the column space, of the word embeddings obtained from the Singular Value Decomposition of PPMI matrix.
- Compared Inverse Participation Ratio for SVD and SGNS
- Demonstratede the significant participants of the eigenvectors form semantically coherent groups

Based on article:

“Interpreting Word Embeddings with Eigenvector Analysis”Jamin Shin Andrea Madotto Pascale Fung