DATA IN MOTION

GloVe Additional Studies

22.08.2023

GloVe Additional Studies

Part Contents

Section Contents

1 GloVe: Additional Studies

Introduction

GloVe 50d Wiki+Gigaword

GloVe 100d Wiki+Gigaword

GloVe 200d Wiki+Gigaword

GloVe 300d Wiki+Gigaword

GloVe 300d Common Crawl

Summary

Conclusions

Introduction

- The pre-trained GloVe models come in different forms;
- In particular, they vary on the number and source of the trained tokens and on the dimensionality of the final vectors;
- We want to study the difference for our use-cases among the different GloVe models.

Introduction

We are going to compute the accuracy and the optimal cosine distance threshold for:

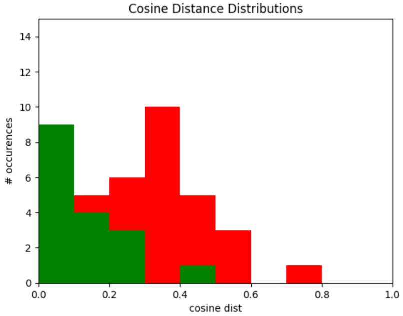
- GloVe 50d, 100d, 200d and 300d vectors taken from Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocabulary uncased);
- GloVe 300d taken from common crawl (42B tokens, 1.9M vocabulary, uncased);
- The downloaded files can be found here.

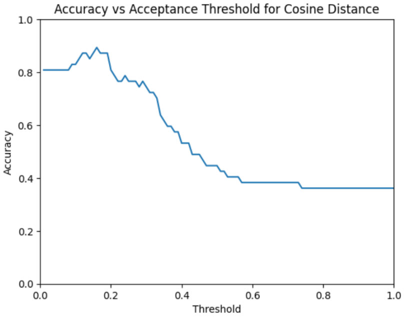
Section Contents

1 GloVe: Additional Studies

- Introduction
- GloVe 50d Wiki+Gigaword
- GloVe 100d Wiki+Gigaword
- GloVe 200d Wiki+Gigaword
- GloVe 300d Wiki+Gigaword
- GloVe 300d Common Crawl
- Summary
- Conclusions

GloVe 50d Wiki+Gigaword





GloVe 50d Wiki+Gigaword

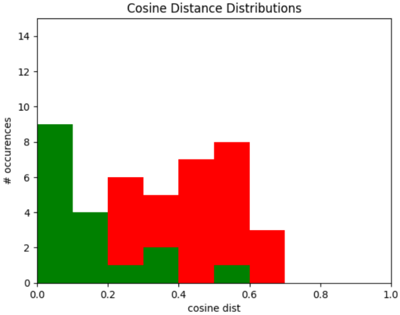
Optimal Threshold	Accuracy	Recall	Precision
0.16	0.89	0.76	0.93

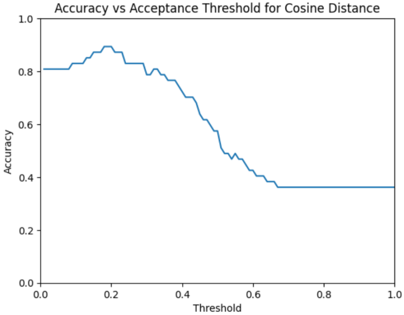
Section Contents

1 GloVe: Additional Studies

- Introduction
- GloVe 50d Wiki+Gigaword
- GloVe 100d Wiki+Gigaword
- GloVe 200d Wiki+Gigaword
- GloVe 300d Wiki+Gigaword
- GloVe 300d Common Crawl
- Summary
- Conclusions

GloVe 100d Wiki+Gigaword





GloVe 100d Wiki+Gigaword

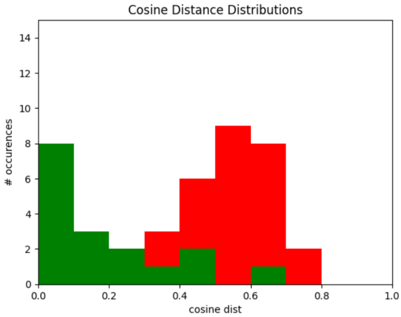
Optimal Threshold	Accuracy	Recall	Precision
0.18	0.89	0.71	1.0

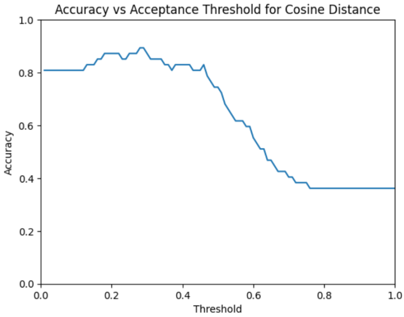
Section Contents

1 GloVe: Additional Studies

- Introduction
- GloVe 50d Wiki+Gigaword
- GloVe 100d Wiki+Gigaword
- GloVe 200d Wiki+Gigaword
- GloVe 300d Wiki+Gigaword
- GloVe 300d Common Crawl
- Summary
- Conclusions

GloVe 200d Wiki+Gigaword





GloVe 200d Wiki+Gigaword

Optimal Threshold	Accuracy	Recall	Precision
0.28	0.89	0.76	0.93

Section Contents

1 GloVe: Additional Studies

Introduction

GloVe 50d Wiki+Gigaword

GloVe 100d Wiki+Gigaword

GloVe 200d Wiki+Gigaword

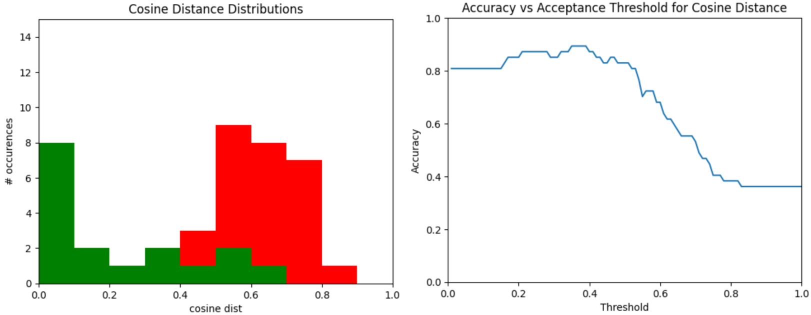
GloVe 300d Wiki+Gigaword

GloVe 300d Common Crawl

Summary

Conclusions

GloVe 300d Wiki+Gigaword



GloVe 300d Wiki+Gigaword

Optimal Threshold	Accuracy	Recall	Precision
0.35	0.89	0.76	0.93

Section Contents

1 GloVe: Additional Studies

Introduction

GloVe 50d Wiki+Gigaword

GloVe 100d Wiki+Gigaword

GloVe 200d Wiki+Gigaword

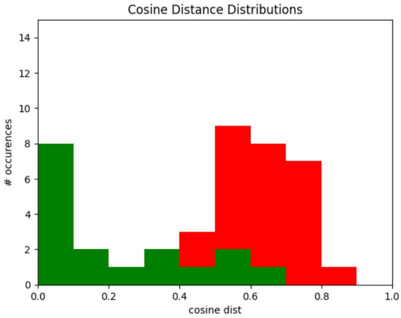
GloVe 300d Wiki+Gigaword

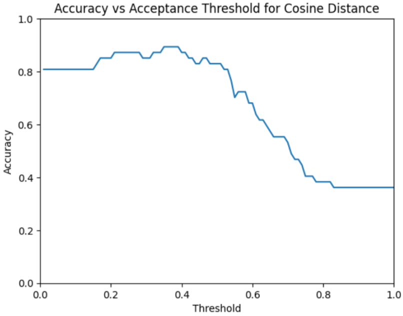
GloVe 300d Common Crawl

Summary

Conclusions

GloVe 300d Common Crawl





GloVe 300d Co mmon Crawl

Optimal Threshold	Accuracy	Recall	Precision
0.35	0.89	0.76	0.93

Section Contents

1 GloVe: Additional Studies

Introduction

GloVe 50d Wiki+Gigaword

GloVe 100d Wiki+Gigaword

GloVe 200d Wiki+Gigaword

GloVe 300d Wiki+Gigaword

GloVe 300d Common Crawl

Summary

Conclusions

Section Contents

1 GloVe: Additional Studies

- Introduction
- GloVe 50d Wiki+Gigaword
- GloVe 100d Wiki+Gigaword
- GloVe 200d Wiki+Gigaword
- GloVe 300d Wiki+Gigaword
- GloVe 300d Common Crawl
- Summary
- Conclusions

Summary

Model	Threshold	Accuracy	Recall	Precision
50d Wiki+Gigaword	0.16	0.89	0.76	0.93
100d Wiki+Gigaword	0.18	0.89	0.71	1.0
200d Wiki+Gigaword	0.28	0.89	0.76	0.93
300d Wiki+Gigaword	0.35	0.89	0.76	0.93
300d Common Crawl	0.35	0.89	0.76	0.93

Section Contents

1 GloVe: Additional Studies

- Introduction
- GloVe 50d Wiki+Gigaword
- GloVe 100d Wiki+Gigaword
- GloVe 200d Wiki+Gigaword
- GloVe 300d Wiki+Gigaword
- GloVe 300d Common Crawl
- Summary
- Conclusions

Conclusions

- No considerable differences have been observed by changing model;
- We have also to consider the loading time for such models due to their size (171MB of the 50d to 5G of 300d Common Crawl);
- It's not worth to use a more complex model that requires much time to load since the difference in performance is basically none.

Conclusion

Useful Links

OSGi Working Group

Working Group: www.osgi.org
WG Blog: www.osgi.org/blog
Twitter: @osgiwg
Bndtools: bndtools.org

Data In Motion

Web: www.datainmotion.com
Blog: datainmotion.com/blog
Twitter: @motion_data

Jürgen Albert

Email: j.albert@data-in-motion.biz

Mark Hoffmann

Email:
m.hoffmann@data-in-motion.biz