**Hyperparameter Tuning of CoCo Embeddings**

**Hyperparameters**

Those that can be varied (parameters used in paper brackets):

GloVe: vocab\_min\_count, vector size (300; tried 10-300), window size (10; tried 2-10), max iter (50 when d < 300, 100 when d=300), X\_max (100)

Mittens: mittens\_params (0.1), window size (10), embedding\_dim, mincount (300), max\_iter

Those that have been varied :

Mittens\_params, 0.05, 0.1

Embedding\_dim: 50, 100, 300

Window size (changed GloVe to 10 once when simultaneously changing embedding dim to 300, should try to change systematically)

MAX\_ITER

X\_max

Those could potentially impact the outcome and to be varied :

Move mittens window size smaller

**Hyperparameter History**

**300D Version**

GloVe

VOCAB\_MIN\_COUNT=50

VECTOR\_SIZE=300

MAX\_ITER=15

WINDOW\_SIZE=10

BINARY=2

NUM\_THREADS=12

X\_MAX=10

Mittens

mittens\_params = 0.1

window\_size = 10

embedding\_dim = 300

mincount = 150

max\_iter = 1000

This version performed worse than initial set up – most lost its significance. The output file is lost and can be generated again by using these hyperparameters.

**100D Versions**

GloVe

VOCAB\_MIN\_COUNT=5

VECTOR\_SIZE=100

MAX\_ITER=15

WINDOW\_SIZE=15

BINARY=2

NUM\_THREADS=12

X\_MAX=10

Mittens

mittens\_params = 0.1

window\_size = 10

embedding\_dim = 100

mincount = 150

max\_iter = 1000

This version performs worse than the initial set-up: i\_we lost its significance

**50D Versions**

1)

This is supposed to be the original version – needs to be run again if needed – the current embed\_orig is finetuned on a version of the vectors.txt that’s already been lost. Need to ensure original and finetuned embeddings match up.

GloVe

VERBOSE=2

MEMORY=4.0

VOCAB\_MIN\_COUNT=5

VECTOR\_SIZE=50

MAX\_ITER=15

WINDOW\_SIZE=15

BINARY=2

NUM\_THREADS=8

X\_MAX=10

Mittens

mittens\_params = 0.1

window\_size = 10

embedding\_dim = 50

mincount = 150

max\_iter = 1000

2)

GloVe

VERBOSE=2

MEMORY=4.0

VOCAB\_MIN\_COUNT=5

VECTOR\_SIZE=50

MAX\_ITER=15

WINDOW\_SIZE=15

BINARY=2

NUM\_THREADS=12

X\_MAX=10

mittens\_params = 0.05

window\_size = 10

embedding\_dim = 50

mincount = 150

max\_iter = 1000

This version doesn’t seem too different from using mittens\_params=0.1 but is a lot slower to train

3)

GloVe (Time = 74 mins)

VOCAB\_MIN\_COUNT=5

VECTOR\_SIZE=50

MAX\_ITER=30

WINDOW\_SIZE=10

BINARY=2

NUM\_THREADS=12

X\_MAX=10

Mittens (Time = 9hrs)

mittens\_params = 0.1

window\_size = 10

embedding\_dim = 50

mincount = 150

max\_iter = 1000

This produces output that gives stronger correlations with identification for both

I\_we and i\_we\_cluster

4)

GloVe (Time = 77mins)

VOCAB\_MIN\_COUNT=5

VECTOR\_SIZE=50

MAX\_ITER=30

WINDOW\_SIZE=10

BINARY=2

NUM\_THREADS=12

X\_MAX=100

Mittens (Time = 9.5hrs)

mittens\_params = 0.1

window\_size = 10

embedding\_dim = 50

mincount = 150

max\_iter = 1000

Lower pvalues and slightly higher correlation using bergami bagozzi. Output named **embedding\_user\_50d\_best.csv and embedding\_quarterly\_50d\_best.csv**

5)

GloVe (Time = 90 mins)

VERBOSE=2

MEMORY=4.0

VOCAB\_MIN\_COUNT=5

**VECTOR\_SIZE=100**

**MAX\_ITER=50**

WINDOW\_SIZE=10

BINARY=2

NUM\_THREADS=12

X\_MAX=100

Mittens (Time = 7.5hrs)

mittens\_params = 0.1

window\_size = 10

embedding\_dim = 100

mincount = 150

max\_iter = 1000

6) *Current*

CORPUS=corpus.txt

VOCAB\_FILE=vocab.txt

COOCCURRENCE\_FILE=cooccurrence.bin

COOCCURRENCE\_SHUF\_FILE=cooccurrence.shuf.bin

BUILDDIR=build

SAVE\_FILE=vectors

VERBOSE=2

MEMORY=4.0

VOCAB\_MIN\_COUNT=5

**VECTOR\_SIZE=50**

**MAX\_ITER=50**

WINDOW\_SIZE=10

BINARY=2

NUM\_THREADS=13

X\_MAX=100

**NOTE TO SELF**

Every time new hyperparams are run, make sure:

1. the same vectors.txt is used across glove\_email.sh, 2\_, and 3\_.
2. input & output match across these three files.
3. important output files are not overwritten
4. hyperparameter details are written down here

**Validation for spacespace:**

mittens\_\*50d.csv uses mincount\_timechunk = 50, and window size =10. Not great.

Mittens\_\*100d.csv uses mincount\_timechunk = 50, and window size = 5 – these are the old spacespace hyperparams. Not great.

**Current**

Mittens (Time = 9.5hrs)

mittens\_params = 0.1

window\_size = 10

embedding\_dim = 50

**mincount = 600**

max\_iter = 1000