進捗報告

Magnitude で取り扱える Popular Embedding Models を図 1 に示す．GloVe や fastText では複数のコーパスがあり，実験では語彙空間の大きそうな CommonCrawl によるものを使用していたが，現在は Wikipedia コーパスによるものも気になっている．実行時間や容量の関係（1 フォーマットにつき数 GB のデータを要する）ので試していなかったが，自然言語の観点から論じるのであれば，複数試すのも悪くないかとも考えられるので，時間が許す限り試してみたい気持ちもなくはないです．

また，各手法における Word Embedding の演算について，[king+woman-man=queen] などの定番と異なり，色に関しては [deep+pink=deeppink] や [red+orange=yellow] などの演算は難しいかもしれません．

Pre-converted Magnitude Formats of Popular Embeddings Models

Popular embedding models have been pre-converted to the `.magnitude` format for immmediate download and usage:

Contributor	Data	Light (basic support for out-of-vocabulary keys)	Medium *(recommended)* (advanced support for out-of-vocabulary keys)	Heavy (advanced support for out-of-vocabulary keys and faster `most_similar_approx`)
Google - word2vec	Google News 100B	300D	300D	300D
Stanford - GloVe	Wikipedia 2014 + Gigaword 5 6B	50D, 100D, 200D, 300D	50D, 100D, 200D, 300D	50D, 100D, 200D, 300D
Stanford - GloVe	Wikipedia 2014 + Gigaword 5 6B (lemmatized by Plasticity)	50D, 100D, 200D, 300D	50D, 100D, 200D, 300D	50D, 100D, 200D, 300D
Stanford - GloVe	Common Crawl 840B	300D	300D	300D
Stanford - GloVe	Twitter 27B	25D, 50D, 100D, 200D	25D, 50D, 100D, 200D	25D, 50D, 100D, 200D
Facebook - fastText	English Wikipedia 2017 16B	300D	300D	300D
Facebook - fastText	English Wikipedia 2017 + subword 16B	300D	300D	300D
Facebook - fastText	Common Crawl 600B	300D	300D	300D
AI2 - AllenNLP ELMo	ELMo Models	ELMo Models	ELMo Models	ELMo Models
Google - BERT	Coming Soon...	Coming Soon...	Coming Soon...	Coming Soon...

図 1: Magnitude で取り扱える Popular Embedding Models

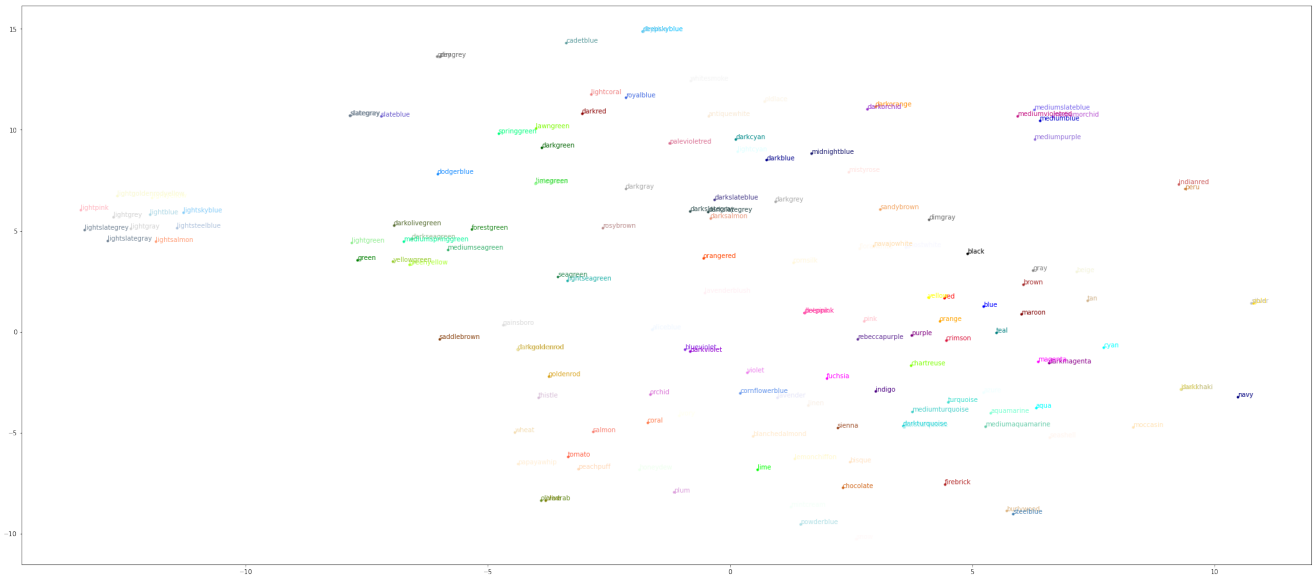

図 **2:** Word2Vec による色見本単語の分散表現 t-SNE

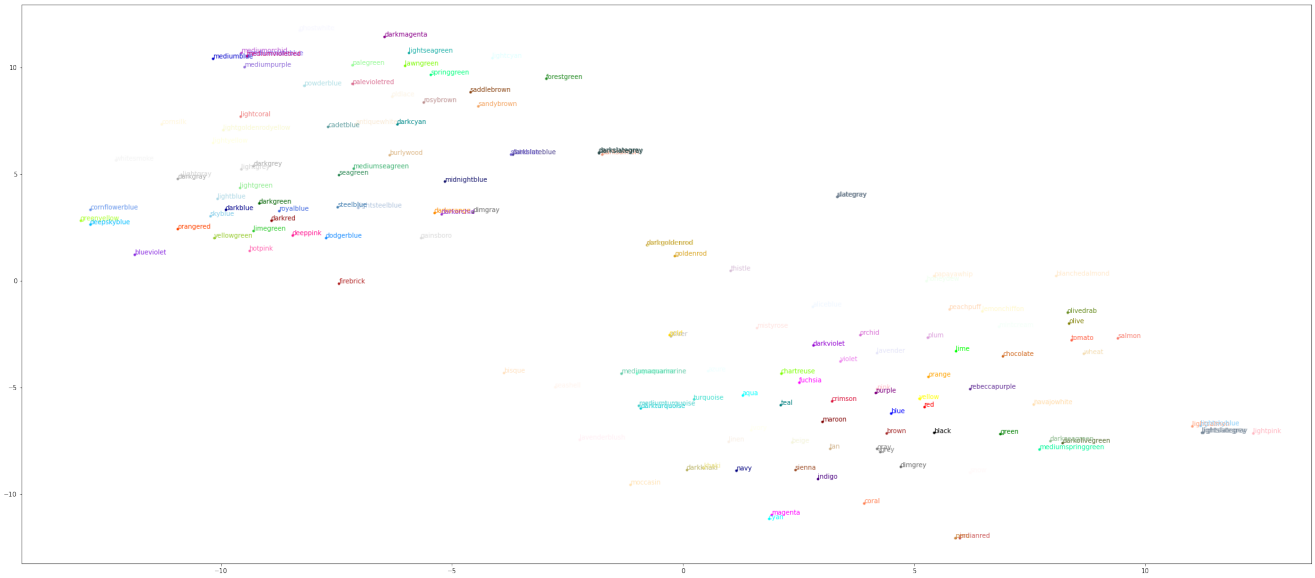

図 **3:** GloVe(CommonCrawl) による色見本単語の分散表現 t-SNE

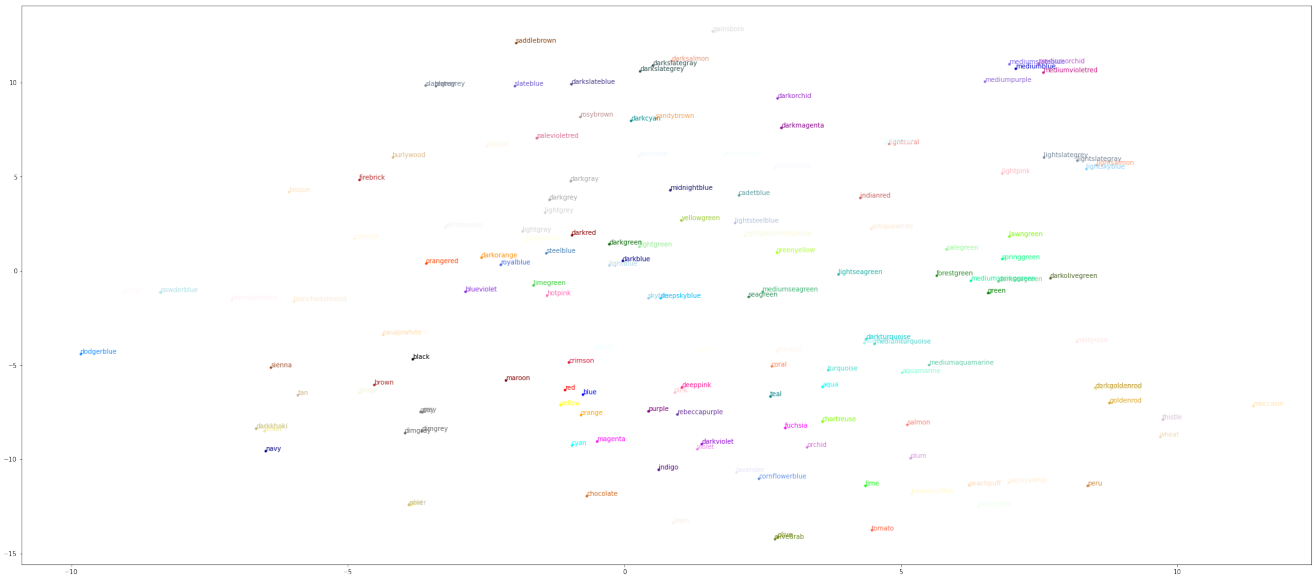

図 4: fastText(CommonCrawl) による色見本単語の分散表現 t-SNE

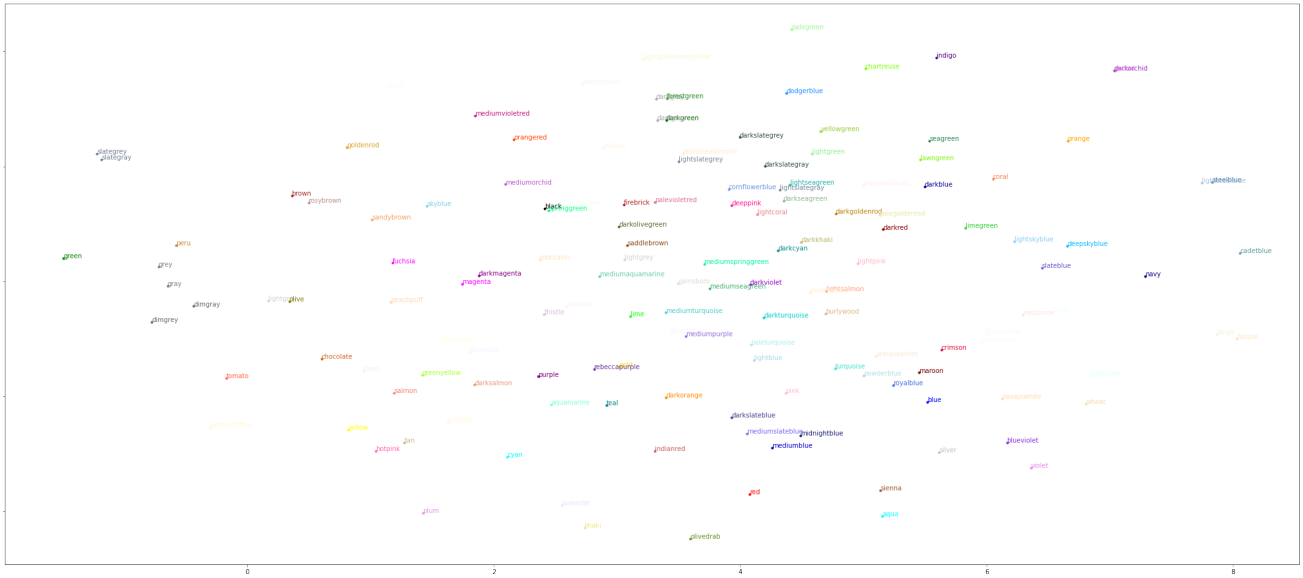

図 5: ElMo による色見本単語の分散表現 t-SNE