Paper Discovery via Topic Modelling

Usman Anwar

Let’s Break The Title

“Paper Discovery” via Topic Modelling:

- Find papers (or documents, in general) that you *do not know* about
- Akin to finding new movies to watch (Netflix), new songs to listen to (Spotify)
- Kind of like a recommender engine but important differences
    - Netflix recommender engine does not analyze the movie itself; it only works on the basis of what kind of ‘user’ you are and what kind of movies users like you watch.

Let’s Break The Title

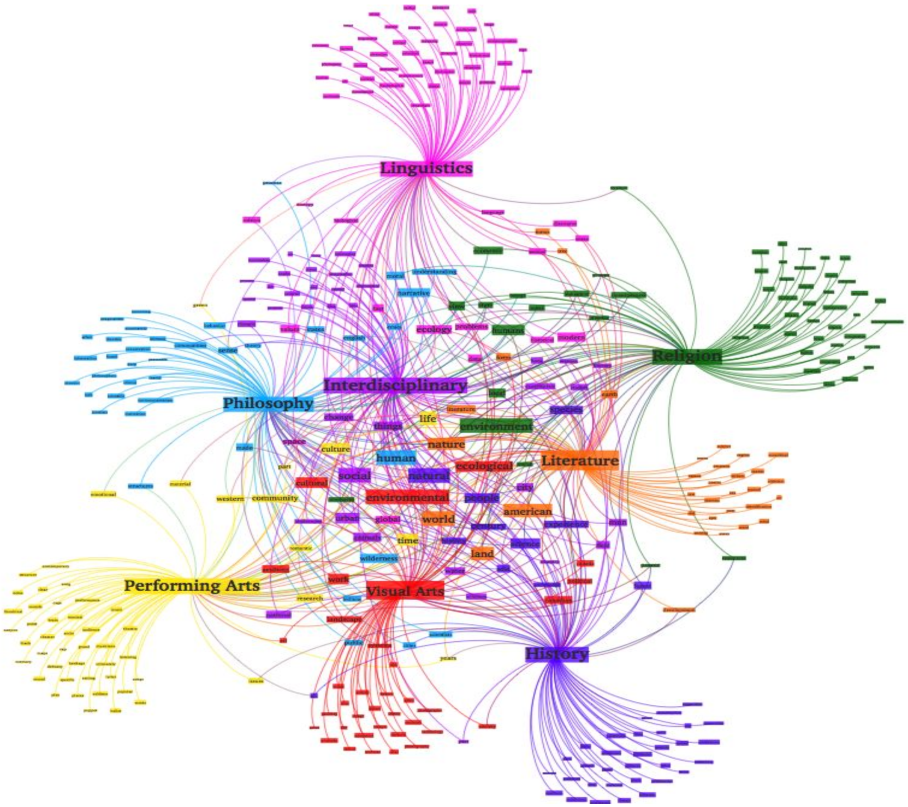
“Paper Discovery” via Topic Modelling:

- Find papers (or documents, in general) that you *do not know* about
- Akin to finding new movies to watch (Netflix), new songs to listen to (Spotify)
- Kind of like a recommender engine but important differences
  - Netflix recommender engine does not analyze the movie itself; it only works on the basis of what kind of ‘user’ you are and what kind of movies users like you watch

Paper Discovery via “Topic Modelling”:

- “In Machine Learning and Natural Language Processing, a topic model is a type of statistical model for discovering abstract “topics” that occur in a collection of documents.

Canonical Example of Topic Modelling



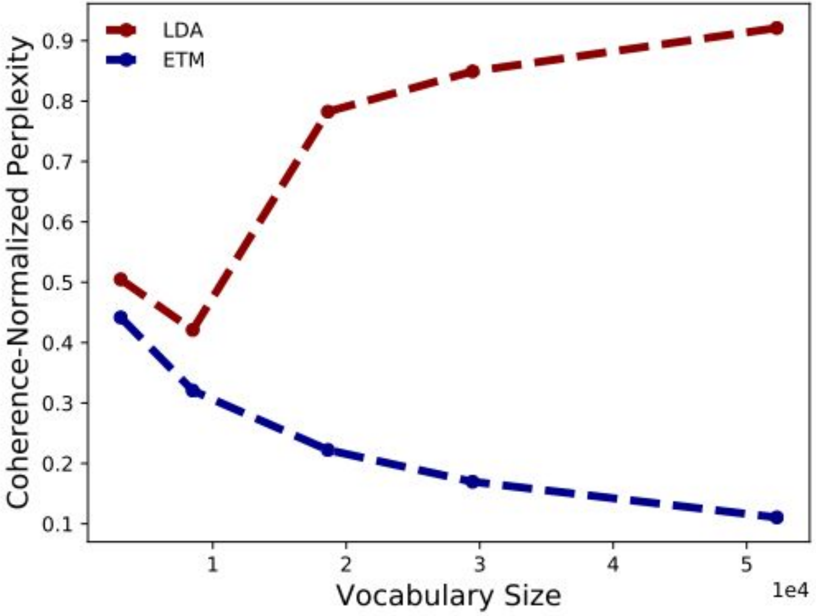
Objectives

- Develop a model that can discover a compact representation for a given paper in terms of some ‘topic’ vectors
  - Compact representation:
    - Memory efficient representation in form a fixed length vector.
  - ‘Topic’ vectors:
    - The set of basis vectors whose combinations can provide the representation of all the documents in the corpus
- Model should have these characteristics:
  - Should be expressive enough to capture topics at the required granularity
    - Technically speaking should have high intra-topic coherence and high inter-topic diversity
  - Should provide fast inference
  - Should be interpretable

Why LDA Does Not Cut It

- Latent Dirichlet Allocation is the classical method for topic modelling
- However, LDA has numerous fatal flaws:
  - It does not scale well with data.
  - Computationally expensive.
  - Compared to neural network, LDA does not preserve linear regularities of data well.

On right, we show the performance of LDA compared with Embedded Topic Model (a neural network based topic model) with respect to data. Note that while LDA performance deteriorates with respect to vocabulary size, neural network based model continues to show improvement with increase in vocabulary size.



Data

- For testing implementations and comparing with reported results:
    - 20 newsgroup dataset - canonical dataset for topic modelling
- For paper discovery task:
    - Nips papers dataset: https://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015
        - 11463 papers from 1987-2015 published in NIPS

- Preprocessing
    - Using spacy’s pipeline for tokenization, lemmatization and stop words removal
    - Further, we remove title, author names and affiliations and references list from paper text

Model

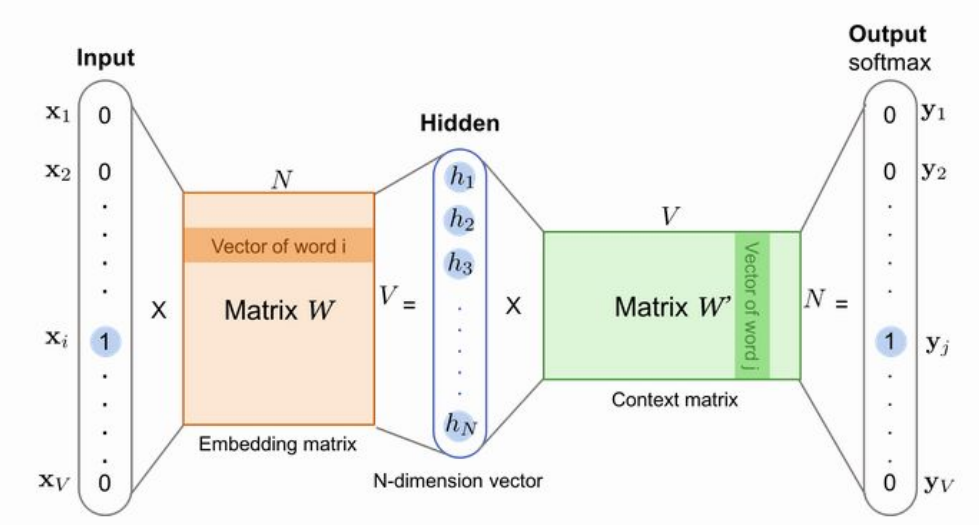
Key Idea: Word vectors have been shown to be quite successful in capturing latent structure of the language at word level, can we somehow extend the concept of embeddings to attain a topic model over documents?

There are two papers that incorporate this key idea:

- Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2019). Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.
  - One of the authors is David Blei (co-author of the original LDA paper).
  - Modifies the continuous bag of words approach of word2vec.
- Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.
  - Has around 80 citations.
  - Modifies skip gram approach of word2vec.

Model: Quick Recap of Word2Vec

- Train a model to predict a word based on its surrounding words (skip-gram) or predict surrounding words based on given word (continuous bag of words).



Model: Quick Recap of Word2Vec

- Train a model to predict a word based on its surrounding words (skip-gram) or predict surrounding words based on given word (continuous bag of words).
- Negative samping loss (shown below) is most commonly used to train word vectors. Intuitively, it maximizes the dot product of words that co-occur and minimizes the dot product of words that do not co-occur.

$$\mathcal{L}_\theta = -[\log \sigma({v'_w}^\top v_{w_I}) + \sum_{\substack{i=1 \\ \tilde{w}_i \sim Q}}^{N} \log \sigma(-{v'_{\tilde{w}_i}}^\top v_{w_I})]$$

Model: lda2vec

- Lda2vec modifies the negative sampling skip gram objective:
  - It posits topics as vectors in the same embedding space as word vectors and documents are linear combination of topic vectors.

$$d_j = p_{j0} t_0 + p_{j1} t_k + \ldots + p_{jk} t_k$$

  - It loosely conditions the generation of target word on document by modifying the pivot word vector as follows

$$v = v_w + d_w \text{ where}$$

$$v_w = \text{original word embedding of pivot word}$$

$$d_w = \text{embedding of the document to which}$$

$$\text{pivot-target word pair belongs}$$

  - Further, it adds a Dirichlet Likelihood Loss.

$$\mathcal{L}^d = \lambda \sum_{jk} (\alpha - 1) p_{jk}$$

Model: Embedded Topic Model (ETM)

- ETM also posits topics as vectors in the same embedding space as word vectors and document vectors as linear combination of topic vectors. However, there are important differences from lda2vec:
  - It uses logistic normal as a prior
    - Logistic normal captures correlated topics better.
  - Its training scheme is much more stable
    - It tightly conditions target word on the document in probabilistic terms and maximizes the following variational lower bound over the word embeddings U and topic embeddings t where p_d is topic proportion vector for document d and w_{nd} is nth word of dth document and W_d is dth document to train a neural network q with ϕ as its variational parameters.

$$\mathcal{L}(t,U,\phi) = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{nd}|p_d, U, t)] - \sum_{d=1}^{D} KL(q(p_d; W_d, \phi)||p(p_d))$$

Results - lda2vec

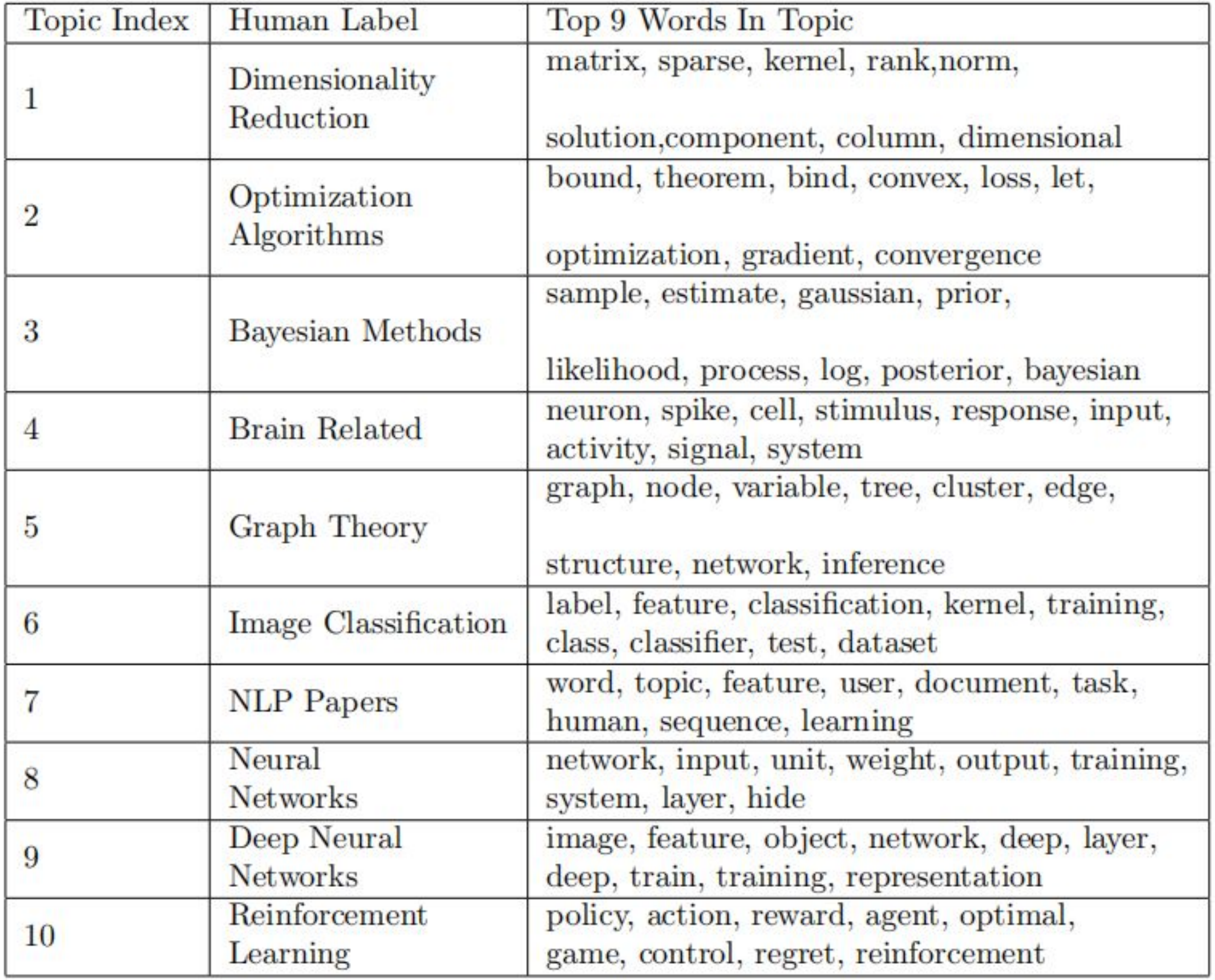
- Lda2vec does not work: distribution of topic embeddings in lda2vec quickly suffers from mode collapse i.e. model only produces one or two unique topics.
- This is because the conditioning imposed by lda2vec on producing the target word conditioned on ‘document’ embedding is too weak to prevent
- Consider the forward pass of word2vec and lda2vec for context word c, pivot/target word p

$$\begin{aligned} p_{word2vec} &= \mathcal{C}(E(c)) \\ p_{lda2vec} &= \mathcal{C}(E(c) + d) \end{aligned}$$

- Model learns to ignore ‘d’ by replacing it with a constant by making all topic embeddings identical

Results - ETM

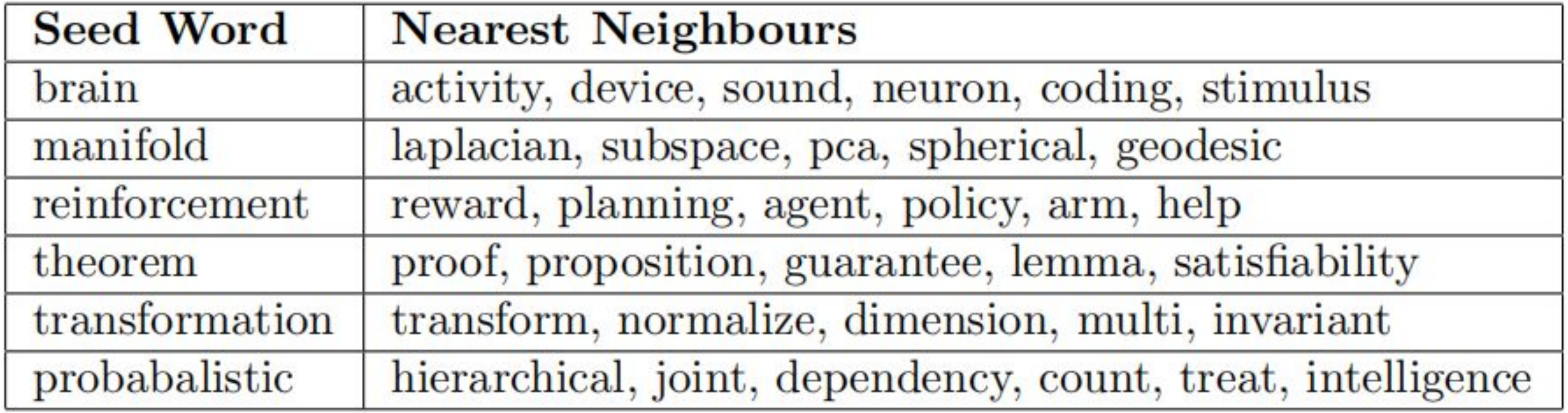
Topics Discovered By Model



Topic Index	Human Label	Top 9 Words In Topic
1	Dimensionality Reduction	matrix, sparse, kernel, rank,norm, solution,component, column, dimensional
2	Optimization Algorithms	bound, theorem, bind, convex, loss, let, optimization, gradient, convergence
3	Bayesian Methods	sample, estimate, gaussian, prior, likelihood, process, log, posterior, bayesian
4	Brain Related	neuron, spike, cell, stimulus, response, input, activity, signal, system
5	Graph Theory	graph, node, variable, tree, cluster, edge, structure, network, inference
6	Image Classification	label, feature, classification, kernel, training, class, classifier, test, dataset
7	NLP Papers	word, topic, feature, user, document, task, human, sequence, learning
8	Neural Networks	network, input, unit, weight, output, training, system, layer, hide
9	Deep Neural Networks	image, feature, object, network, deep, layer, deep, train, training, representation
10	Reinforcement Learning	policy, action, reward, agent, optimal, game, control, regret, reinforcement

Results - ETM

Inspecting Word Embeddings: Nearest Neighbours of Seed Words

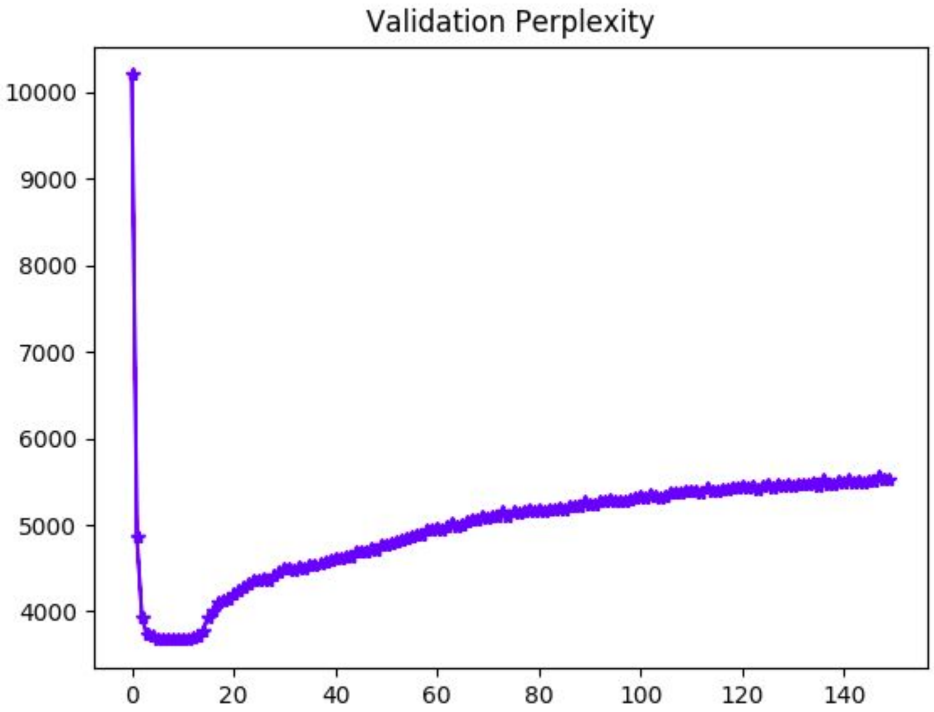


Seed Word	Nearest Neighbours
brain	activity, device, sound, neuron, coding, stimulus
manifold	laplacian, subspace, pca, spherical, geodesic
reinforcement	reward, planning, agent, policy, arm, help
theorem	proof, proposition, guarantee, lemma, satisfiability
transformation	transform, normalize, dimension, multi, invariant
probabalistic	hierarchical, joint, dependency, count, treat, intelligence

Results - ETM

Validation Perplexity: *Measure the degree to which a probability distribution predicts a sample. Lower the value, better the learned distribution.*

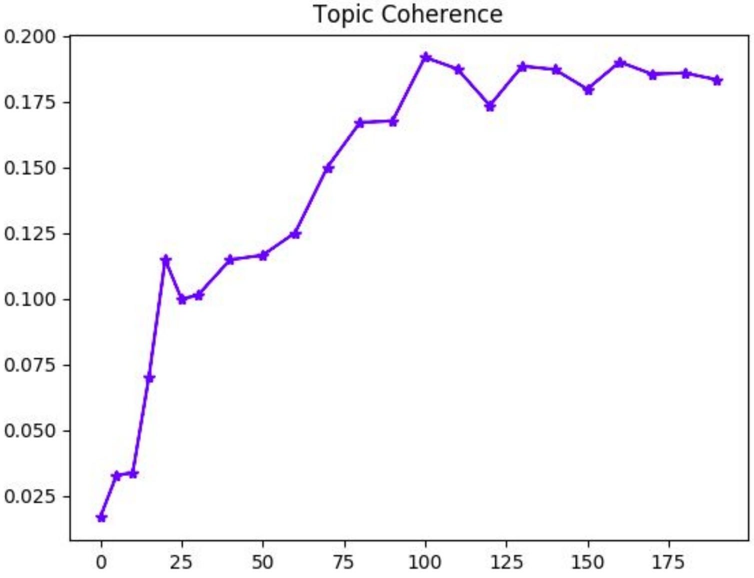
Perplexity is a general way to measure how good a language model is; however, it does not correlate with human interpretable models.

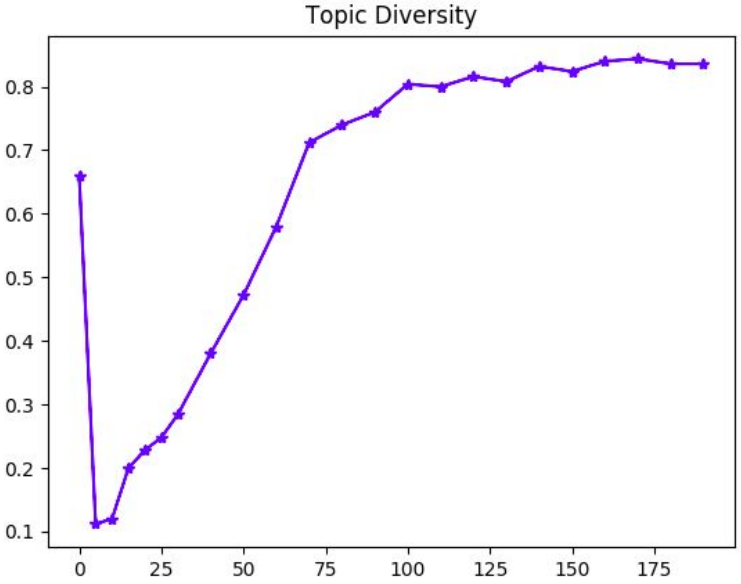


Results - ETM

Topic Coherence: *Measures the average internal coherence of topics.*

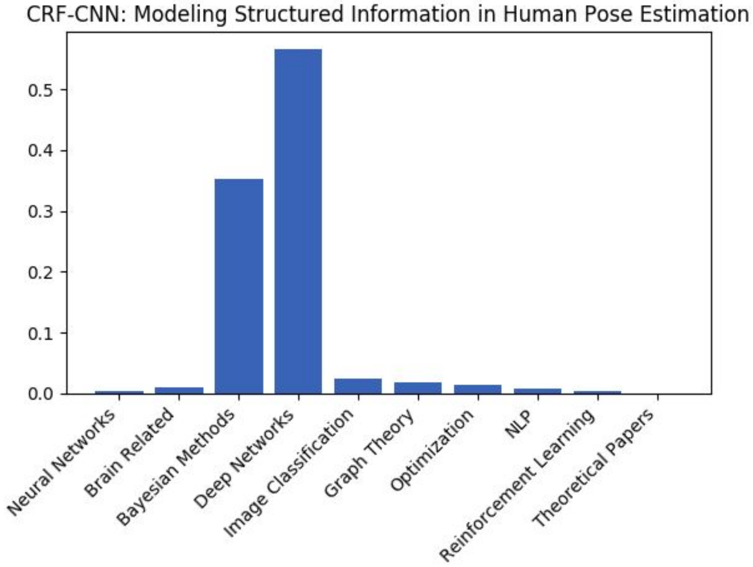
Topic Diversity: *Measures how different topics are*

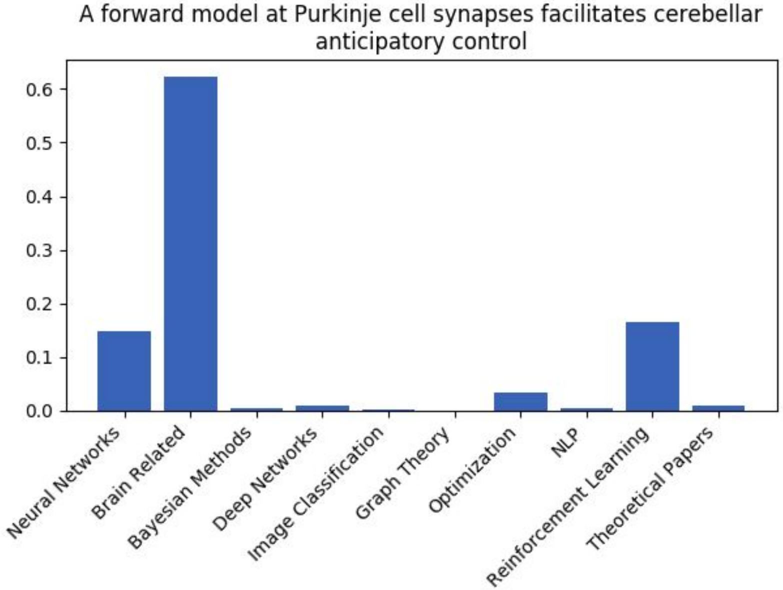




Results - ETM

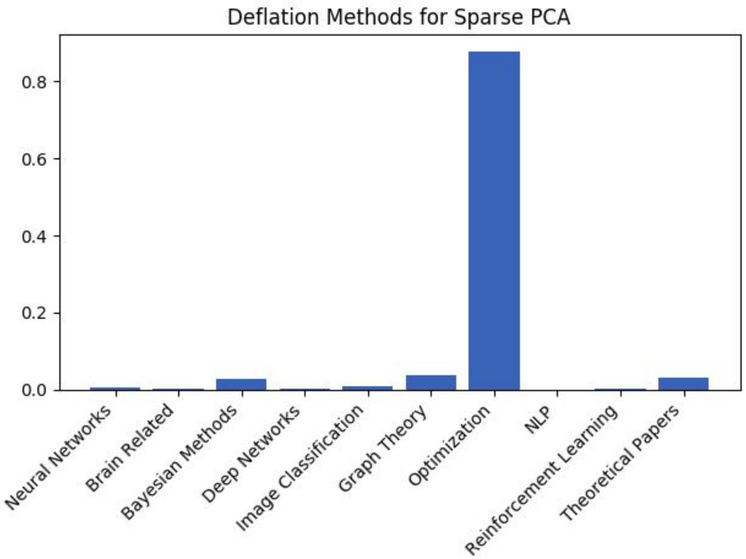
(Topic Distributions Of Some Papers)

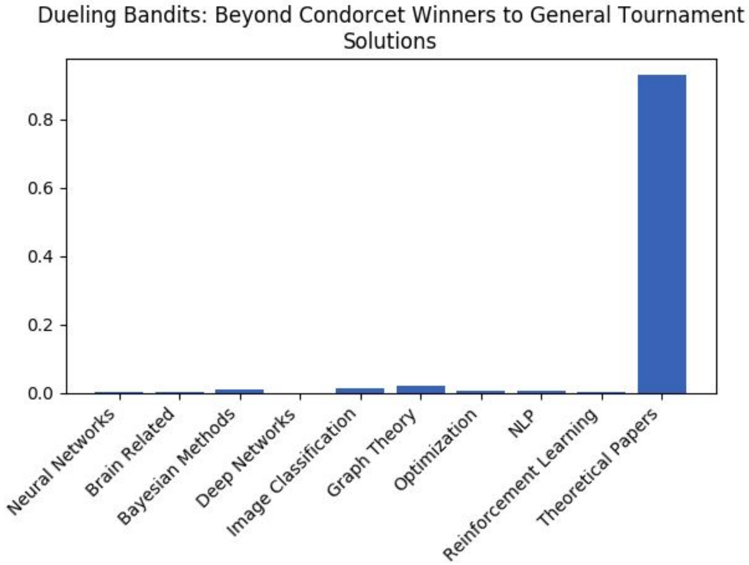




Results - ETM

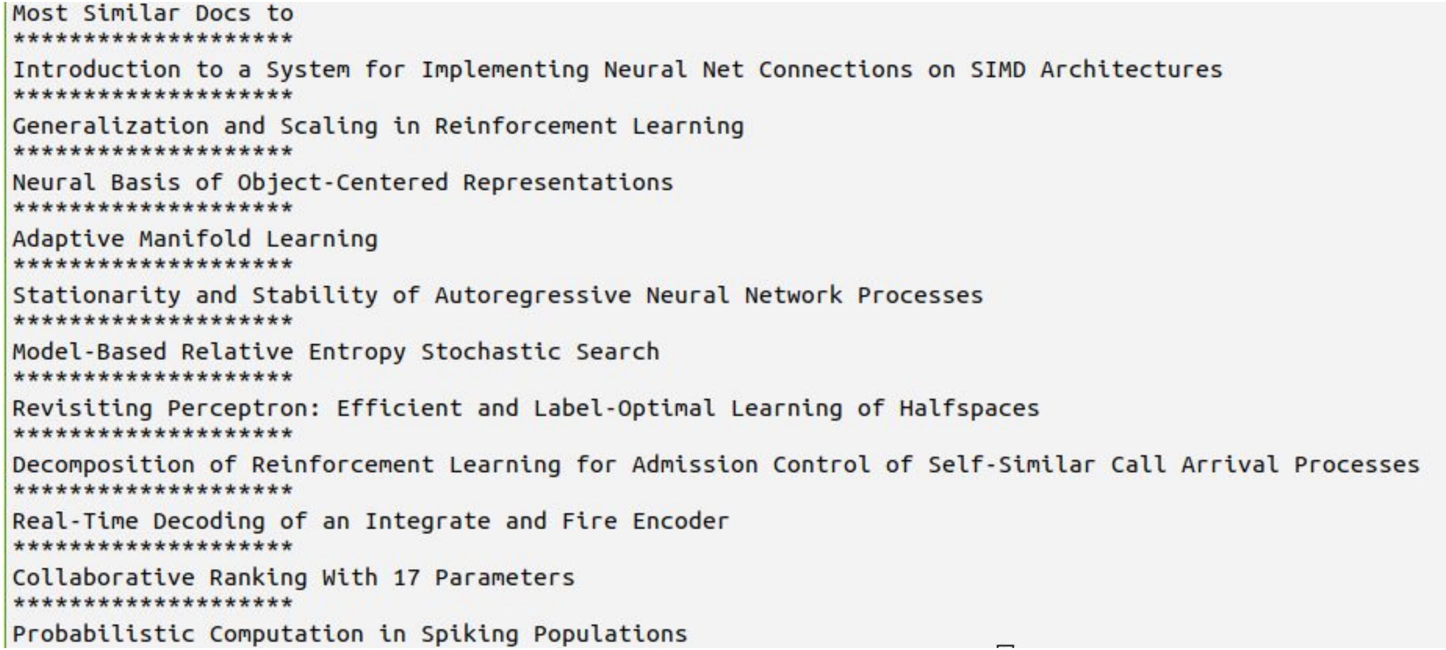
(Topic Distributions Of Some Papers)





Paper Discovery

Learn topic distribution from seed paper and provide similar papers from the dataset (database)



```
Most Similar Docs to
********************
Introduction to a System for Implementing Neural Net Connections on SIMD Architectures
********************
Generalization and Scaling in Reinforcement Learning
********************
Neural Basis of Object-Centered Representations
********************
Adaptive Manifold Learning
********************
Stationarity and Stability of Autoregressive Neural Network Processes
********************
Model-Based Relative Entropy Stochastic Search
********************
Revisiting Perceptron: Efficient and Label-Optimal Learning of Halfspaces
********************
Decomposition of Reinforcement Learning for Admission Control of Self-Similar Call Arrival Processes
********************
Real-Time Decoding of an Integrate and Fire Encoder
********************
Collaborative Ranking With 17 Parameters
********************
Probabilistic Computation in Spiking Populations
```

Conclusion

- Neural topic modelling approaches has the potential to provide improved performance by leveraging the vast amount of data available.

Future Directions

- Hierarchical Topic Modelling With Non-Euclidean Embeddings: Euclidean space is ‘flat’ and does not allow to incorporate hierarchical structure. Recent work on learning word embeddings on spherical and hyperbolic manifolds has shown that these manifolds are naturally more suitable to learning better word embeddings.
- Topic modelling for paper discovery may benefit by incorporating metadeta such as publishing venue, publication year, authors etc.
- Combining topic modelling and summarization in a single product may significantly ease the process of browsing documents in multiple fields: legal documents, medical reports etc.

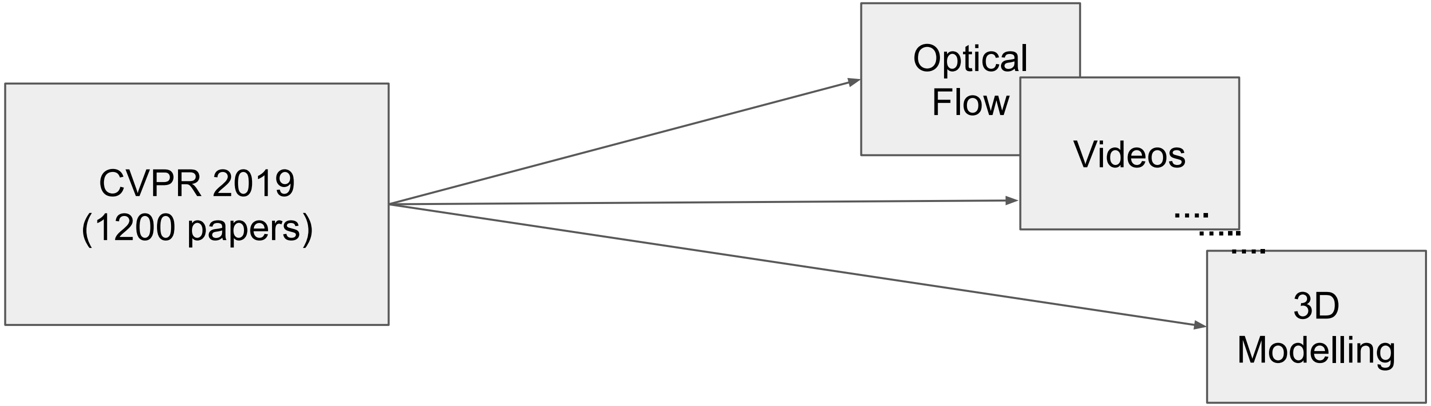
Questions?
Suggestions?

Use Case 1: Recommending Papers

- Users interest in different topics may be gauged in some way
  - Based on past history or by getting a user to fill in a form
- Topic Based Recommendations: User can then be shown papers from topics of his/her interest.
- Paper Based Recommendations: If user has shown interest in a particular paper, we can discover nearest neighbours of that paper in embedding space and recommend them.

Use Case 2: Information Summarization

- Consider a user that is looking to browse some journal or conference proceeding (basically a collection of papers on the order of 100's); by using a topic model, we can assign labels to papers and this can significantly ease the process of finding relevant or interesting papers for the user.


CVPR 2019
(1200 papers)
Optical Flow
Videos
....
.....
....
3D Modelling