NYU

Gaussian Dirichlet Process is all you need

Gyanesh Gupta, Rohith Mukku, Rahul Ahuja

12.10.21

Issues with current embeddings

- Capture sense awareness
  - Word2Vec/glove - No sense information
  - GPT/BERT/ELMO - Sense aware, but too contextual

Issues with current embeddings

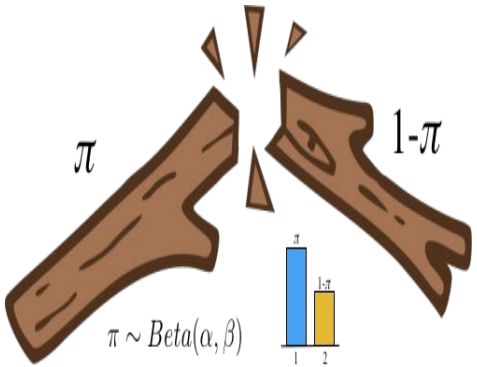
- Capture sense awareness
  - Word2Vec/glove - No sense information
  - GPT/BERT/ELMO - Sense aware, but too contextual
- No notion of hierarchy
  - Man is a broad concept, statistician is not
  - Are all bugs insects ?

Issues with current embeddings

- Capture sense awareness
  - Word2Vec/glove - No sense information
  - GPT/BERT/ELMO - Sense aware, but too contextual
- No notion of hierarchy
  - Man is a broad concept, statistician is not
  - Are all bugs insects ?
- How many senses ?
  - Mostly determined by trial and error (Multimodal Word Distributions, GMMs, *Athiwaratkun and Wilson*, 2017)
  - No control over how fine our notion of “sense” is

Our Solution : Dirichlet Process

- (Probabilistic) Word embeddings as a mixture of multiple Gaussians
- Each Gaussian represents a word sense
- Number of senses to be found using “stick breaking process” (not manually)
- All above parameters to be learned using Expectation-Maximization Algorithm



Methodology

Large Corpus

Methodology

Large Corpus

River Bank

Bank of America

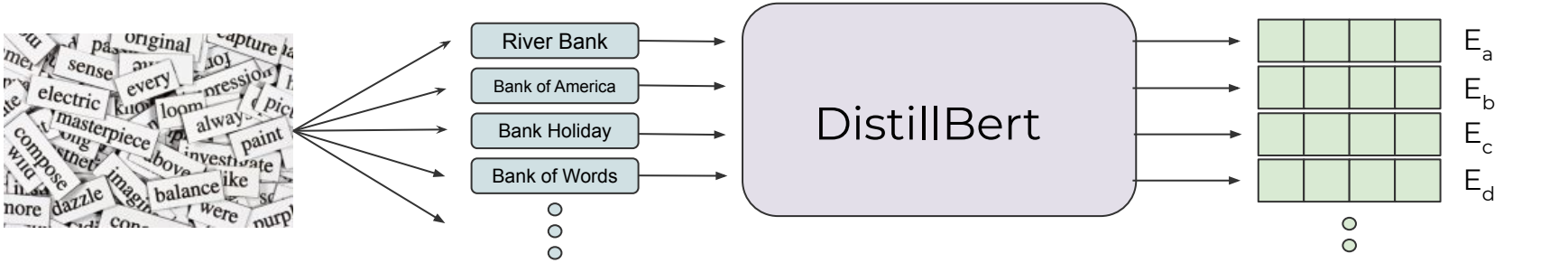
Bank Holiday

Bank of Words

Selected Words

NYU

Methodology



Large Corpus

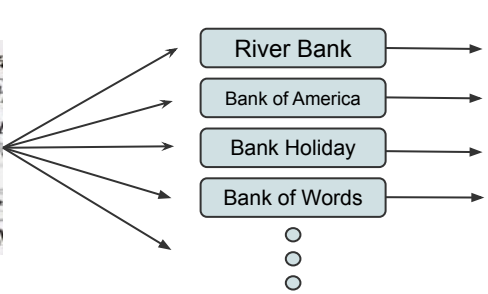
Selected Words

Model

Context Aware Embeddings

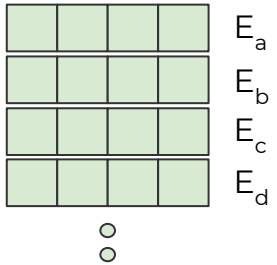
Methodology

Large Corpus

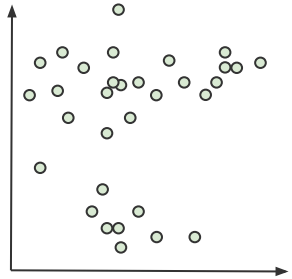


Selected Words

Model

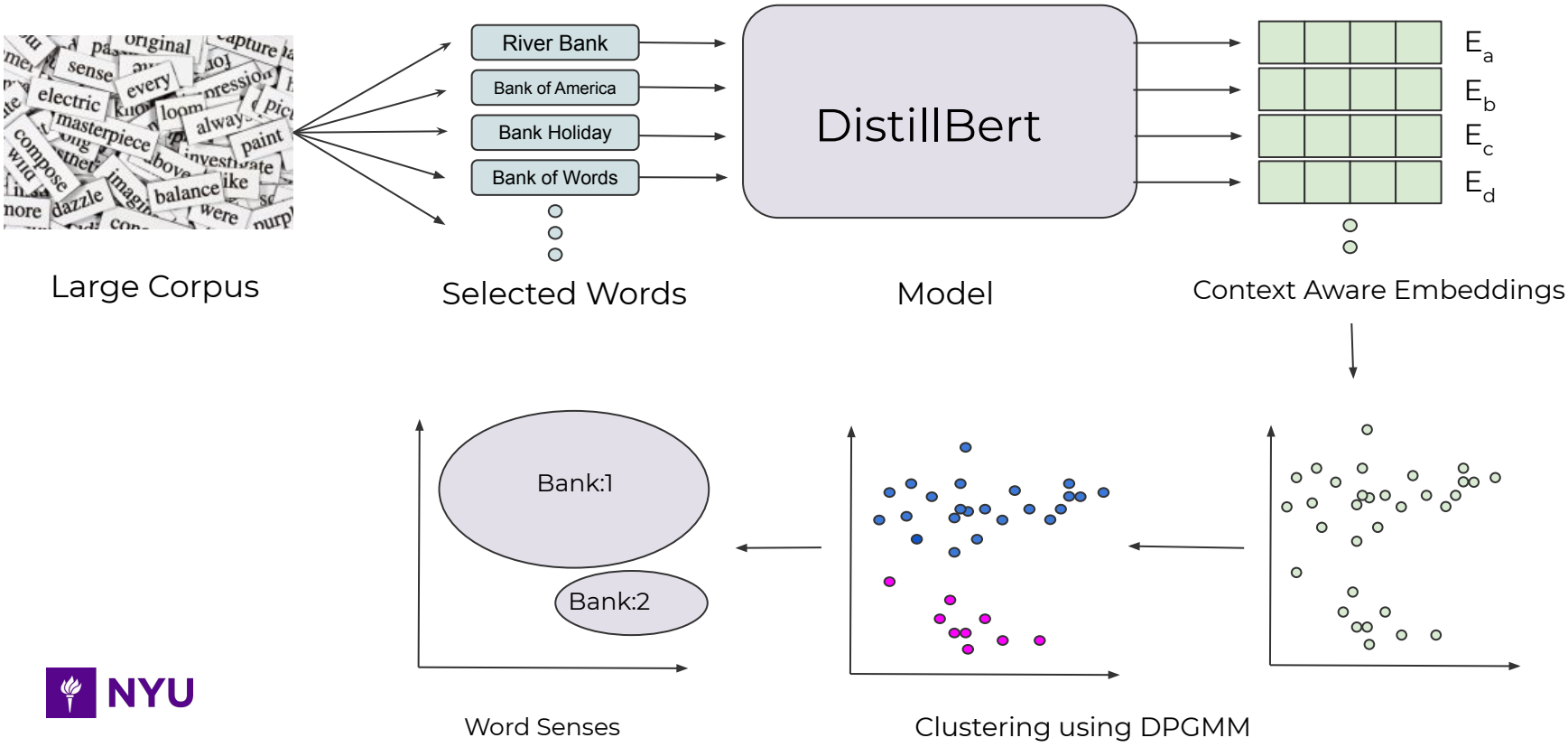


Context Aware Embeddings

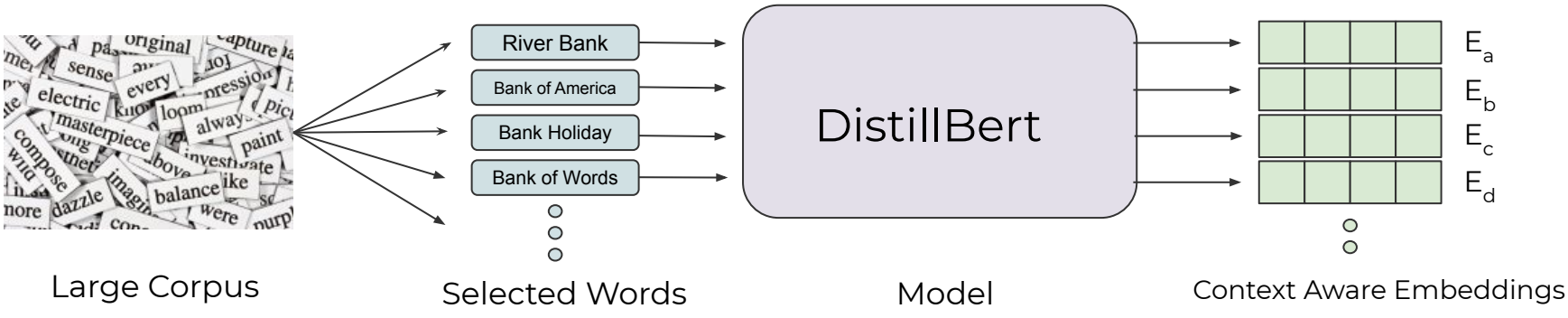


NYU

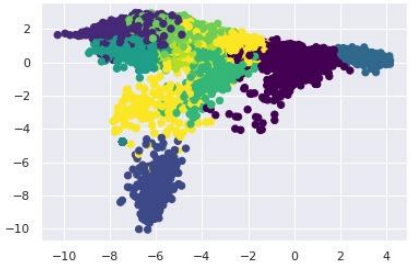
Methodology



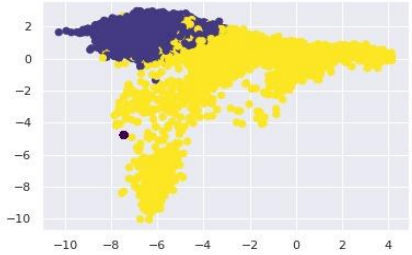
Methodology



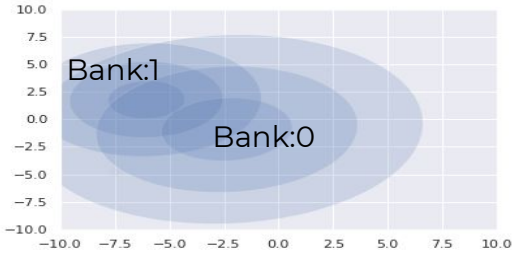
Key Findings/Analysis



GMM for “bank”



DPGMM for “bank”

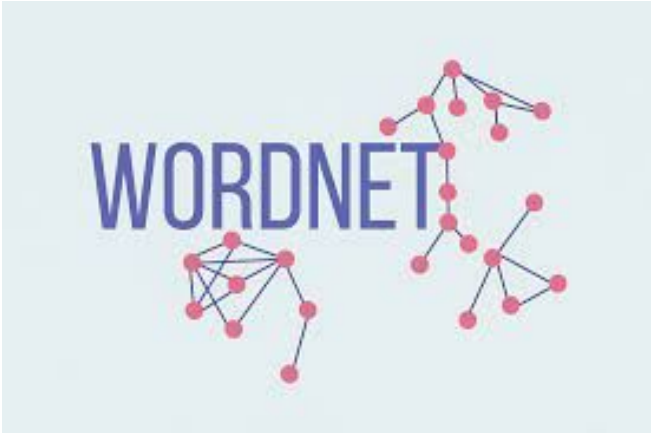


Gaussians for “bank

- How to perform Entailment or find broadness of a word ? KL Divergence of word distributions

Word-Pair	KL Divergence
Cat - animal	602.288
Animal - cat	1766.199
Airplane-animal	720.583

The Bigger Picture



More in the report...

NYU


NYU
Thank you