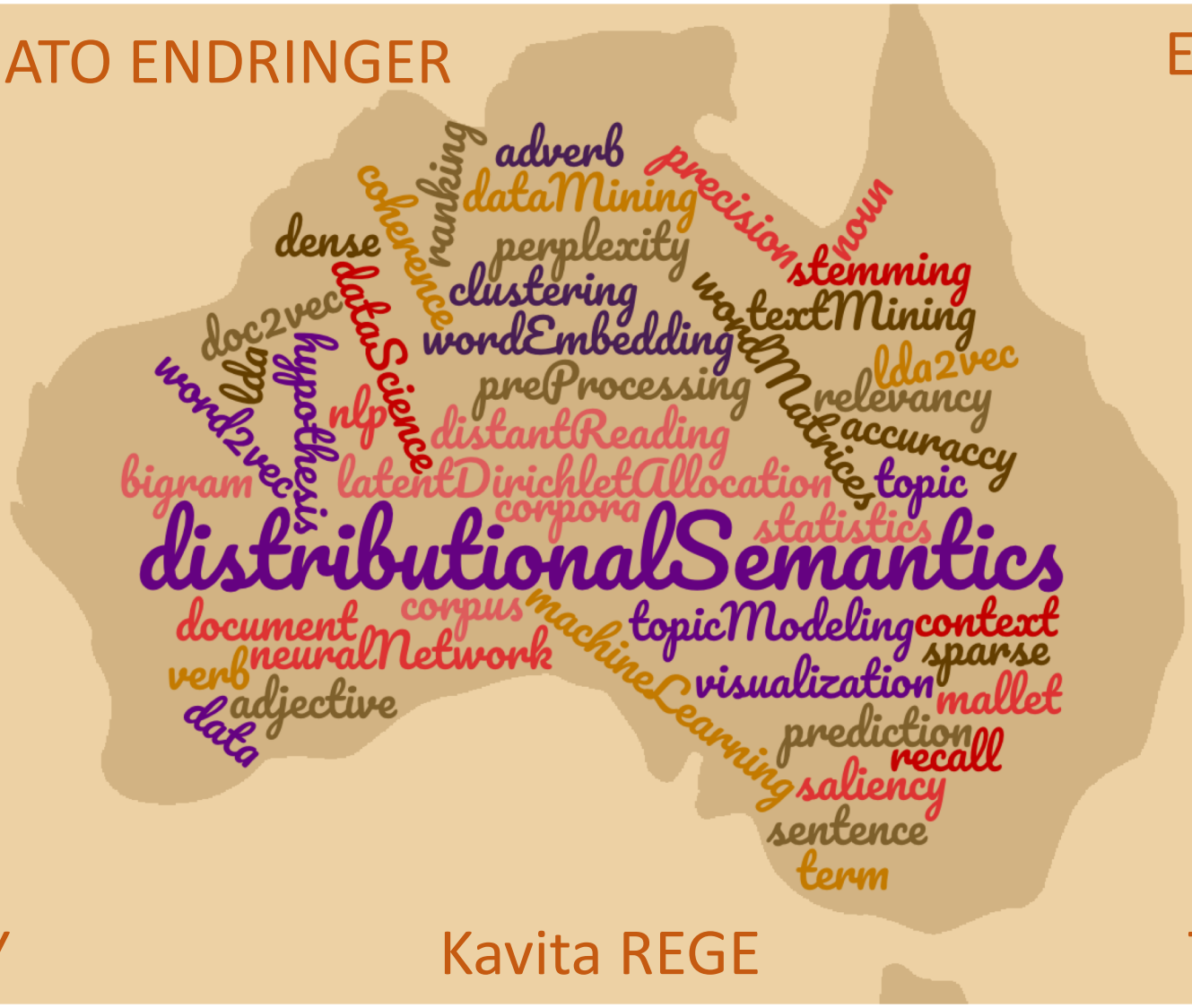
Dioni REBONATO ENDRINGER

Ema KEPUSKA



Elizabeth JOY

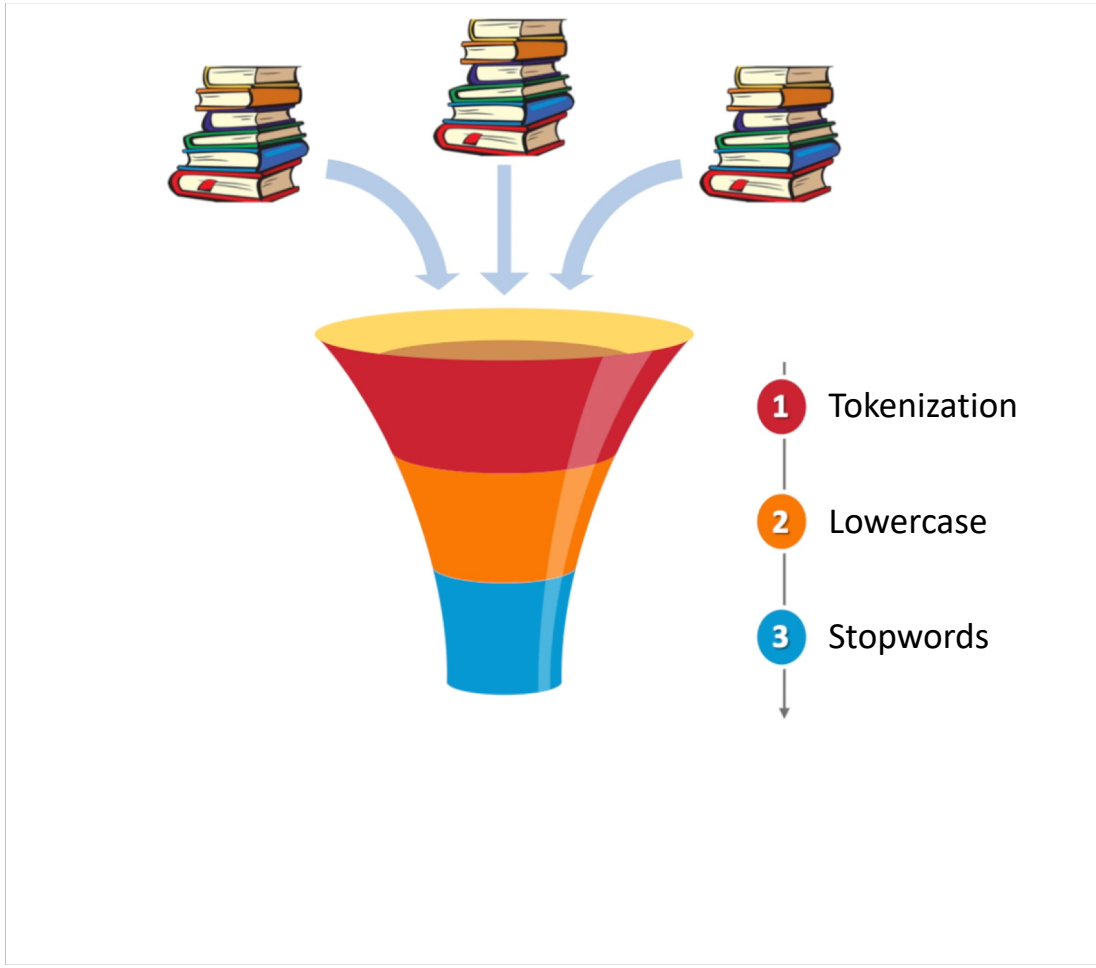
Kavita REGE

Tolga SAGLIK

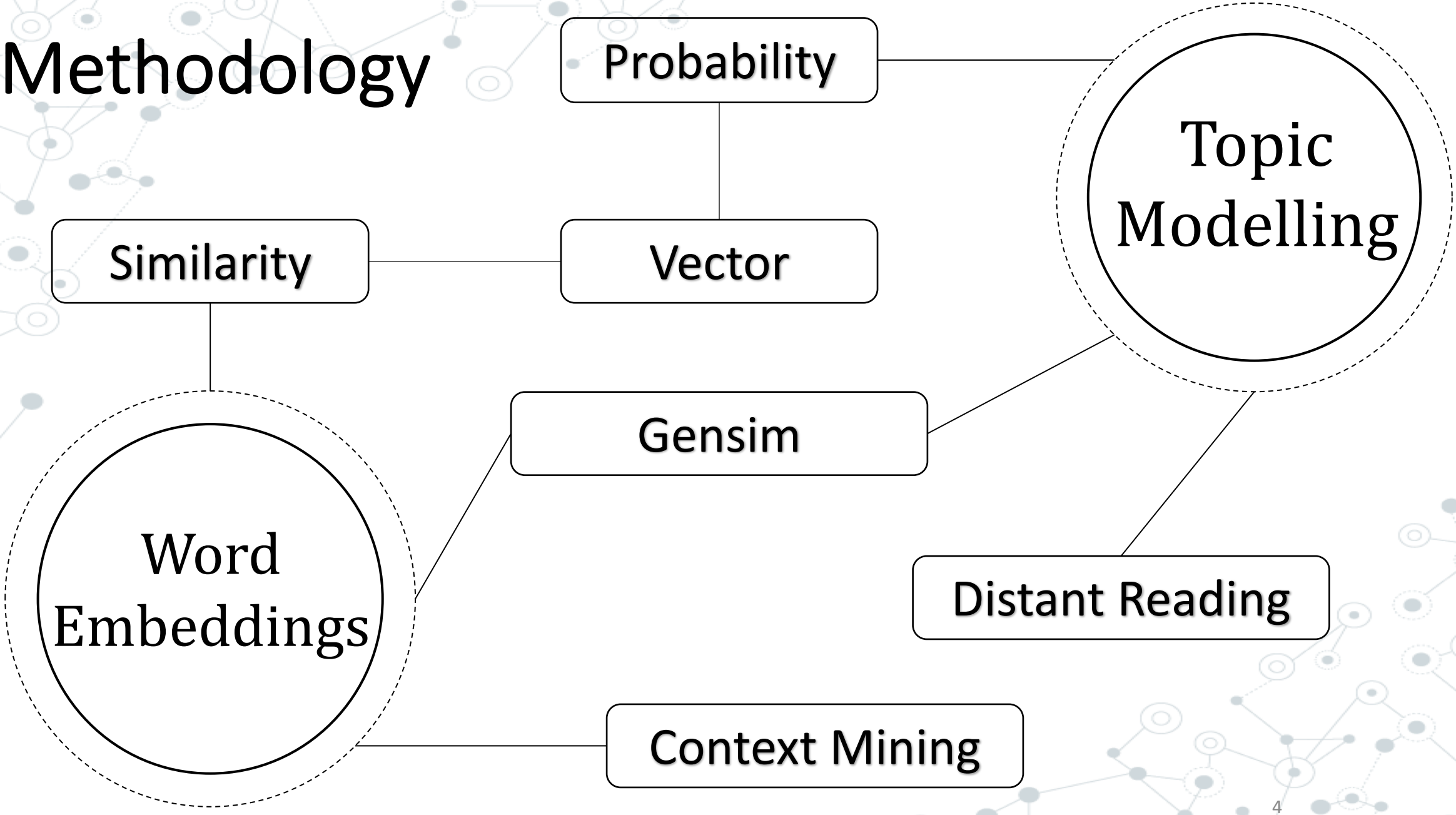
Advisor: Ekaterina Kamlovskaya

Motivation

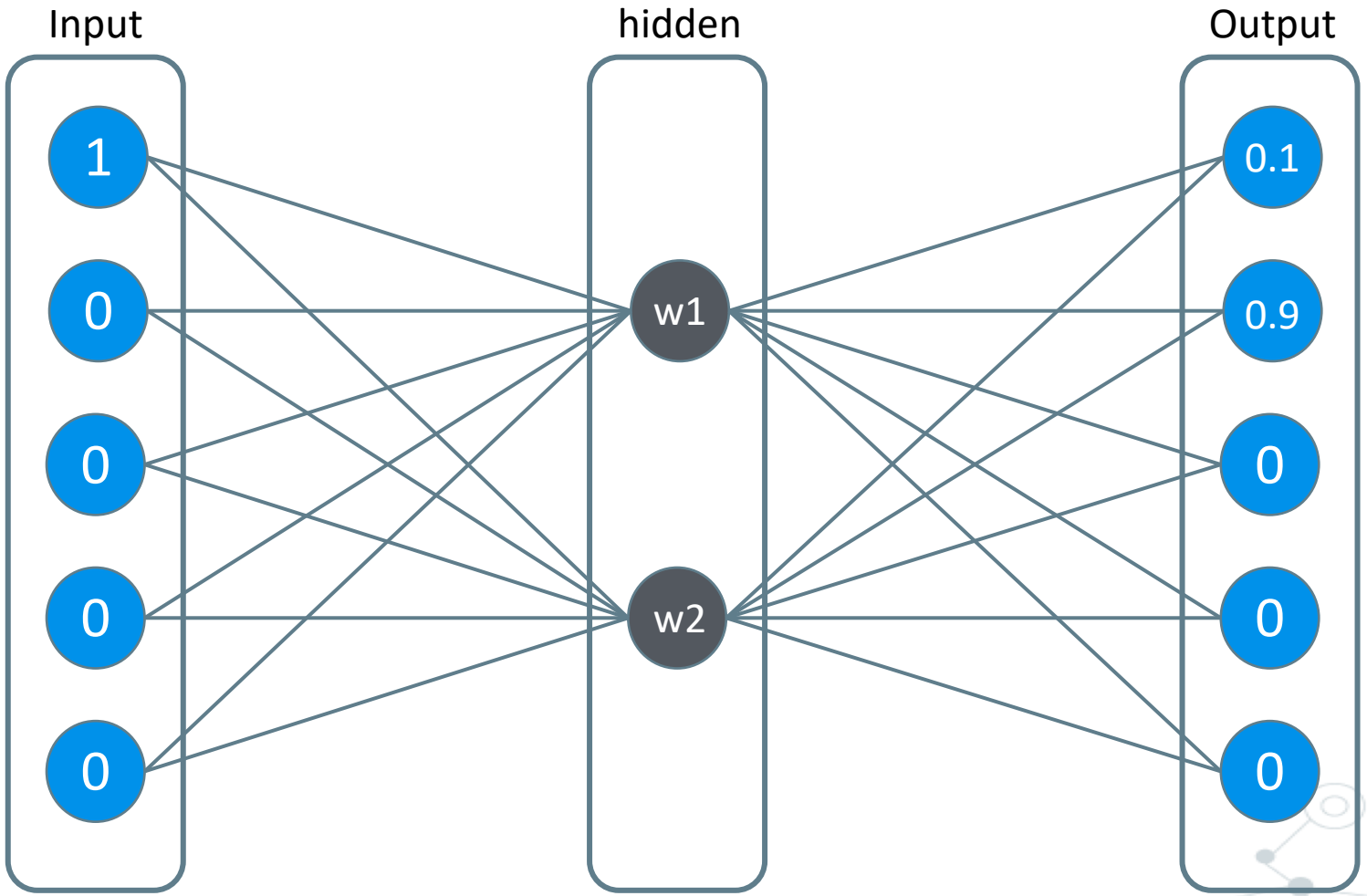
Data Preprocessing



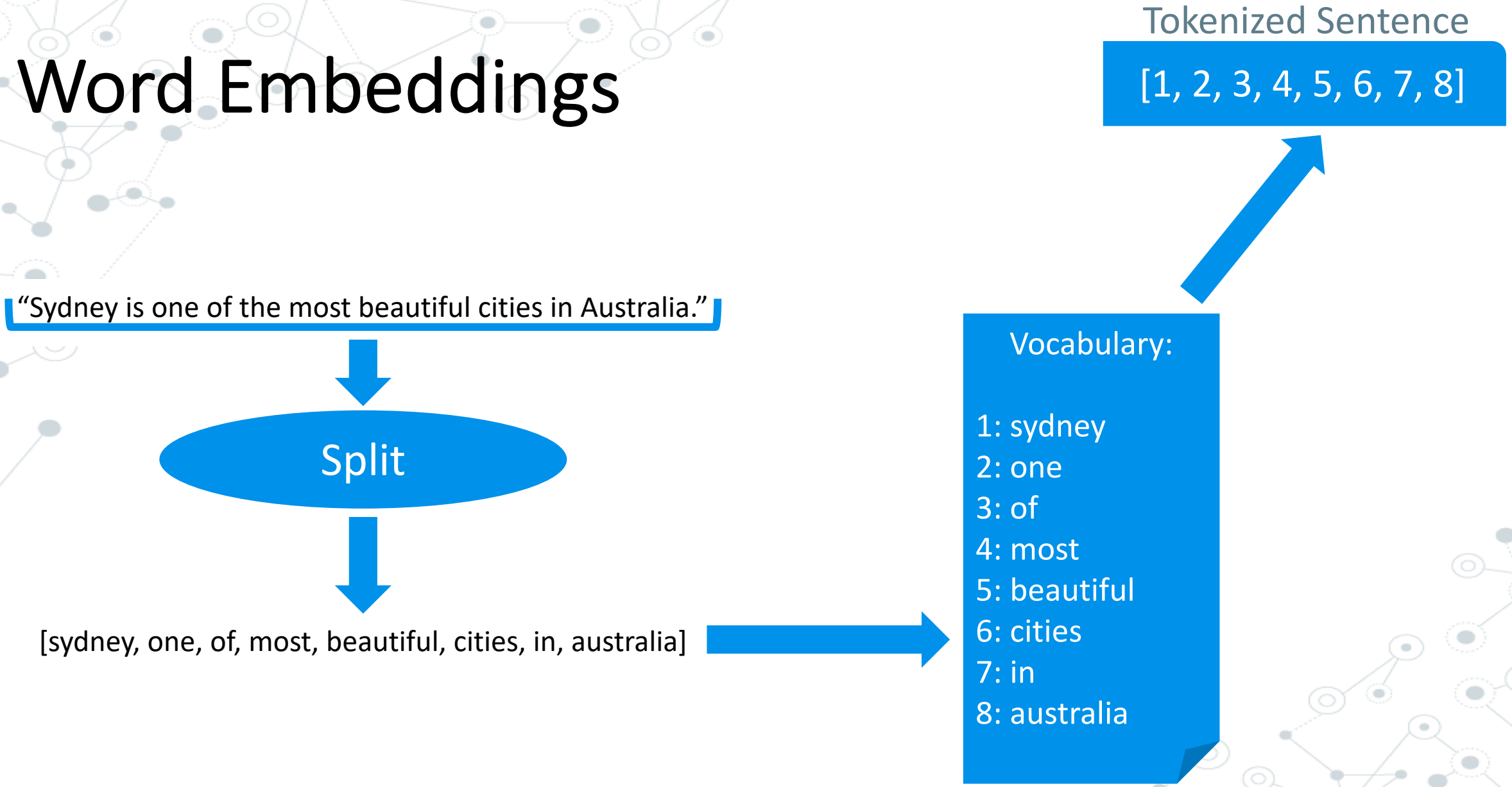
Methodology



Neural Networks



Word Embeddings



Word2Vec

[0, 0, 0, …, 1, 0, 0, …]

Vocabulary Length Vector

[-1.2, 0.23, …, -0.73, 0.1]

"Embedding Size" Length Vector

Embedding Matrix

1: sydney	[0.20, -0,23, 0.63]
2: one	[0.21, 0.10, -0.20]
3: of	[-0.22, 0.20, 0.15]
4: most	[0.81, 0.14, -0.25]
5: beautiful	[-0.81, 0.34, -0.50]
6: cities	[0.28, -0.26, 0.60]
7: in	[0.12, 0.53, 0.44]
8: australia	[0.35, -0.34, 0.52]

Tokenized Sentence

[1, 2, 3, 4, 5, 6, 7, 8]

Goal **Pred(t|c)**

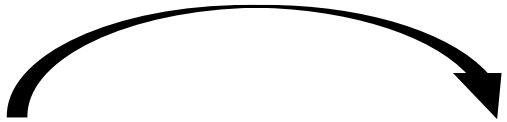
Pred(sydney|cities)

Pred(1|6)

Pred([0.20, -0,23, 0.63]|[0.28, -0.26, 0.60])

Word2Vec

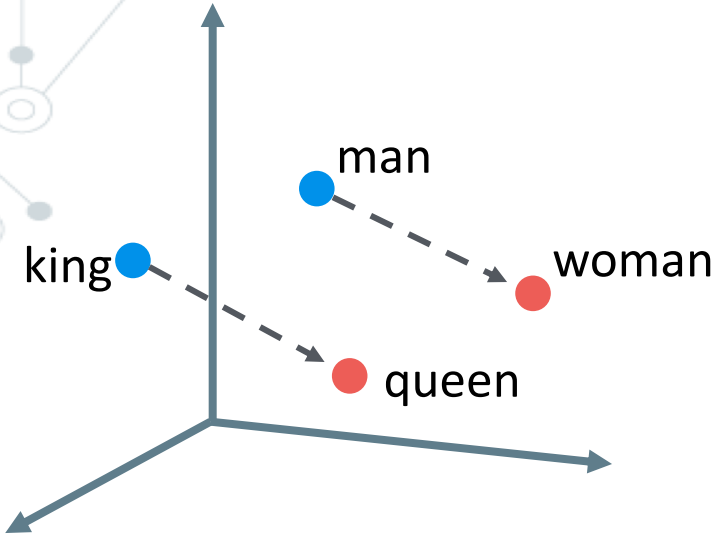
Vectors make us 



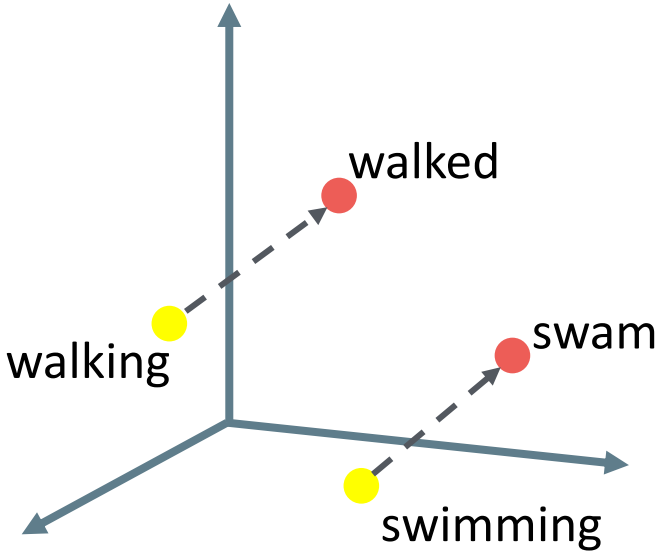
Similarity(“Sydney”, ”Australia”) VERY HIGH

Similarity(“Sydney”, ”Melbourne”) VERY HIGH

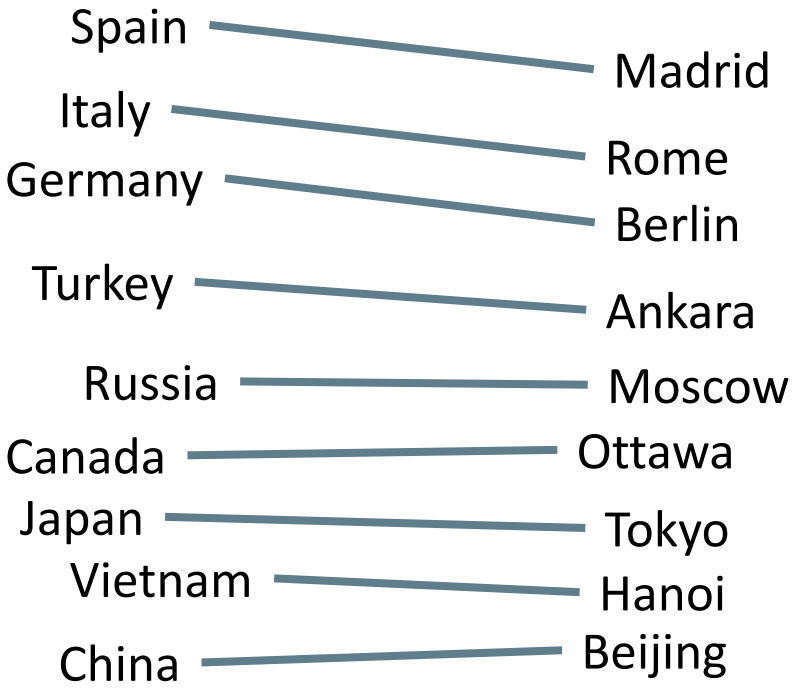
Word Embeddings



Male-Female



Verb tense



Country-Capital

$$v_{king} - v_{man} + v_{woman} \approx v_{queen}$$

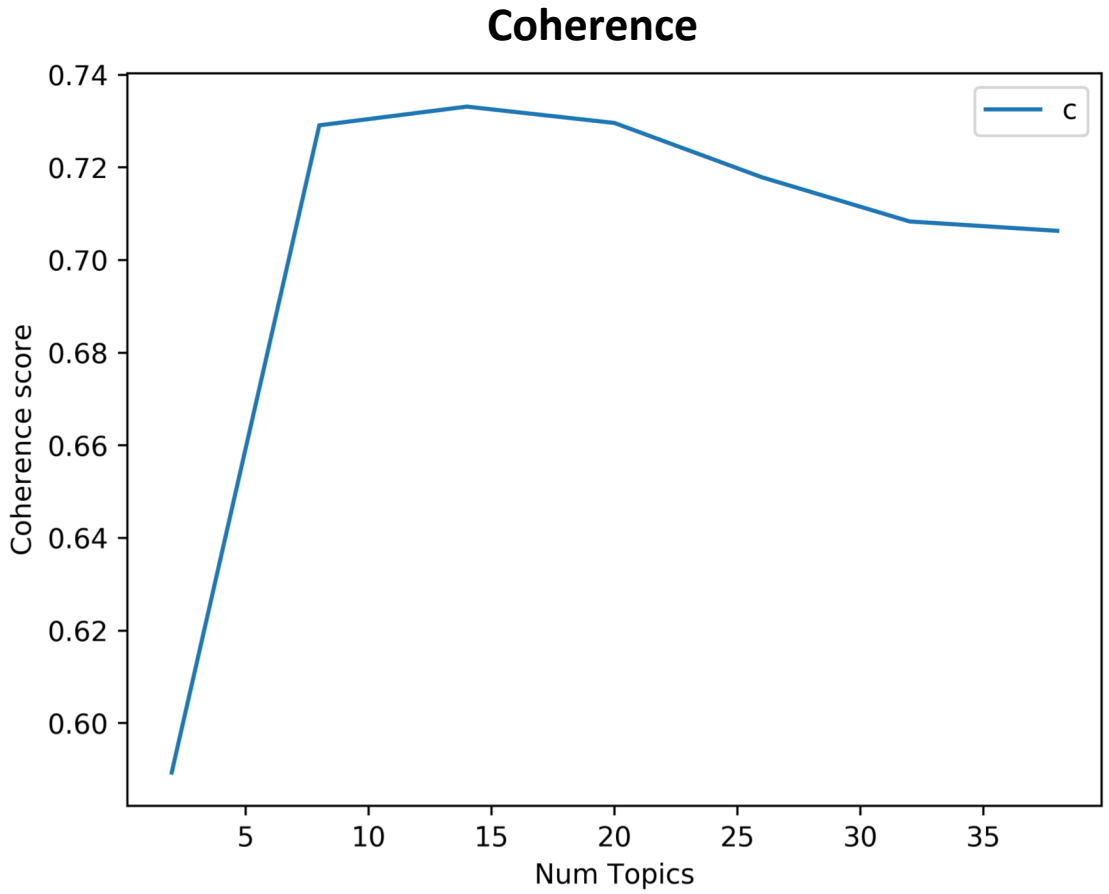
Topic Modelling with LDA

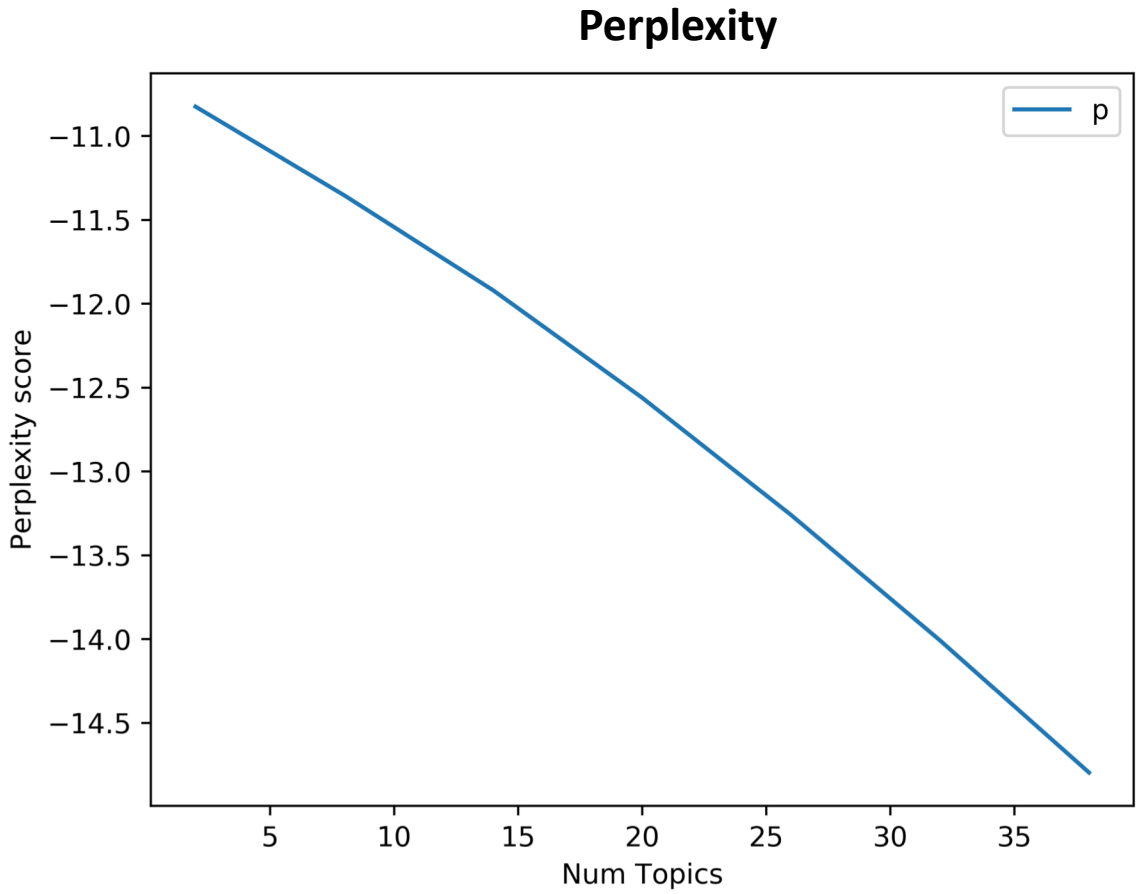
$$P(Z|W,D) = \frac{\#\ of\ word\ W\ in\ topic\ Z + \beta_w}{total\ tokens\ in\ Z + \beta} * (\#\ words\ in\ D\ that\ belong\ to\ Z + \alpha)$$

Topic Example	
0.028*"people"	0.010*"nature"
0.019*"book"	0.010*"knowledge"
0.014*"power"	0.010*"truth"
0.012*"fact"	0.009*"sense"
0.011*"idea"	0.009*"poet"

Results

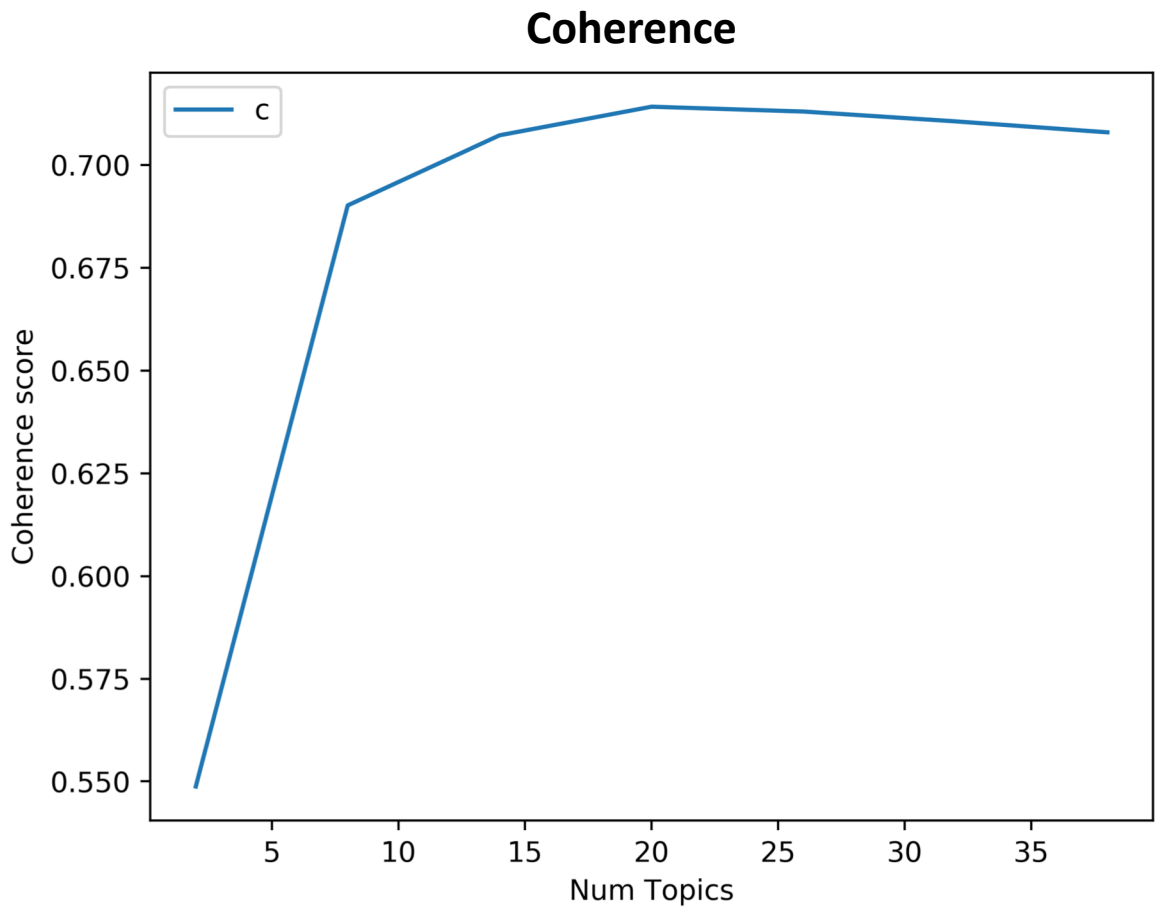
Deciding Number of Topics to Model with NOUNs

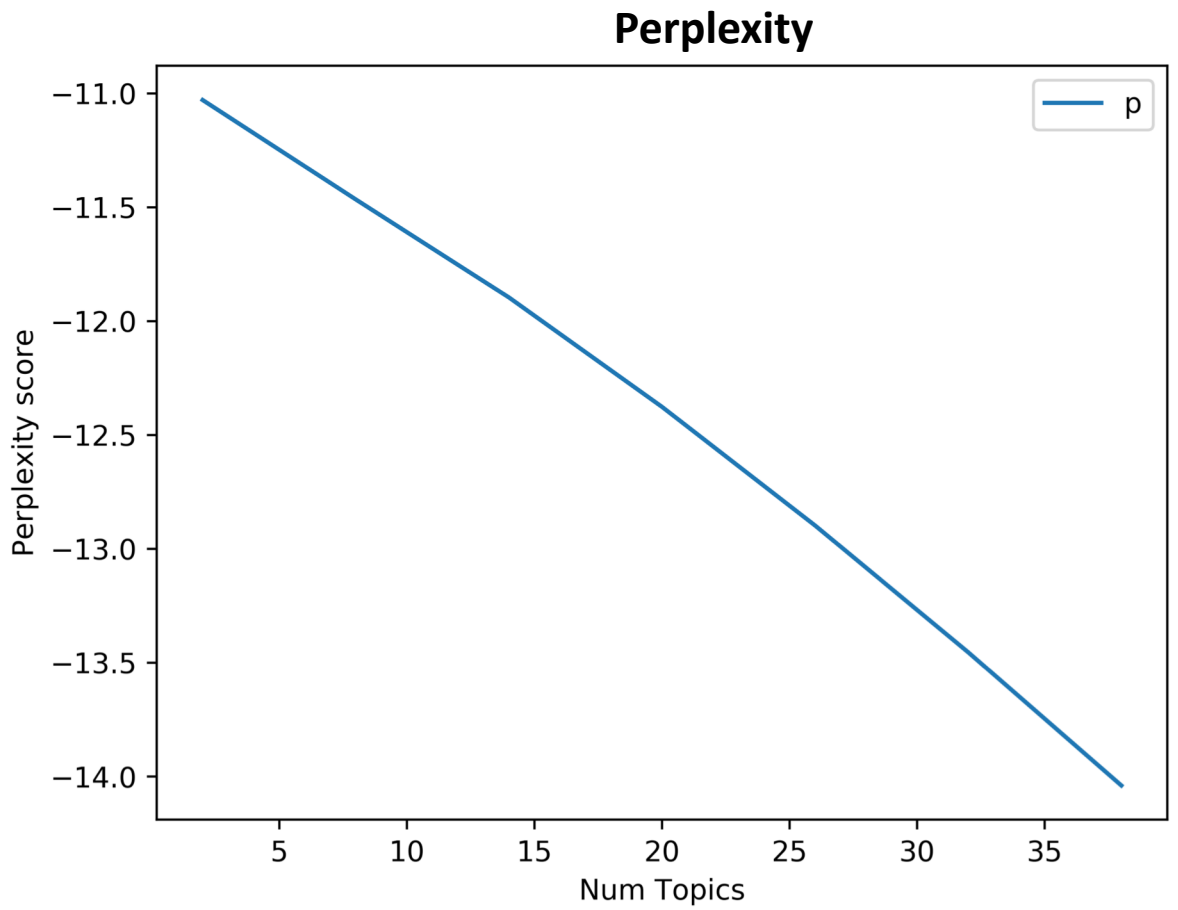




Results

Deciding Number of Topics to Model in NOUN, ADVERB, VERBs

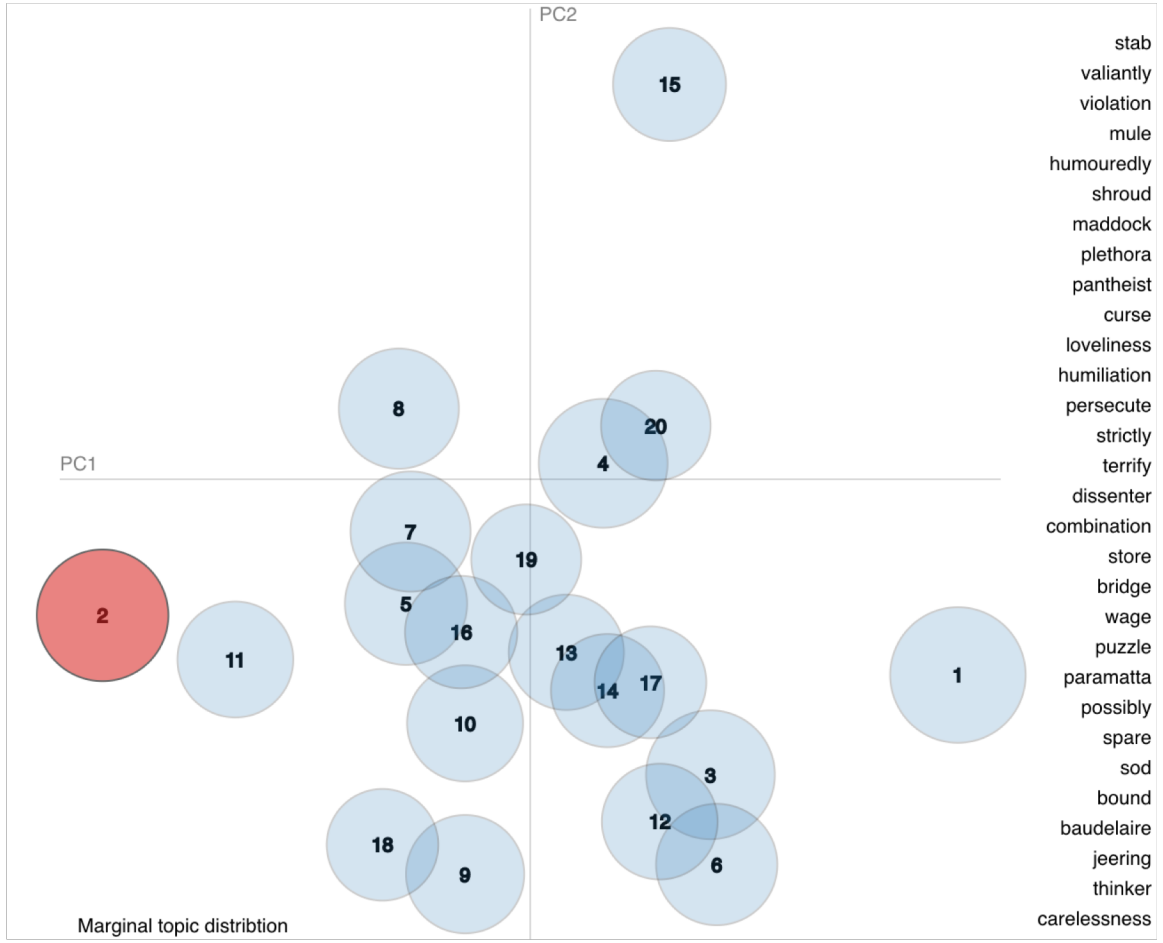




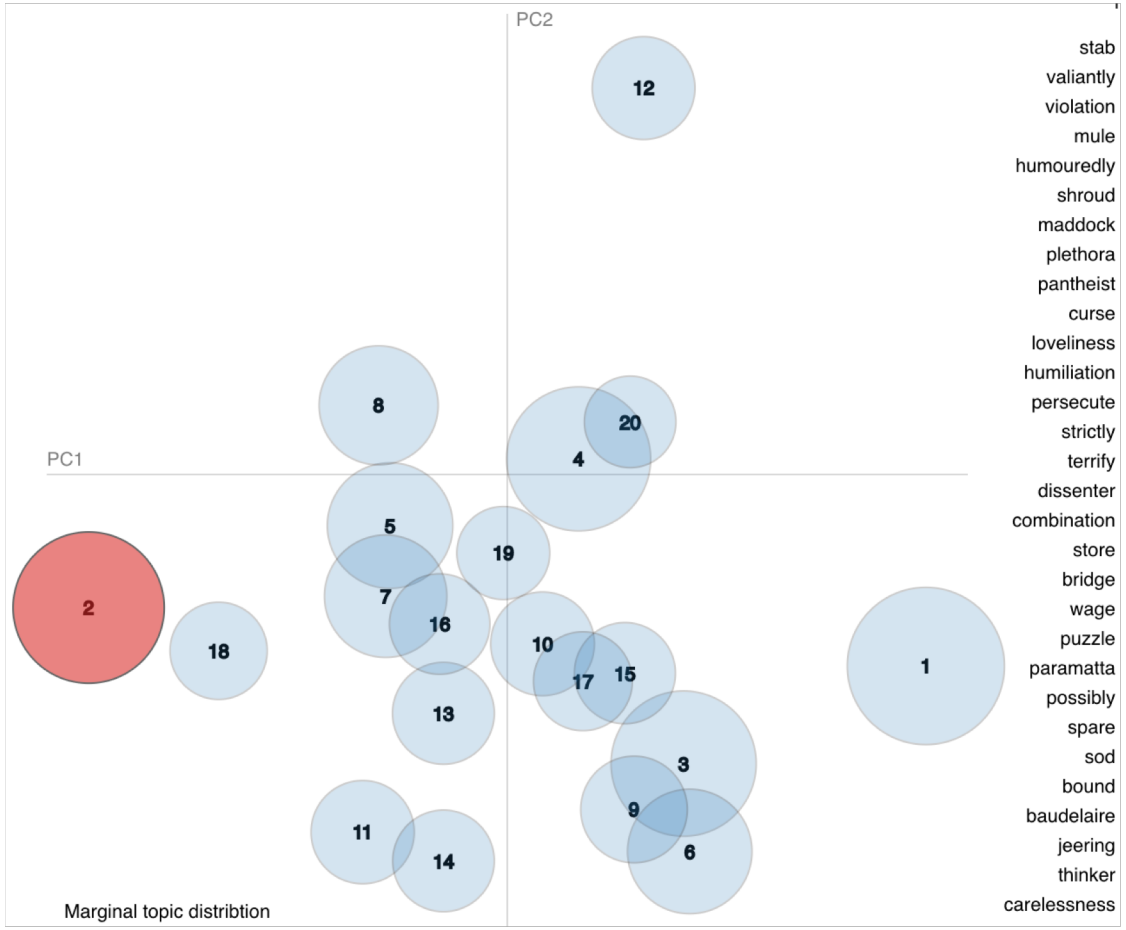
Results

Use LDA model to look-up a single document

Complete Corpus

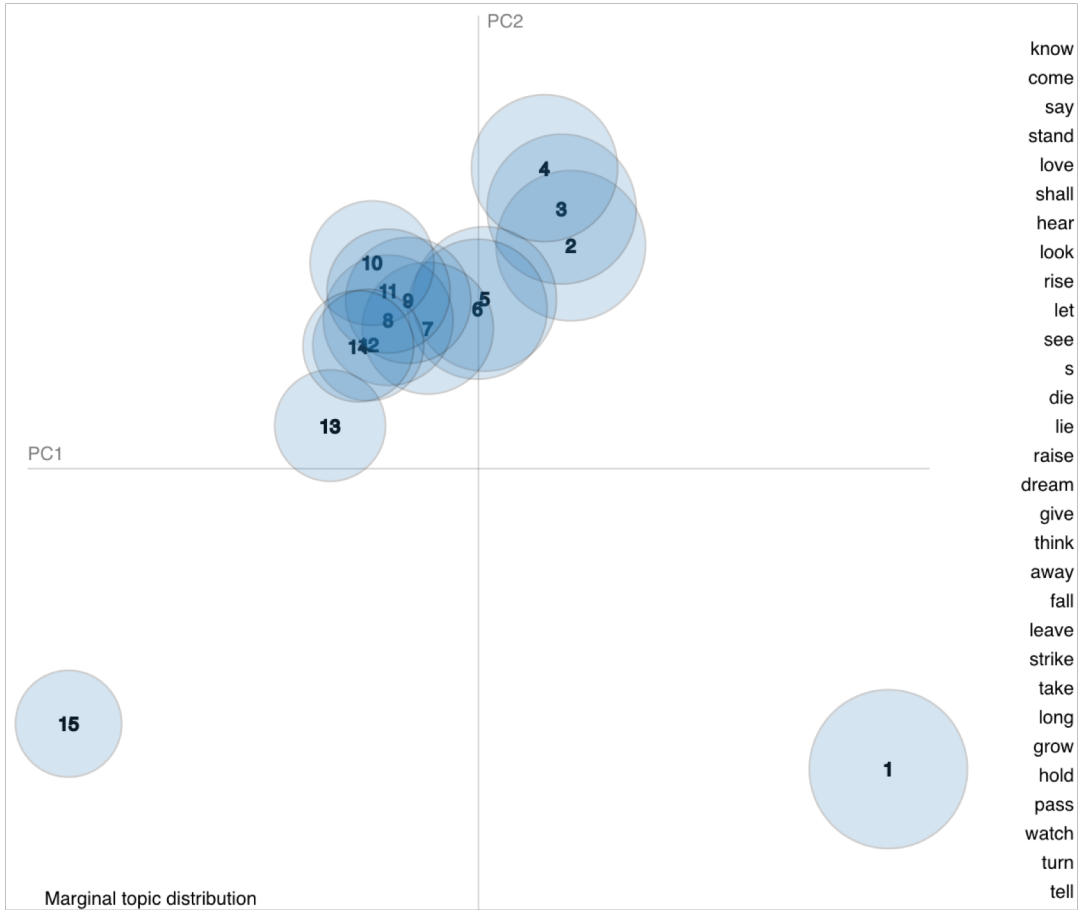


Document: Australian Essays



Why is MALLET more preferable?

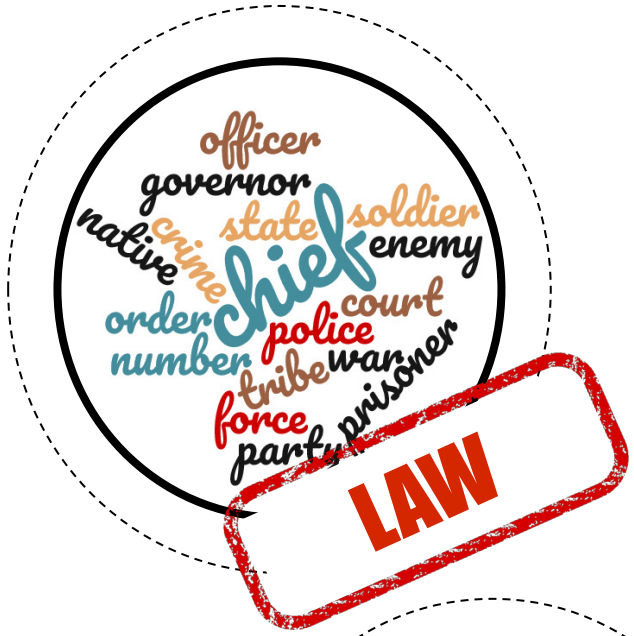
Only Nouns

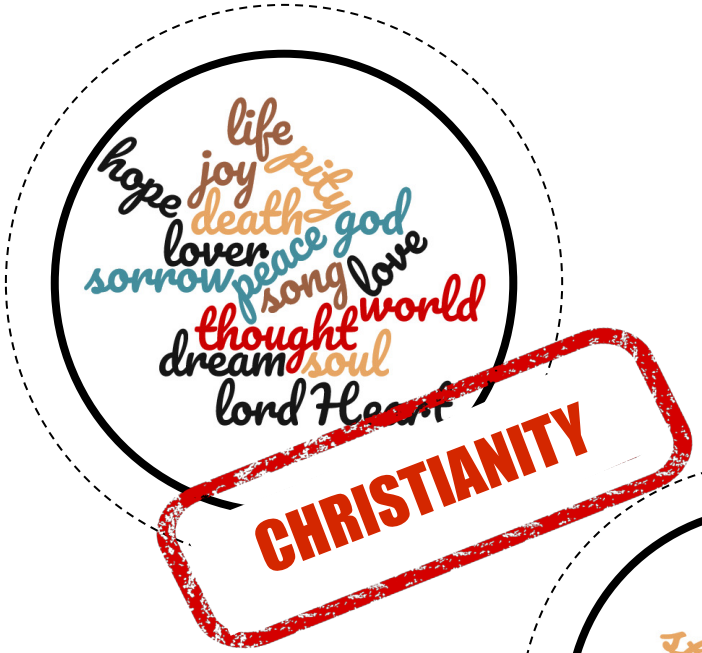


Verbs, Adverbs and Nouns

Results

Visualizing Topics with pyLDAvis





1 word2vec

2 LDA

3 lda2vec

Thank you!

Find our project and slides on GitHub.

https://github.com/tolgasaglik/distributional-semantics-project