

Dioni REBONATO ENDRINGER

Ema KEPUSKA

Elizabeth JOY

Kavita REGE

Tolga SAGLIK

distributional^{corpora}Semantics^{statistics}

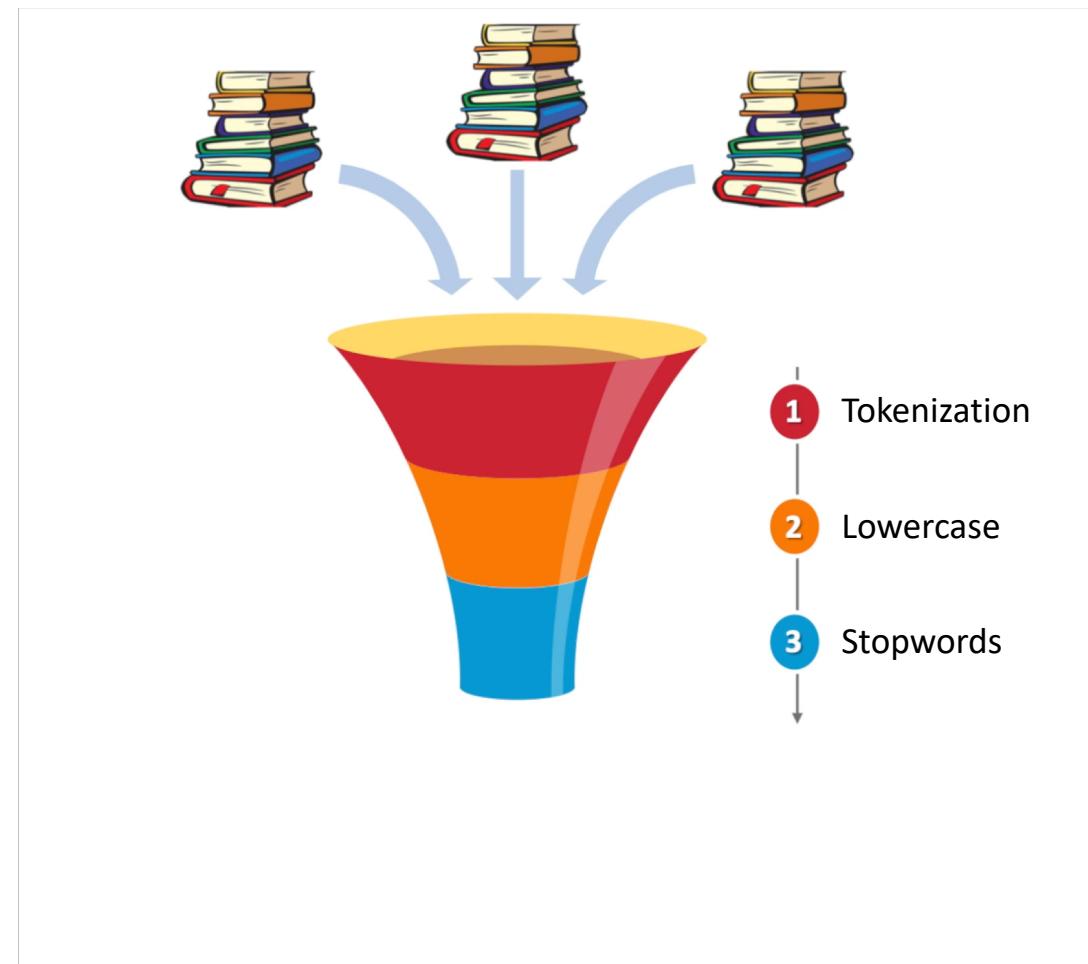
document corpus machine topic Modeling context
verb neural Network learning visualization sparse
adjective mallet prediction recall
data saliency sentence term

Advisor: Ekaterina Kamlovskaya

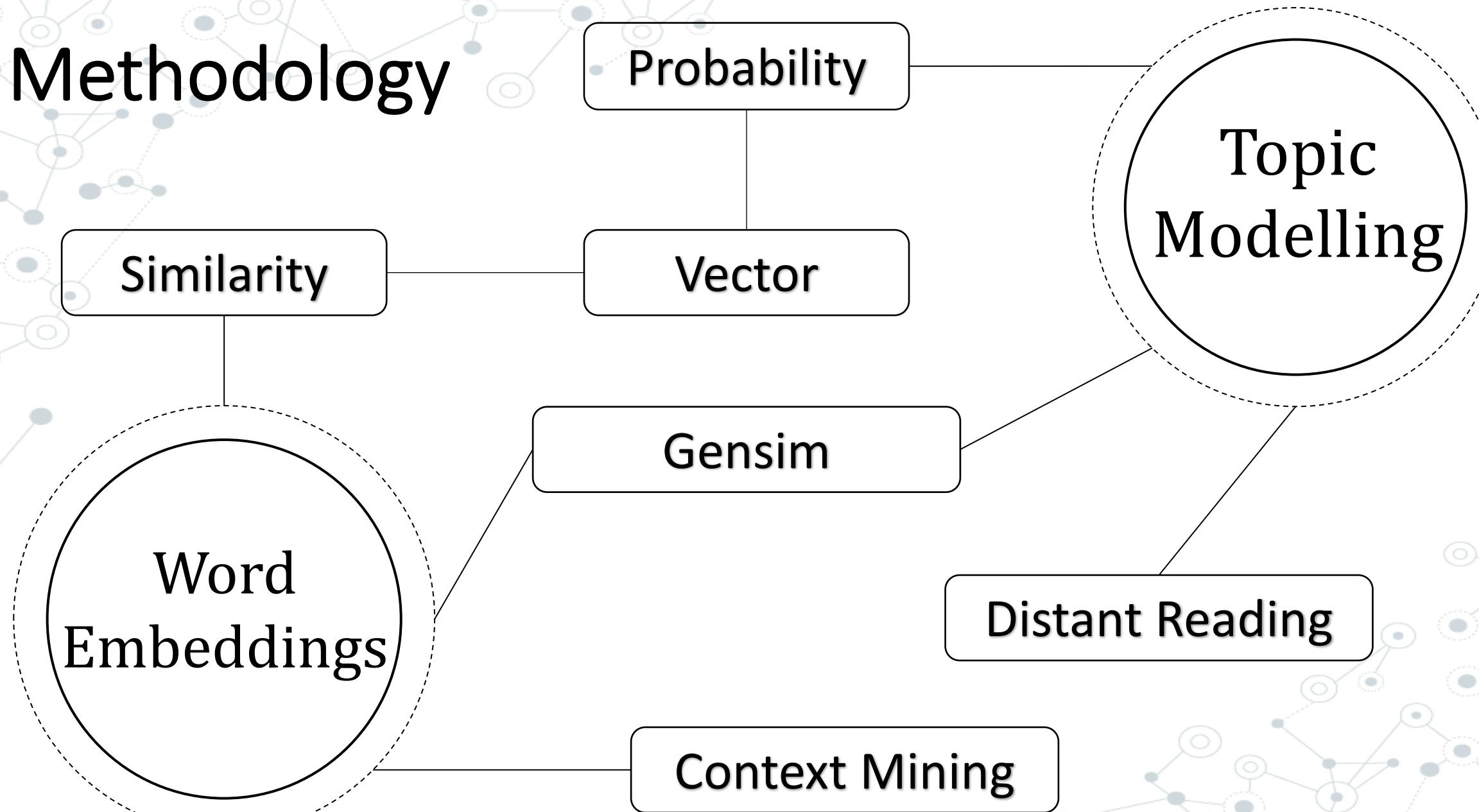
Motivation



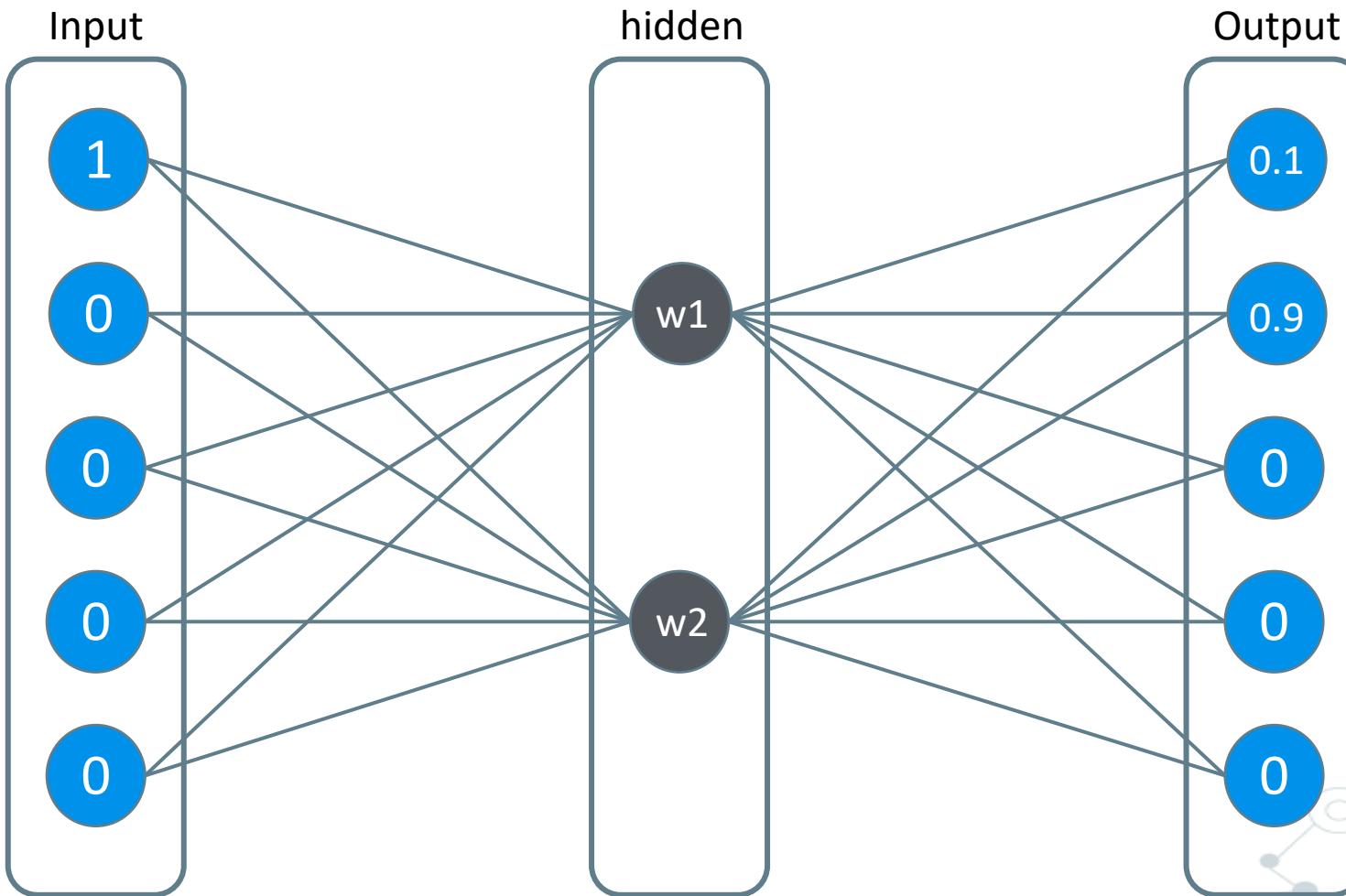
Data Preprocessing



Methodology



Neural Networks



Word Embeddings

“Sydney is one of the most beautiful cities in Australia.”

Split

[sydney, one, of, most, beautiful, cities, in, australia]

Vocabulary:

- 1: sydney
- 2: one
- 3: of
- 4: most
- 5: beautiful
- 6: cities
- 7: in
- 8: australia

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

Word2Vec

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 $\text{Pred}(t|c)$

$\text{Pred}(\text{sydney}|\text{cities})$

$\text{Pred}(1|6)$

$\text{Pred}([0.20, -0.23, 0.63] | [0.28, -0.26, 0.60])$

[0, 0, 0, ..., 1, 0, 0, ...]

[-1.2, 0.23, ..., -0.73, 0.1]

Vocabulary Length Vector

"Embedding Size" Length Vector

Word2Vec

Vectors make us

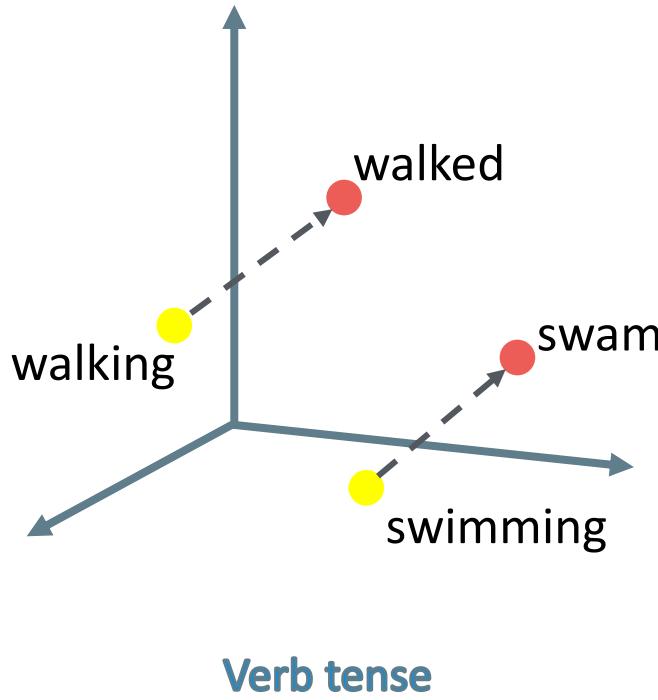
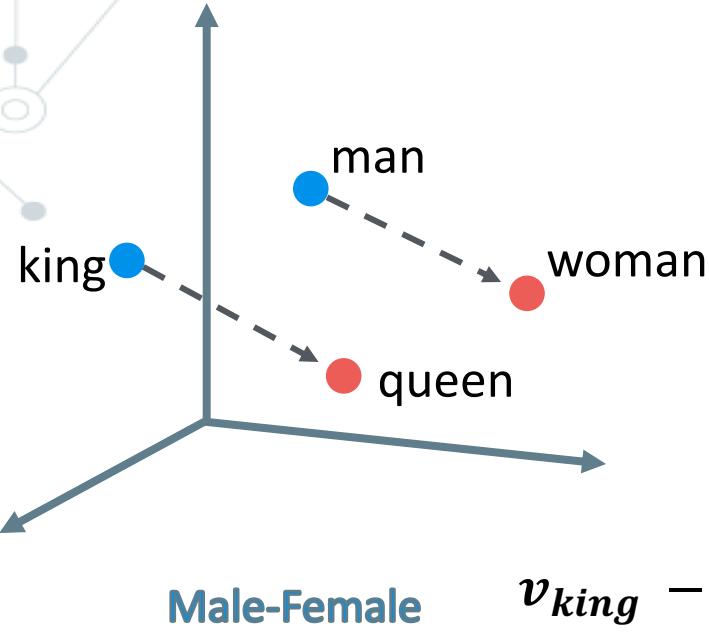


Similarity("Sydney", "Australia") **VERY HIGH**

Similarity("Sydney", "Melbourne") **VERY HIGH**



Word Embeddings



Spain	Madrid
Italy	Rome
Germany	Berlin
Turkey	Ankara
Russia	Moscow
Canada	Ottawa
Japan	Tokyo
Vietnam	Hanoi
China	Beijing

Country-Capital



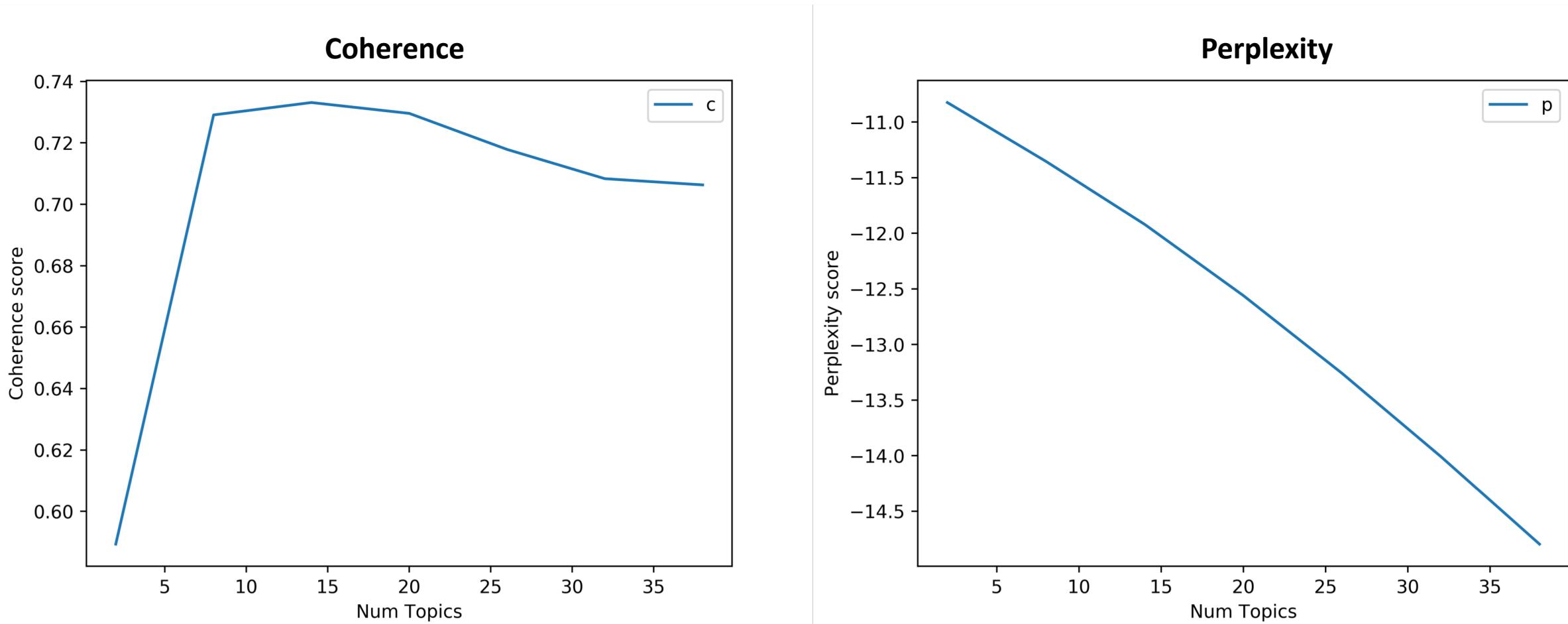
Topic Modelling with LDA

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

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

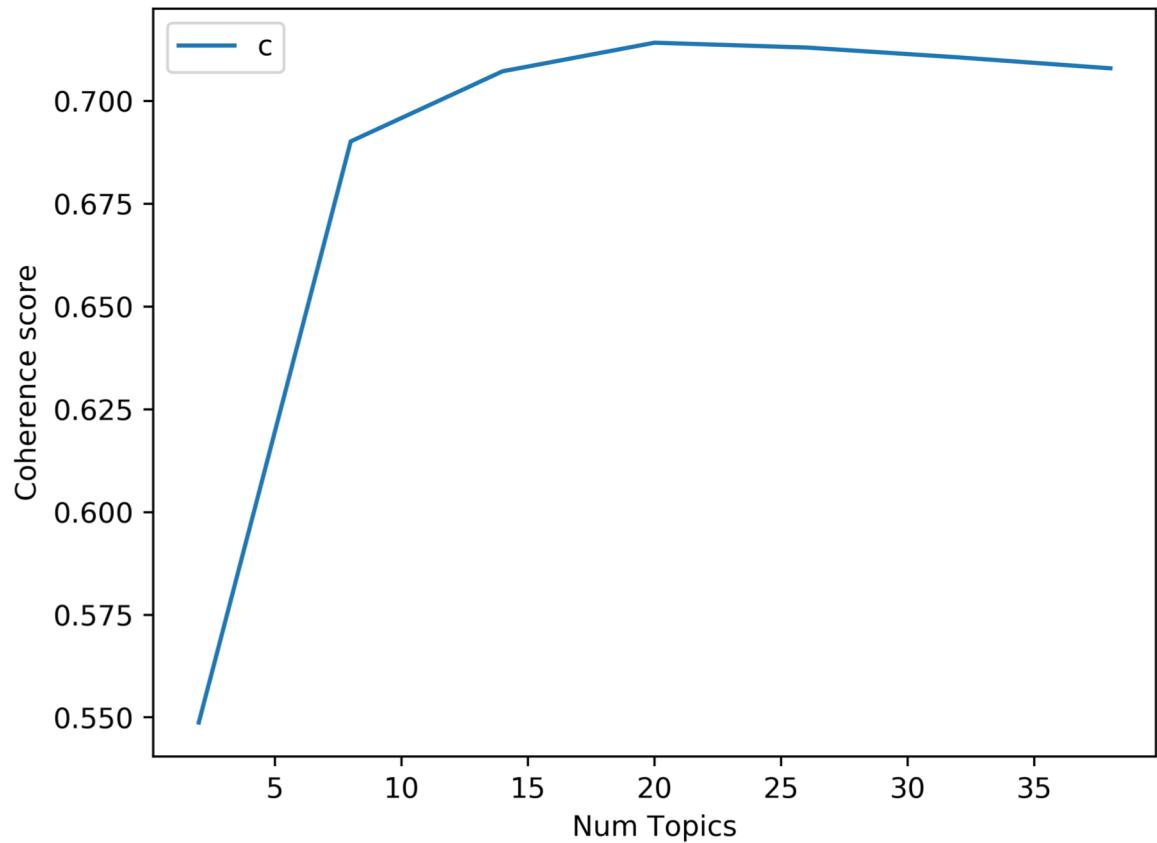
Deciding Number of Topics to Model with NOUNs



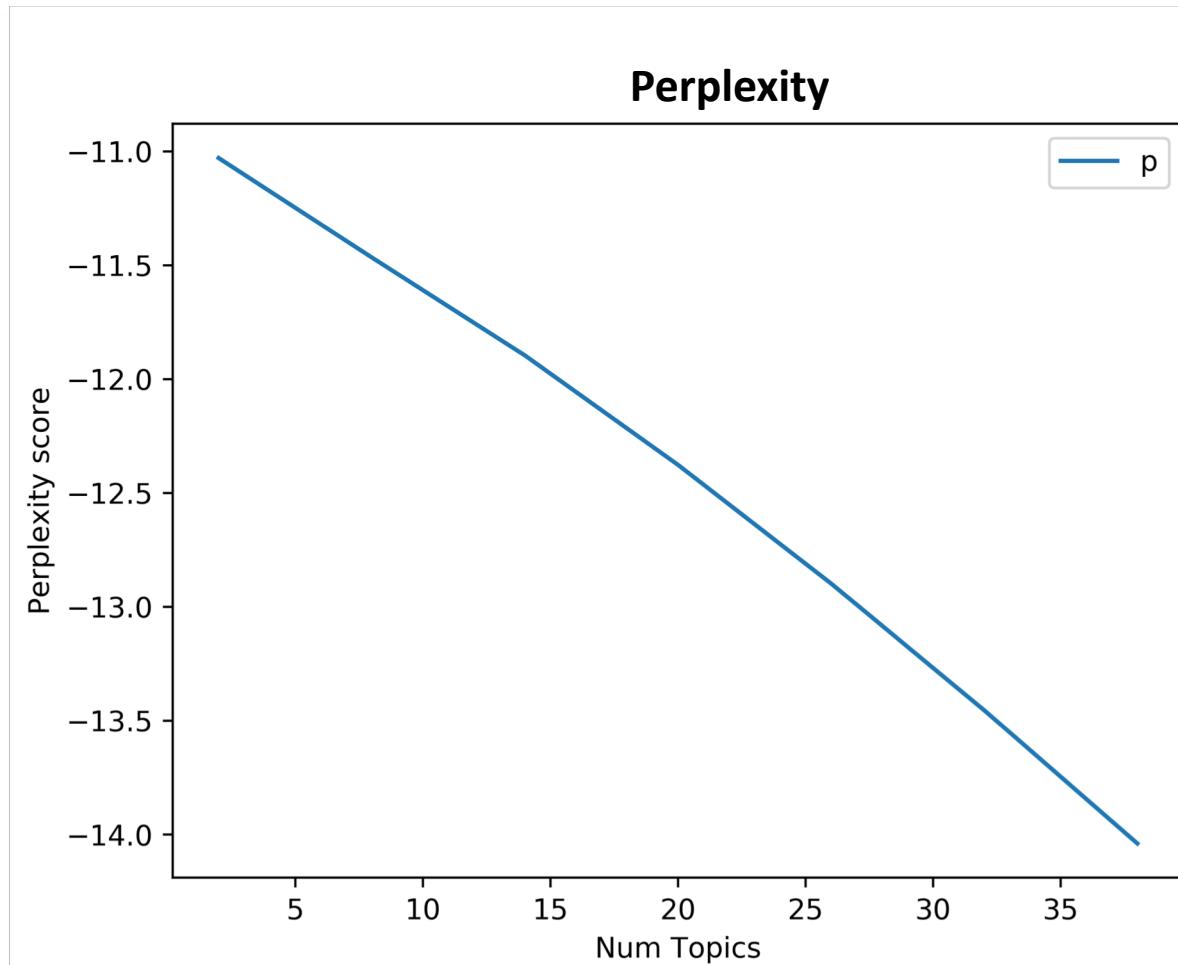
Results

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

Coherence



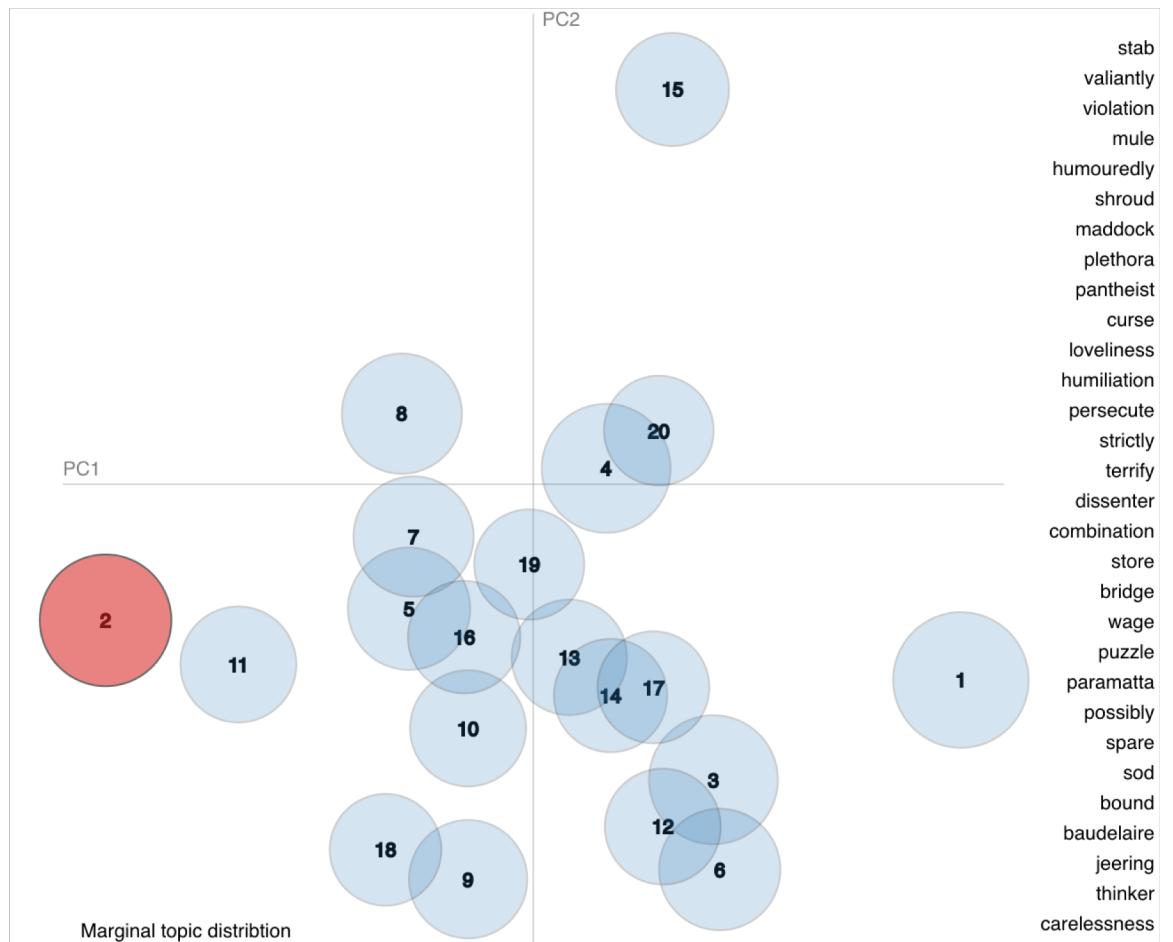
Perplexity



Results

Use LDA model to look-up a single document

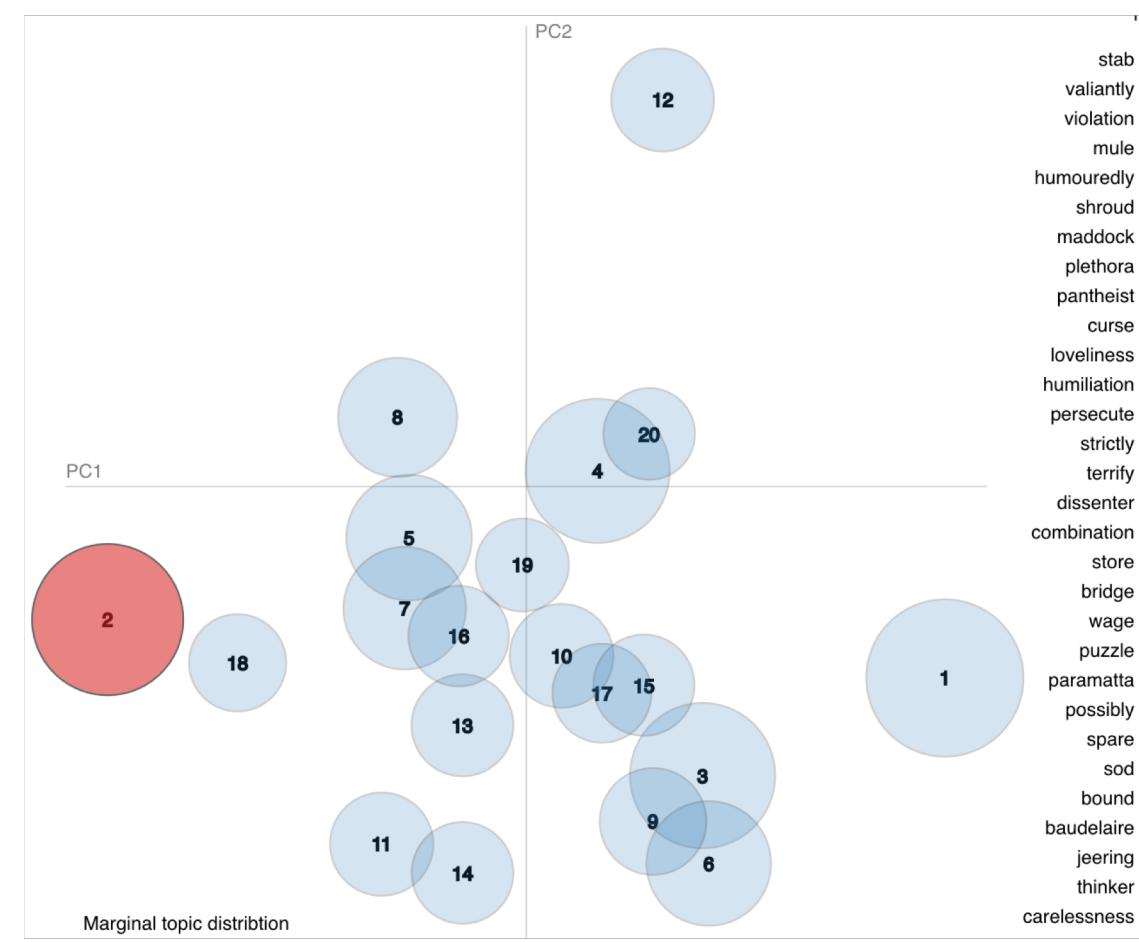
Complete Corpus



stab
valiantly
violation
mule
humouredly
shroud
maddock
plethora
pantheist
curse
loveliness
humiliation
persecute
strictly
terrify
dissenter
combination
store
bridge
wage
puzzle
paramatta
possibly
spare
sod
bound
baudelaire
jeering
thinker
carelessness

13

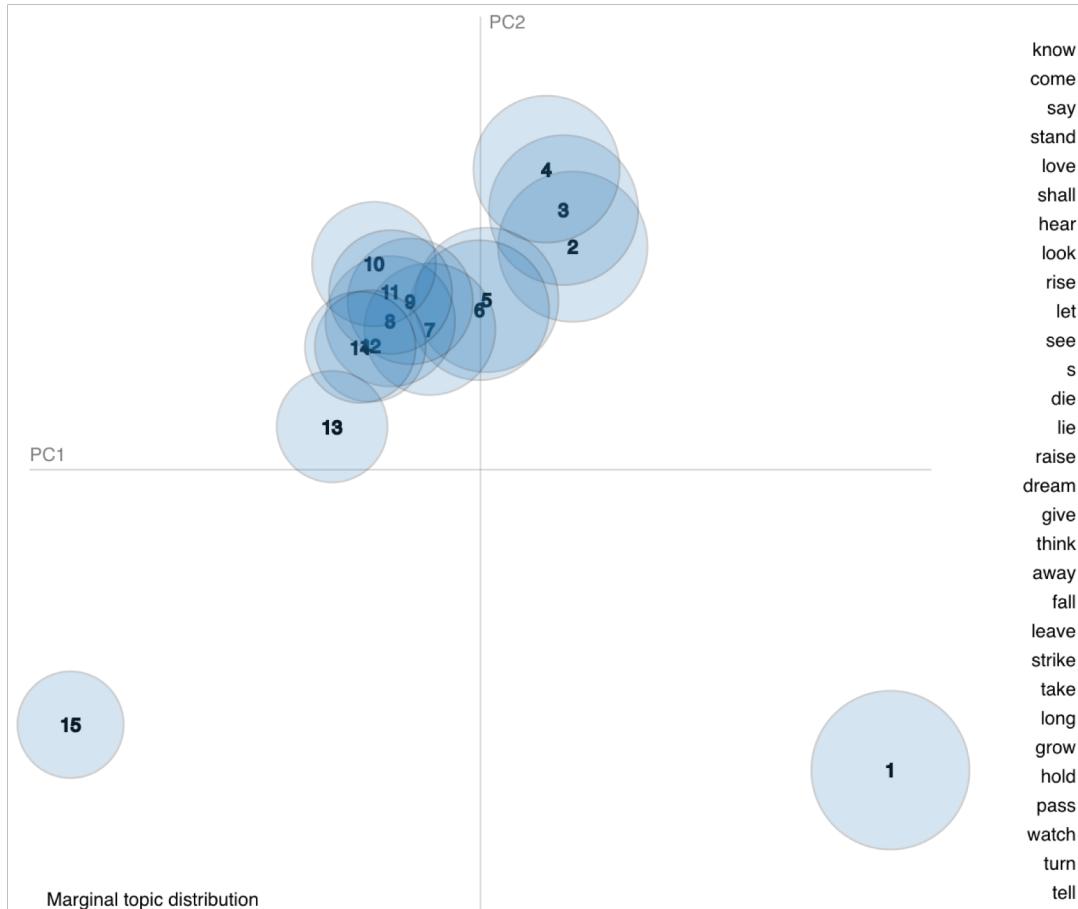
Document: Australian Essays



stab
valiantly
violation
mule
humouredly
shroud
maddock
plethora
pantheist
curse
loveliness
humiliation
persecute
strictly
terrify
dissenter
combination
store
bridge
wage
puzzle
paramatta
possibly
spare
sod
bound
baudelaire
jeering
thinker
carelessness

Why is MALLET more preferable?

Only Nouns



Verbs, Adverbs and Nouns



Results

Visualizing Topics with pyLDAvis

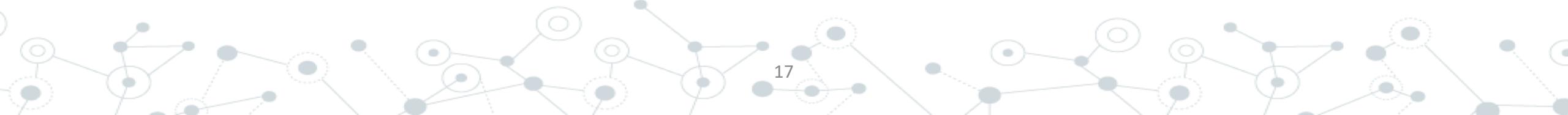


1 word2vec

2 LDA

3 lda2vec

Thank you!



Find our project and slides on GitHub.

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