10. Topic Models II: Beyond LDA (flipped)

DS-GA 1015, Text as Data
Arthur Spirling

April 20, 2021

Housekeeping

Housekeeping

1. HW3 is out. Usual academic honesty rules apply.

Housekeeping

1. HW3 is out. Usual academic honesty rules apply.

2. Back to usual this week: office hours on Wednesday, 1030–1130am.

Housekeeping

1. HW3 is out. Usual academic honesty rules apply.
2. Back to usual this week: office hours on Wednesday, 1030–1130am.

Extensions and Special Cases

Extensions and Special Cases

'vanilla' LDA is a popular model.

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.
But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.
But given topic A, we are more likely to see topic B than C. [CTM]

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).

But time matters for what topic 'means' [DTM]

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).

But time matters for what topic 'means' [DTM]

Assn no covariates

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts. . .

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).

But time matters for what topic 'means' [DTM]

Assn no covariates

But topic prevalence and topic content are $f(X)$ [STM]

Correlated Topic Model

Correlated Topic Model

Motolinia, "Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico" (2021).

Correlated Topic Model

Motolinia, "Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico" (2021). 6,890 legislative sessions in 20 Mexican states from 2012 to 2018:

Correlated Topic Model

Motolinia, "Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico" (2021). 6,890 legislative sessions in 20 Mexican states from 2012 to 2018: theory is that when reelection becomes possible, legislators will talk about particularistic (clientalistic) legislation more.

Correlated Topic Model

Motolinia, "Electoral Accountability and Particularistic Legislation: Evidence from an Electoral Reform in Mexico" (2021). 6,890 legislative sessions in 20 Mexican states from 2012 to 2018: theory is that when reelection becomes possible, legislators will talk about particularistic (clientalistic) legislation more.

CTM "CTM is able to model that a session discussing teachers is more likely to also discuss education than energy" Generally, CTM outperforms LDA (in holdout likelihood sense), and supports more topics.

Results

Results

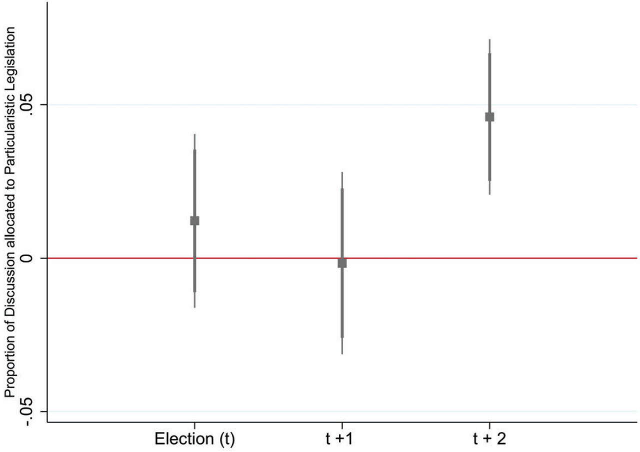
"Long-Term Reelection Incentives increase the proportion of discussion allocated to particularistic legislation by around 1.5 percentage points."

Results

"Long-Term Reelection Incentives increase the proportion of discussion allocated to particularistic legislation by around 1.5 percentage points." And more of an effect as an election gets closer:

Results

"Long-Term Reelection Incentives increase the proportion of discussion allocated to particularistic legislation by around 1.5 percentage points." And more of an effect as an election gets closer:



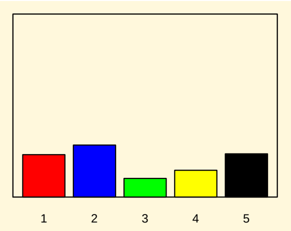
STM: Per Document Topic Distribution (θ)

STM: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.

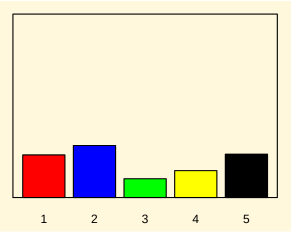
STM: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.



STM: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.



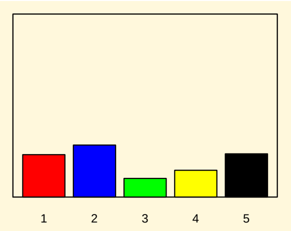
STM, that topic distribution ('prevalence') is a function of the document metadata.

STM: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.

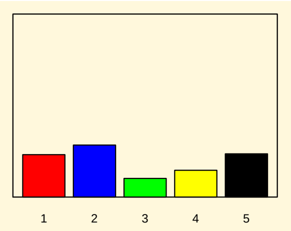
STM, that topic distribution ('prevalence') is a function of the document metadata.

e.g. perhaps male author ($X = 0$) documents have different topics relative to female ($X = 1$) author docs.



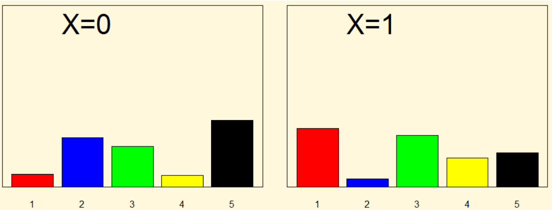
STM: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.



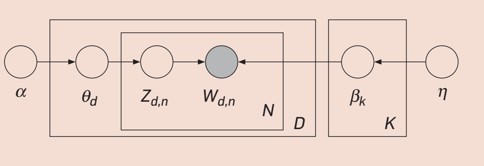
STM, that topic distribution ('prevalence') is a function of the document metadata.

e.g. perhaps male author ($X = 0$) documents have different topics relative to female ($X = 1$) author docs.

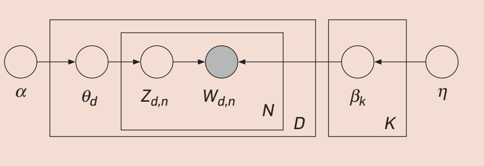


STM

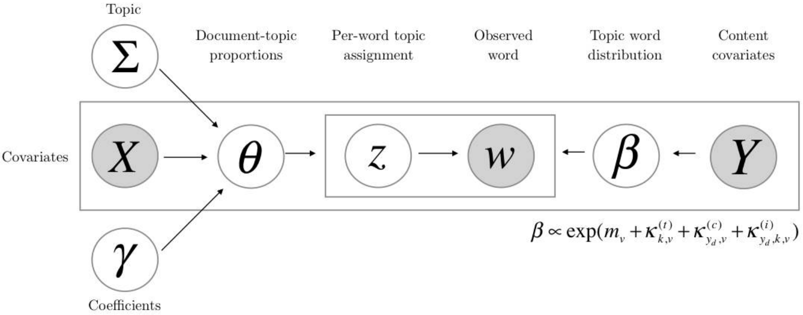
STM



STM



α
θ_d
$Z_{d,n}$
$W_{d,n}$
N
D
β_k
K
η



Topic
Document-topic proportions
Per-word topic assignment
Observed word
Topic word distribution
Content covariates
Covariates
Coefficients
$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$

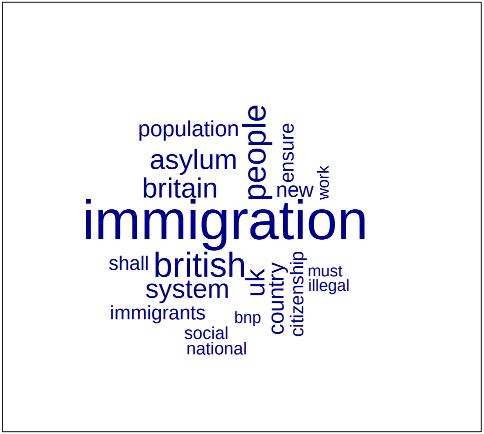
STM: Per Topic Word Distribution (β)

STM: Per Topic Word Distribution (β)

LDA: topic ('immigration') has a given distribution over words.

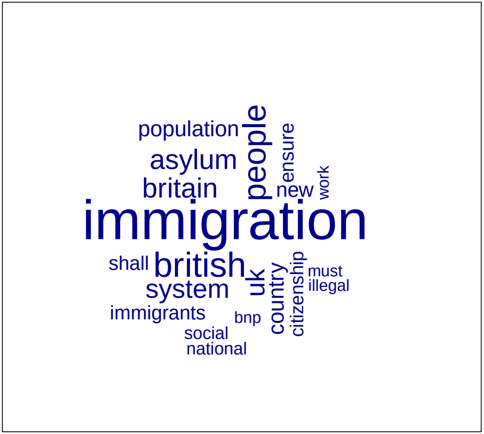
STM: Per Topic Word Distribution (β)

LDA: topic ('immigration') has a given distribution over words.



STM: Per Topic Word Distribution (β)

LDA: topic ('immigration') has a given distribution over words.



STM: that word distribution ('content') is a function of the document metadata.

STM: that word distribution ('content') is a function of the document metadata.

e.g. perhaps right parties ($Y = 0$) talk about a given topic differently to left ($Y = 1$) parties.

STM: that word distribution ('content') is a function of the document metadata.

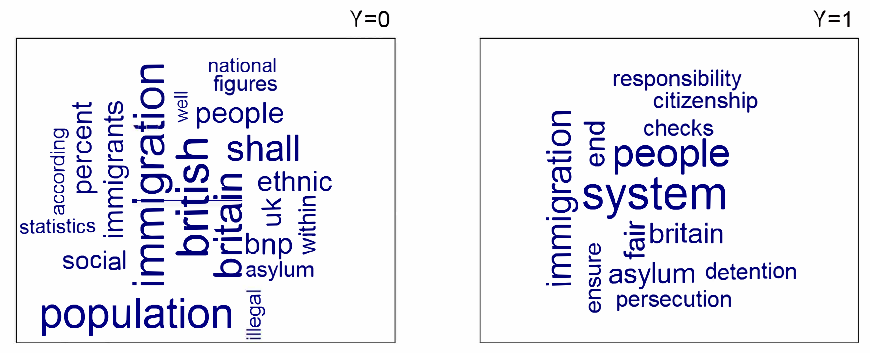
e.g. perhaps right parties ($Y = 0$) talk about a given topic differently to left ($Y = 1$) parties.

In practice, `content` needs to a single discrete variable.

STM: that word distribution ('content') is a function of the document metadata.

e.g. perhaps right parties ($Y = 0$) talk about a given topic differently to left ($Y = 1$) parties.

In practice, `content` needs to a single discrete variable.



In practice. . .

In practice...

"the model allows using topical prevalence covariates (θ),

In practice...

"the model allows using topical prevalence covariates (θ), a topical content covariate (β),

In practice...

"the model allows using topical prevalence covariates (θ), a topical content covariate (β), both,

In practice. . .

"the model allows using topical prevalence covariates (θ), a topical content covariate (β), both, or neither". If neither, it is CTM.

In practice. . .

"the model allows using topical prevalence covariates (θ), a topical content covariate (β), both, or neither". If neither, it is CTM.

NB: content "must be a single variable which defines a discrete partition of the dataset (each document is in one and only one group)".

In practice...

"the model allows using topical prevalence covariates (θ), a topical content covariate (β), both, or neither". If neither, it is CTM.

NB: content "must be a single variable which defines a discrete partition of the dataset (each document is in one and only one group)". In practice, this variable often appears in prevalence covariates too.

In practice...

"the model allows using topical prevalence covariates (θ), a topical content covariate (β), both, or neither". If neither, it is CTM.

NB: content "must be a single variable which defines a discrete partition of the dataset (each document is in one and only one group)". In practice, this variable often appears in prevalence covariates too.

$\rightarrow$ generally seems the case that people are more interested in prevalence than content effects.

“Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India’s Village Assemblies”

Parthasarathy et al model topics of *gram sahbas* (village governing council) meetings in India.

“Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India’s Village Assemblies”

Parthasarathy et al model topics of *gram sahbas* (village governing council) meetings in India. Want to estimate deliberative quality of these meetings, and who has advantages/disadvantages in terms of agenda setting etc.

"Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies"

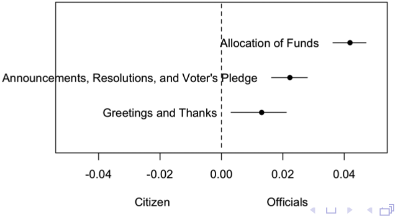
Parthasarathy et al model topics of *gram sahbas* (village governing council) meetings in India. Want to estimate deliberative quality of these meetings, and who has advantages/disadvantages in terms of agenda setting etc.

Use STM with topic prevalence a function of gender of the speaker + the position of the speaker + reservation status of the village council president (female and/or Scheduled Caste)

“Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India’s Village Assemblies”

Parthasarathy et al model topics of *gram sahbas* (village governing council) meetings in India. Want to estimate deliberative quality of these meetings, and who has advantages/disadvantages in terms of agenda setting etc.

Use STM with topic prevalence a function of gender of the speaker + the position of the speaker + reservation status of the village council president (female and/or Scheduled Caste)



Exercise

Exercise

1. In the dynamic topic model, any changes over time are ‘smooth’. Give an example of a time series problem where you want topics to change over time, but not in a smooth way.

Exercise

1. In the dynamic topic model, any changes over time are 'smooth'. Give an example of a time series problem where you want topics to change over time, but not in a smooth way.

2. In social science, we might be interested in the dynamic topic model because both its αs (feeds topic proportions) and the βs (feeds topic distribution) are allowed to change. What types of substantive problems do changes in these two different parameters help us model?

Exercise

1. In the dynamic topic model, any changes over time are ‘smooth’. Give an example of a time series problem where you want topics to change over time, but not in a smooth way.

2. In social science, we might be interested in the dynamic topic model because both its αs (feeds topic proportions) and the βs (feeds topic distribution) are allowed to change. What types of substantive problems do changes in these two different parameters help us model?

3. Are the ‘effects’ in the STM causal? If not, why not, and can you give a scenario where they would be?

Embeddings

Recap

Recap

What is the 'distributional hypothesis'?

Recap

What is the ‘distributional hypothesis’? What does it say about e.g. `cup` and `tea` and `coffee`?

Recap

What is the 'distributional hypothesis'? What does it say about e.g. `cup` and `tea` and `coffee`?

How can we operationalize a 'context'?

Recap

What is the 'distributional hypothesis'? What does it say about e.g. `cup` and `tea` and `coffee`?

How can we operationalize a 'context'?

What is a 'target' (word)?

Recap

What is the ‘distributional hypothesis’? What does it say about e.g. `cup` and `tea` and `coffee`?

How can we operationalize a ‘context’?

What is a ‘target’ (word)?

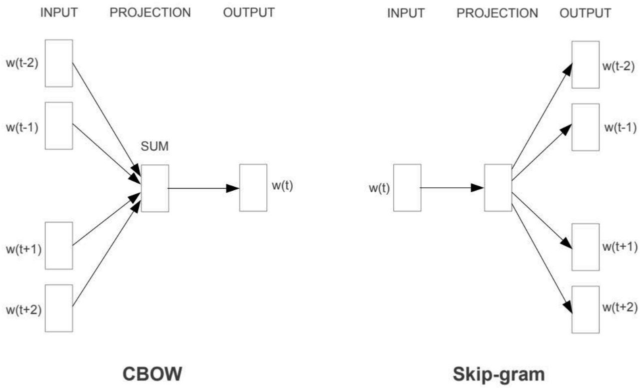
What is a ‘one hot encoding’?

word2vec architectures

What is the difference between CBOW and Skipgram in terms of inputs/outputs?

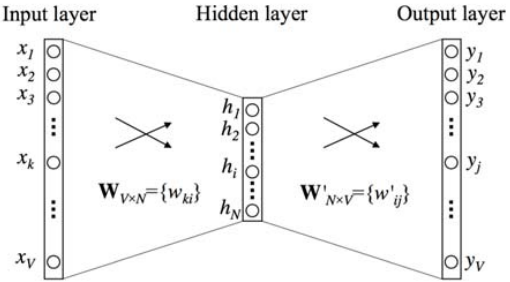
word2vec architectures

What is the difference between CBOW and Skipgram in terms of inputs/outputs?



CBOW Schematic

CBOW Schematic



Word Embeddings

Word Embeddings

What, literally, is an embedding for a word?

Word Embeddings

What, literally, is an embedding for a word?

A word embedding is a vector,

Word Embeddings

What, literally, is an embedding for a word?

A word embedding is a vector, and is a set of estimated weights from a neural network model.

Word Embeddings

What, literally, is an embedding for a word?

A word embedding is a vector, and is a set of estimated weights from a neural network model.

What is the neural network? what is it taking as an input? What is it trying to predict?

Word Embeddings

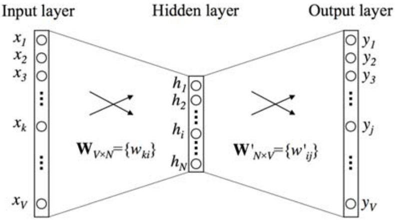
What, literally, is an embedding for a word?

A word embedding is a vector, and is a set of estimated weights from a neural network model.

What is the neural network? what is it taking as an input? What is it trying to predict?

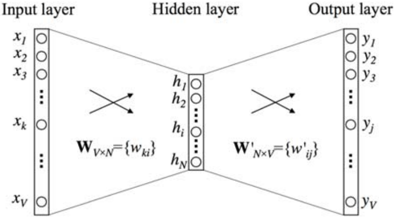
Window length, vector length

Window length, vector length



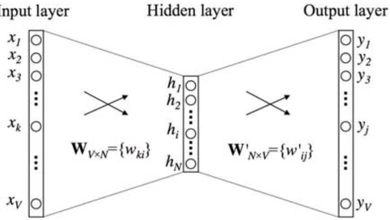
Word embeddings are usually expressed in terms of window length and length of the vector:

Window length, vector length



Word embeddings are usually expressed in terms of window length and length of the vector: e.g. 6-300 is window length of 6 (each side) and vector length of 300.

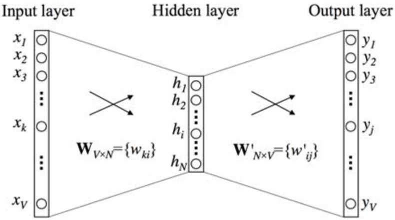
Window length, vector length



Word embeddings are usually expressed in terms of window length and length of the vector: e.g. 6-300 is window length of 6 (each side) and vector length of 300.

Q How do you think window length affects what we capture? Would you expect smaller or larger windows to capture syntactic relationships (e.g. have-having)? what about topical relationships (e.g. Biden-President)?

Window length, vector length

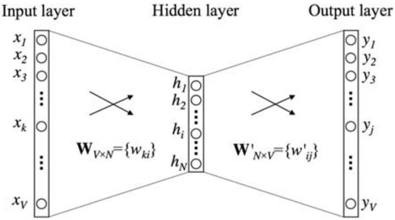


Word embeddings are usually expressed in terms of window length and length of the vector: e.g. 6-300 is window length of 6 (each side) and vector length of 300.

Q How do you think window length affects what we capture? Would you expect smaller or larger windows to capture syntactic relationships (e.g. have-having)? what about topical relationships (e.g. Biden-President)?

Q What does the length of the vector reflect?

Window length, vector length



Word embeddings are usually expressed in terms of window length and length of the vector: e.g. 6-300 is window length of 6 (each side) and vector length of 300.

Q How do you think window length affects what we capture? Would you expect smaller or larger windows to capture syntactic relationships (e.g. have-having)? what about topical relationships (e.g. Biden-President)?

Q What does the length of the vector reflect? How should that be chosen?

GloVe Embeddings

GloVe Embeddings

Word2vec uses local context windows: it goes word-to-word trying to predict the next word(s) from the surrounding ones.

GloVe Embeddings

Word2vec uses local context windows: it goes word-to-word trying to predict the next word(s) from the surrounding ones.

Pennington et al suggest that one should look at the co-occurances of the words in the documents directly.

GloVe Embeddings

Word2vec uses local context windows: it goes word-to-word trying to predict the next word(s) from the surrounding ones.

Pennington et al suggest that one should look at the co-occurances of the words in the documents directly.

Imagine creating a large matrix of the words $\times$ contexts in which they are found (other words), and then factorizing that matrix.

Some evidence GloVe is more stable and does better on some tasks.

Training Embeddings

Training Embeddings

Embeddings need quite a lot of text to train: e.g. want to disambiguate meanings from contexts.

Training Embeddings

Embeddings need quite a lot of text to train: e.g. want to disambiguate meanings from contexts. You can download pre-trained,

Training Embeddings

Embeddings need quite a lot of text to train: e.g. want to disambiguate meanings from contexts. You can download pre-trained, or get the code and train locally.

Training Embeddings

Embeddings need quite a lot of text to train: e.g. want to disambiguate meanings from contexts. You can download pre-trained, or get the code and train locally.

Word2Vec is trained on the Google News dataset ($\sim$ 100B words, 2013)

Training Embeddings

Embeddings need quite a lot of text to train: e.g. want to disambiguate meanings from contexts. You can download pre-trained, or get the code and train locally.

Word2Vec is trained on the Google News dataset ($\sim$ 100B words, 2013)

Different versions of GloVe are trained on different things: Wikipedia (2014) + Gigaword (6B words), Common Crawl, Twitter

Training Embeddings

Embeddings need quite a lot of text to train: e.g. want to disambiguate meanings from contexts. You can download pre-trained, or get the code and train locally.

Word2Vec is trained on the Google News dataset ($\sim$ 100B words, 2013)

Different versions of GloVe are trained on different things: Wikipedia (2014) + Gigaword (6B words), Common Crawl, Twitter

Various other pre-trained versions out there, and you can fit your own (locally).

Questions

Questions

1 When would we use pre-trained, when would we use locally trained ? What could go wrong if we use the 'wrong' one?

Questions

1 When would we use pre-trained, when would we use locally trained ? What could go wrong if we use the 'wrong' one?

2 The downloadable embeddings for both W2V and GloVe were trained in 2013/4. Does this matter?

Can you give an example of a word that had a specific meaning then but a different one before or after that time?

Analogies

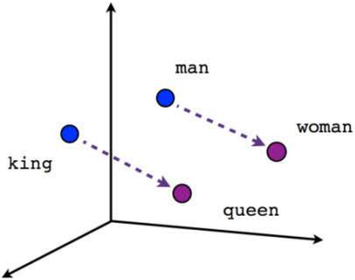
Analogies

These real valued vectors can be used for analogies and related tasks,

Analogies

These real valued vectors can be used for analogies and related tasks,

$$v_{\text{queen}} - v_{\text{woman}} + v_{\text{men}} \approx v_{\text{king}}$$



Male-Female

Cultural Bias: Caliskan et al, 2017

Cultural Bias: Caliskan et al, 2017

Drawing on idea of Implicit Association Test ‘reaction times’,

Cultural Bias: Caliskan et al, 2017

Drawing on idea of Implicit Association Test ‘reaction times’, authors consider (cosine) distance between embeddings of words.

Cultural Bias: Caliskan et al, 2017

Drawing on idea of Implicit Association Test 'reaction times', authors consider (cosine) distance between embeddings of words. Use 'off-the-shelf' GloVe.

Cultural Bias: Caliskan et al, 2017

Drawing on idea of Implicit Association Test 'reaction times', authors consider (cosine) distance between embeddings of words. Use 'off-the-shelf' GloVe.

Idea is that e.g. if (stereotypically) European American names are closer to pleasant terms than (stereotypically) African American names, this implies culture has in-built biases.

Cultural Bias: Caliskan et al, 2017

Drawing on idea of Implicit Association Test ‘reaction times’, authors consider (cosine) distance between embeddings of words. Use ‘off-the-shelf’ GloVe.

Idea is that e.g. if (stereotypically) European American names are closer to pleasant terms than (stereotypically) African American names, this implies culture has in-built biases.

Many experiments here: insects v fruits, musical instruments v weapons, gender v family/career.

Cultural Bias: Caliskan et al, 2017

Drawing on idea of Implicit Association Test 'reaction times', authors consider (cosine) distance between embeddings of words. Use 'off-the-shelf' GloVe.

Idea is that e.g. if (stereotypically) European American names are closer to pleasant terms than (stereotypically) African American names, this implies culture has in-built biases.

Many experiments here: insects v fruits, musical instruments v weapons, gender v family/career. Essentially all as expected:

Cultural Bias: Caliskan et al, 2017

Drawing on idea of Implicit Association Test 'reaction times', authors consider (cosine) distance between embeddings of words. Use 'off-the-shelf' GloVe.

Idea is that e.g. if (stereotypically) European American names are closer to pleasant terms than (stereotypically) African American names, this implies culture has in-built biases.

Many experiments here: insects v fruits, musical instruments v weapons, gender v family/career. Essentially all as expected: statistically significant differences in associations.

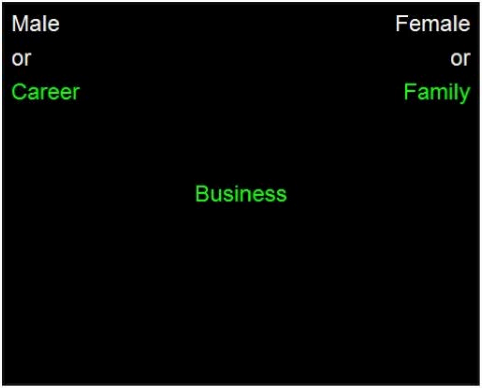
Implicit Association Test

Implicit Association Test

Basic idea is to compare speed of association (via keystroke), depending on what word is paired with what gender/race category.

Implicit Association Test

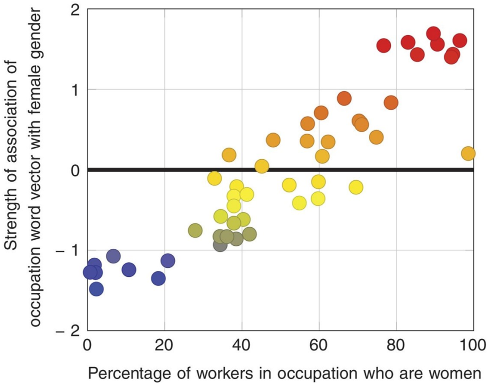
Basic idea is to compare speed of association (via keystroke), depending on what word is paired with what gender/race category.



Matching to Real World Data

Matching to Real World Data

Can use word embeddings distances to predict real-world participation in various occupations.



Exercise

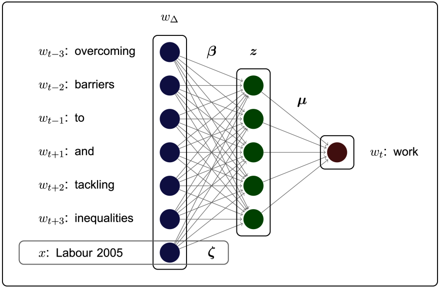
Exercise

1. How do we know whether a word embedding vector is a good representation or not? How could we test the merits of one particular model versus another?

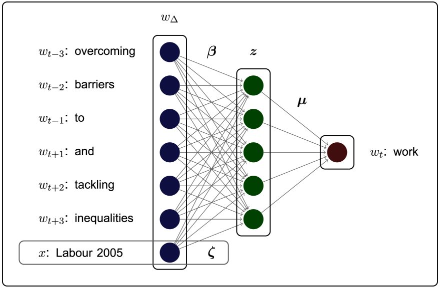
2. Embeddings reflect cultural biases. What does this mean for our work? What should we do about this?

Using metadata: Rheault and Cochrane, 2020

Using metadata: Rheault and Cochrane, 2020

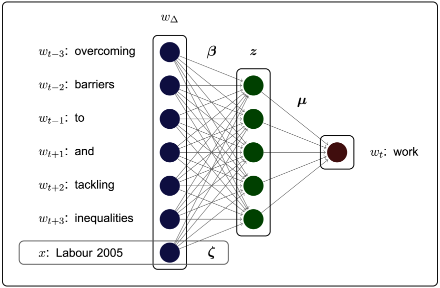


Using metadata: Rheault and Cochrane, 2020



We can add covariates into our embeddings.

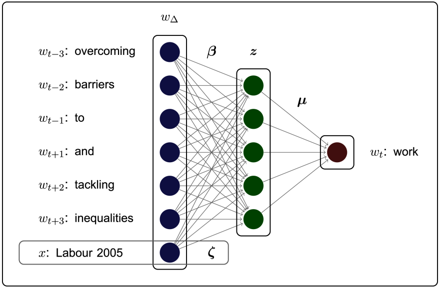
Using metadata: Rheault and Cochrane, 2020



We can add covariates into our embeddings.

Here, ζ is an indicator for party-year. And is of the same dimensions as the embeddings themselves.

Using metadata: Rheault and Cochrane, 2020

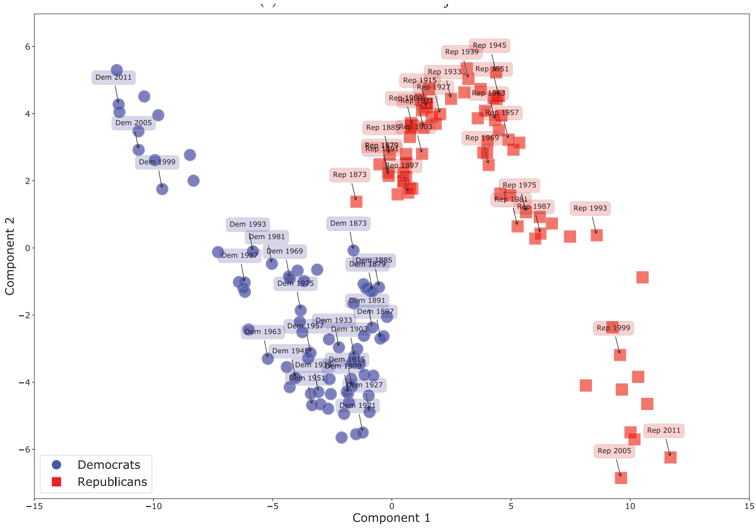


We can add covariates into our embeddings.

Here, ζ is an indicator for party-year. And is of the same dimensions as the embeddings themselves.

We can do a data reduction (e.g. PCA) on that matrix of ζs and then e.g. plot its dimensions. . .

Results for US



Results for UK