Effective Text Classification with word2vec

personagraph

an Intertrust Company

galvanize

Outline

1. Text Classification
   a. The basic problem and standards approaches
   b. Challenges of text classification
2. Deep learning:
   a. Auto-encoding as a signal compressor
3. Doc2Vec as a feature space generator:
   a. What is Word2Vec
   b. Word vectors to Doc2Vec as feature engineering
4. Benchmarking under Label Sparsity and imbalance
   a. Supervised Learning: Document Vectors vs. BOW features
   b. OOS improvement with Doc2Vec engineering under imbalance
5. Conclusions

Document Classification (Supervised)

Preprocessing

Stemming
Capitalization
Punctuation

Feature Selection

Regularization
Feature Importance
Correlation Modeling

Feature Engineering

PCA
Sparsity Filtering
TF-IDF
Transformations/Scaling
BOW
Context Embedding
Auto Encoding

Model Training

Classifiers:
Random Forest
Logistic Regression
Naive Bayes
...
ANN

Document Classification (Unsupervised)

Preprocessing

Stemming
Capitalization
Punctuation

Feature Engineering

PCA
Sparsity Filtering
TF-IDF
Transformations/Scaling
BOW
Context Embedding
Auto Encoding

Unsupervised Training

LDA
Document Clustering
...
Doc2vec

Document Classification

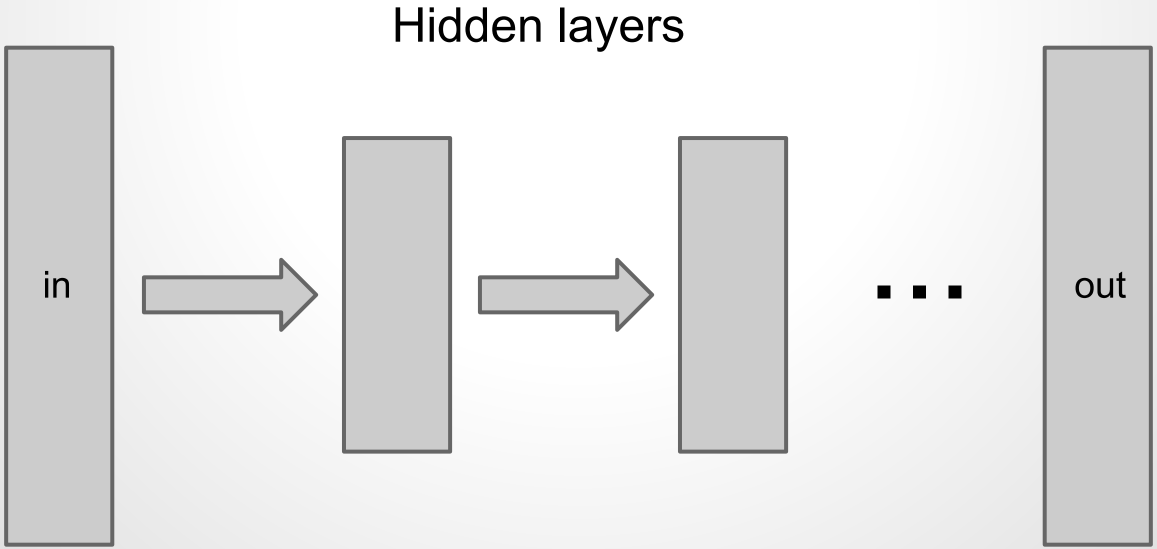
Data Challenges:

1. Data Quality
   a. Data Shape: Feature count ! >> Data count
      i. Curse of dimensionality (supervised and unsupervised)
   b. Data Sparsity:
      i. Documents contain small subset of feature terms
   c. Lack of training examples (supervised):
      i. Too few training examples for each class
      ii. Imbalanced population: count of (+)s << count of (-)s

What is Deep Learning?

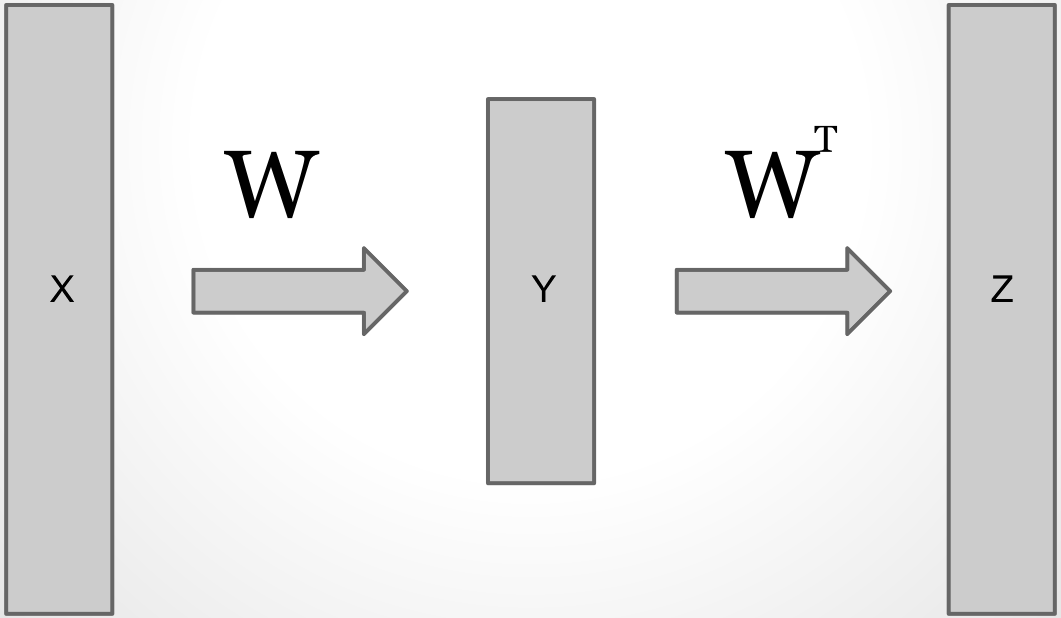
Deep learning is...

Artificial Neural Network w/ multiple hidden layers



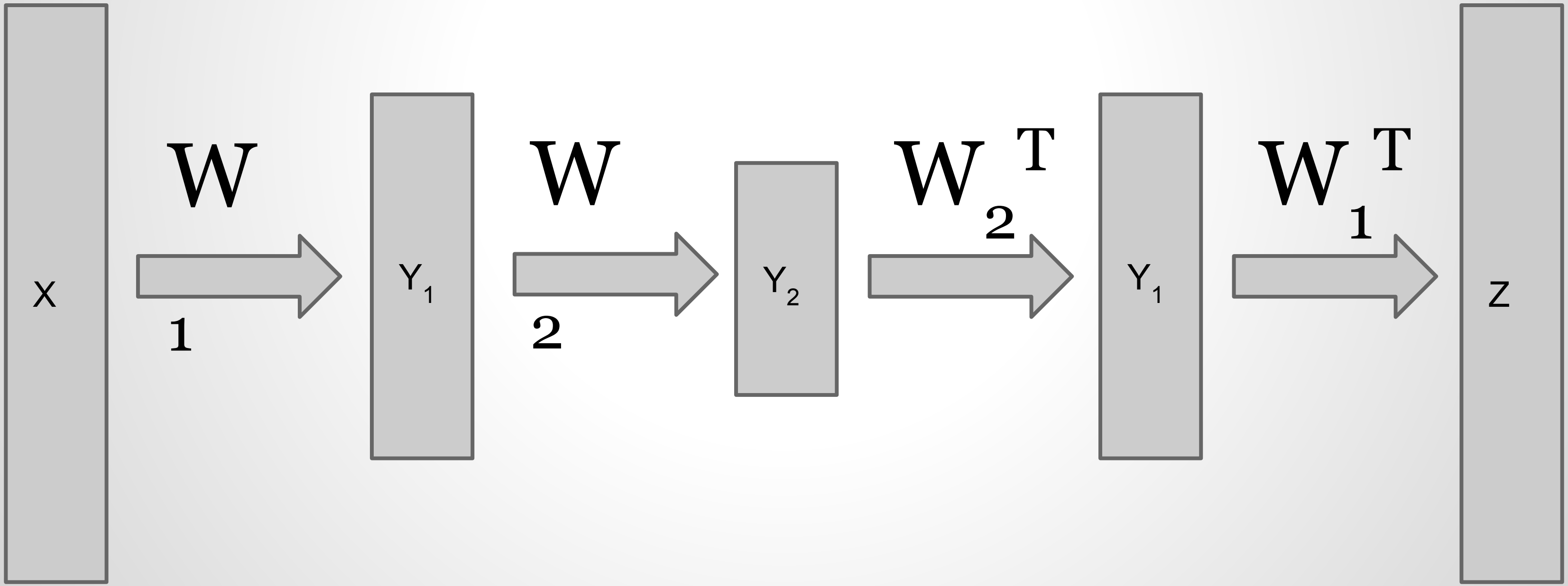
AutoEncoder for Feature Compression

Minimize reconstruction error J = Loss(X,Z)



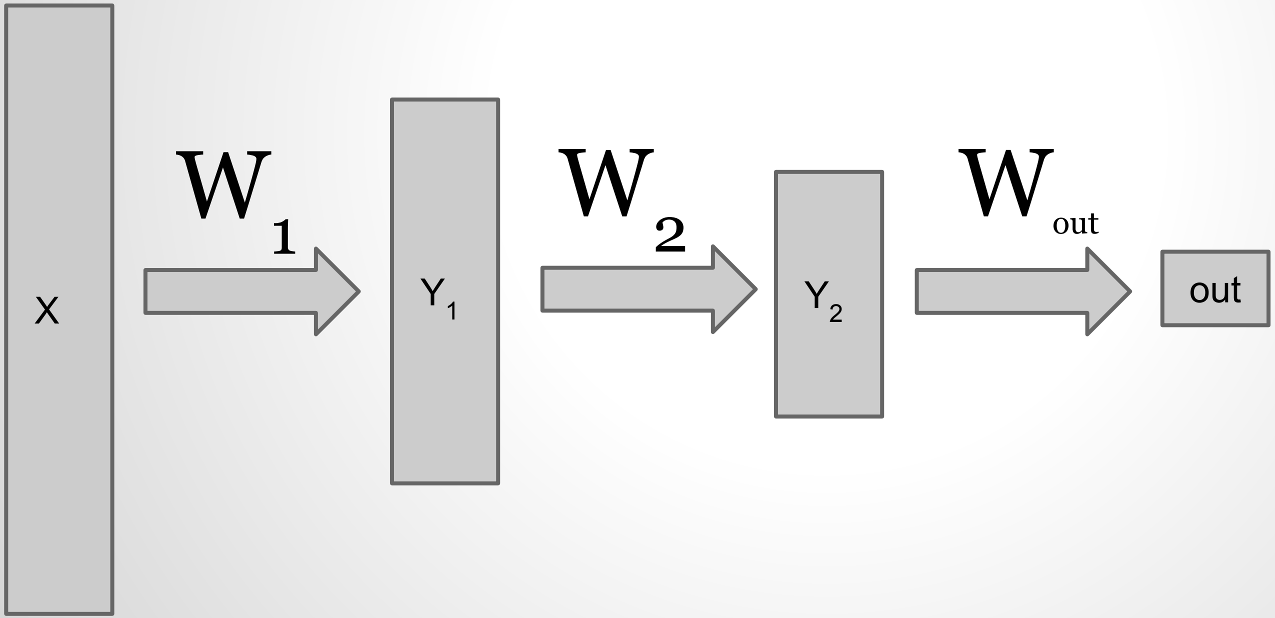
… and repeat until desired depth

Minimize reconstruction error J = Loss(X,Z)



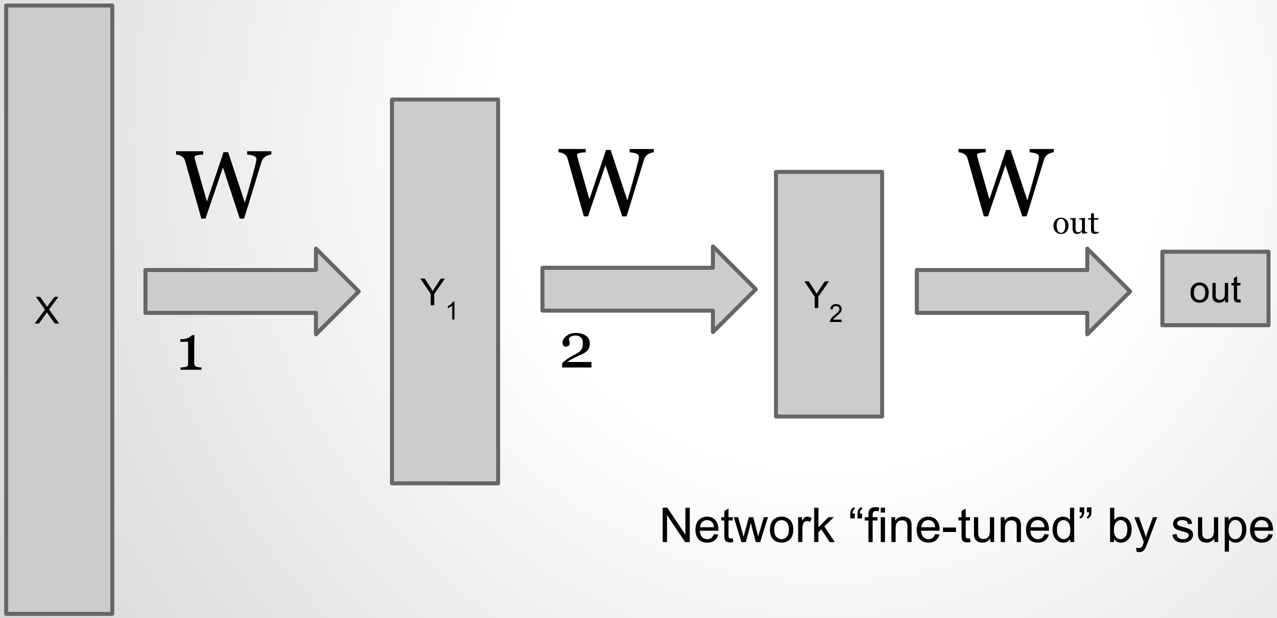
AE for Pre-training of Supervised Net

Minimize prediction error J = Loss(out,label)



AE for Pre-training of Supervised Net

Minimize prediction error J = Loss(out,label)

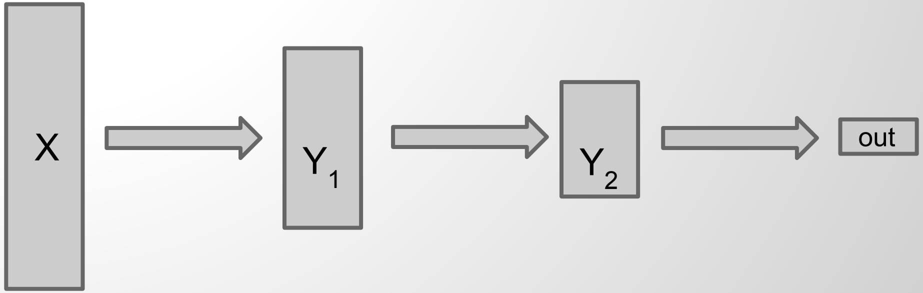


Network “fine-tuned” by supervised BackProp

AE for Pre-training of Supervised Net

Downsides:

- Unstable
- Difficult to implement
- Tuning cost scales with order of taxonomy node count
  - Time consuming
  - Expensive
  - Training label cost



What is Word2Vec?

Continuous vector representation for individual terms:

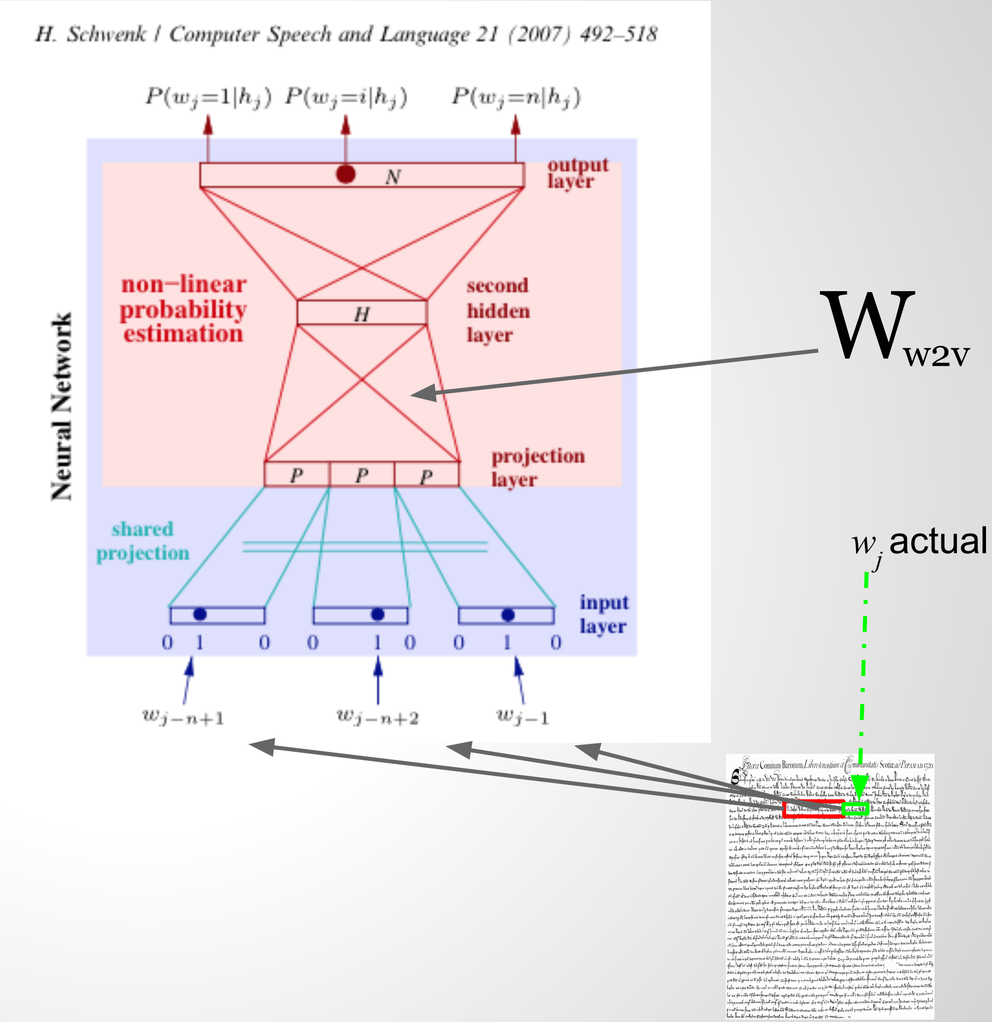
- Trained to specialize in sentence completion
- n-gram or skip gram
- Learns grammar
- Learns conceptual relationships

Word2Vec

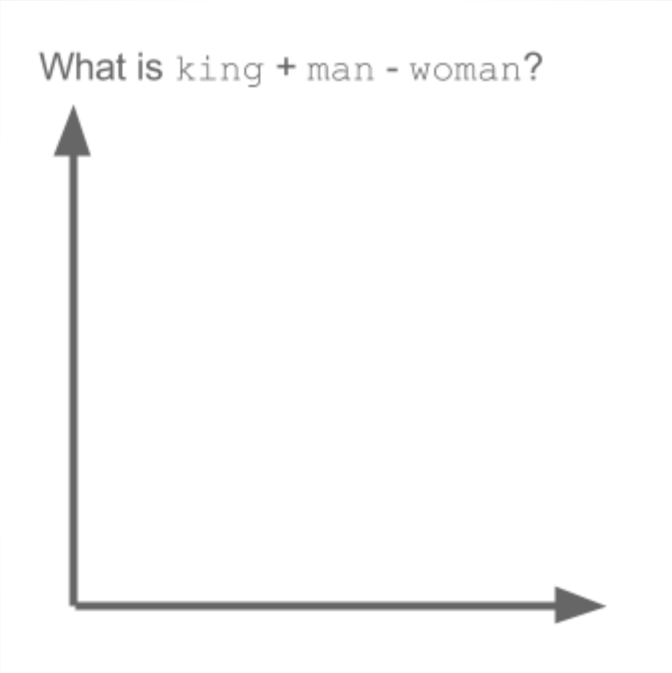
N-gram ANN classifier:

1. Project the "context" h_j (w_{j-n+1} to w_{j-1})
2. Soft-max predictor for output layer
3. Use BackProp algorithm to execute gradient descent to tune ANN loss on the actual w_j

(Can also do a "skip gram")

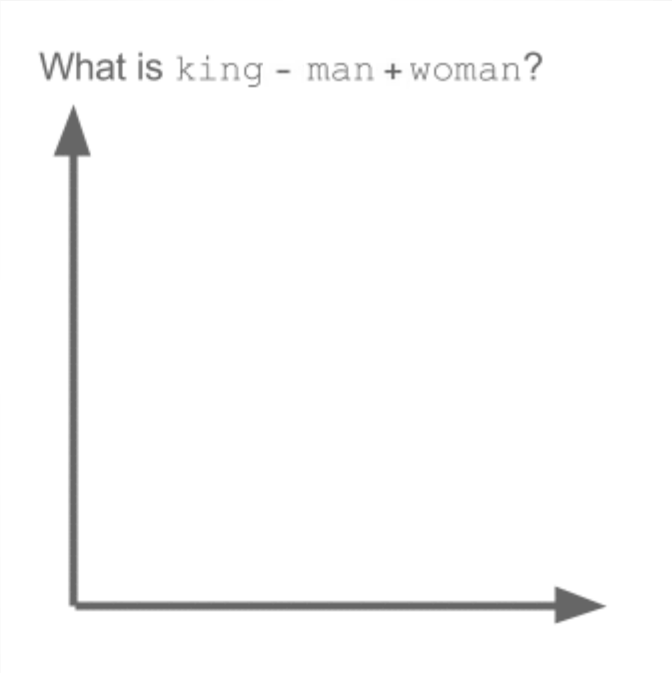


W_{w2v} Matrix Captures Conceptual Relations:



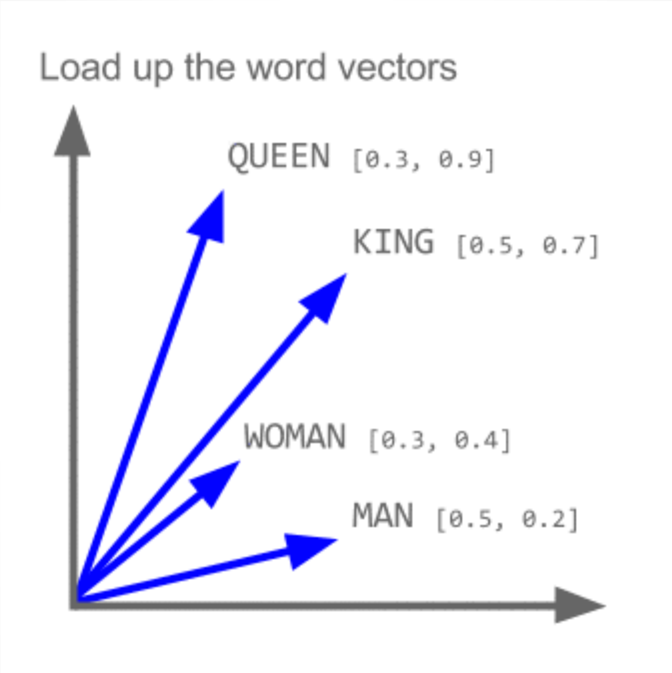
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



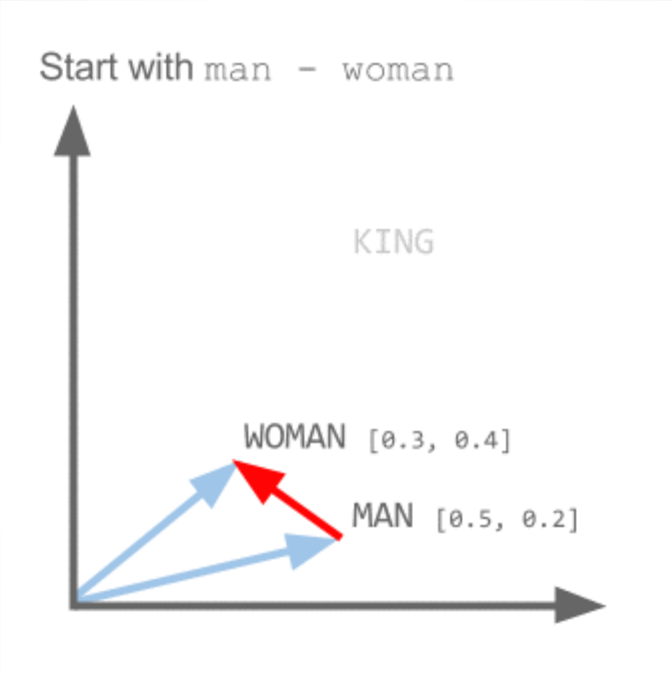
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



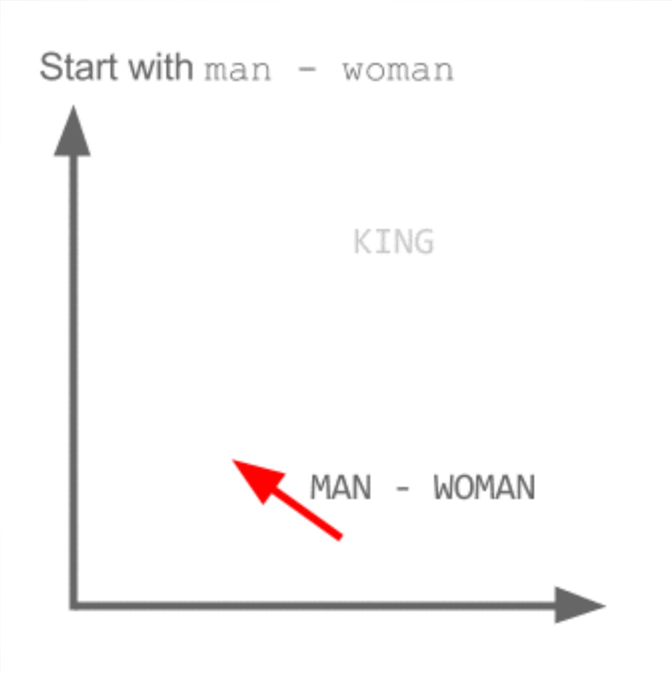
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



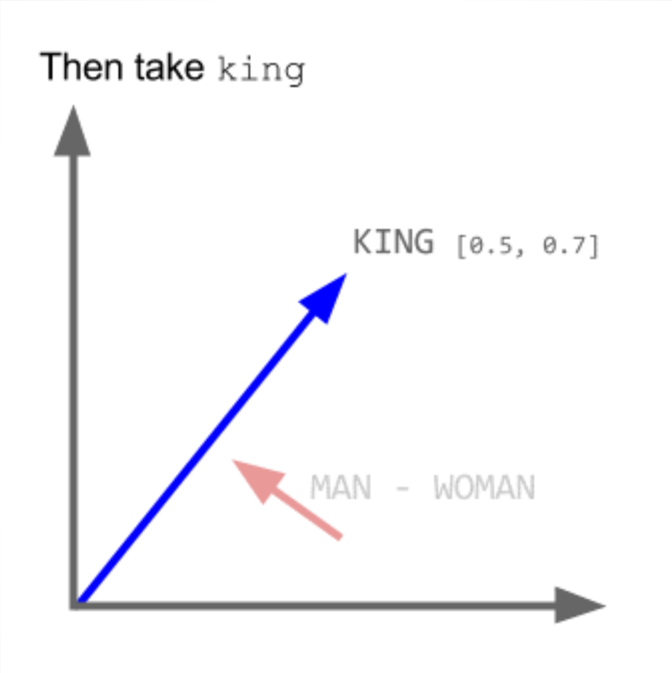
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



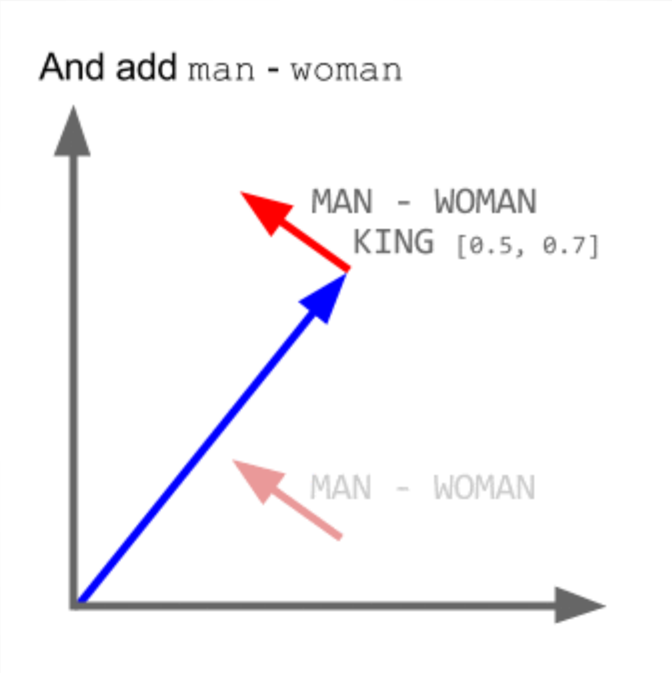
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



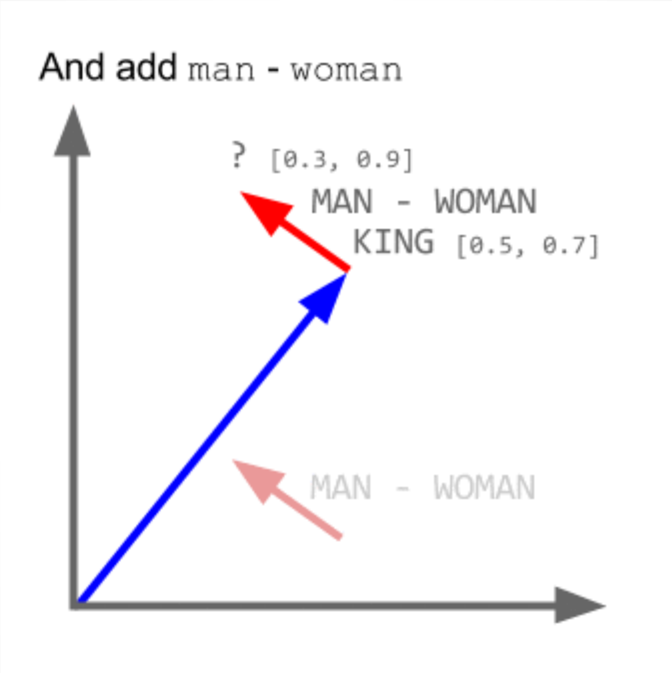
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



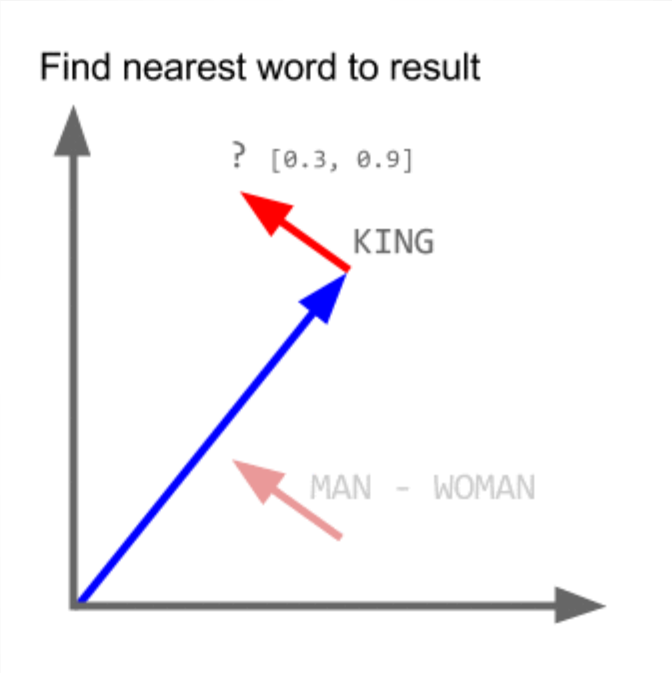
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



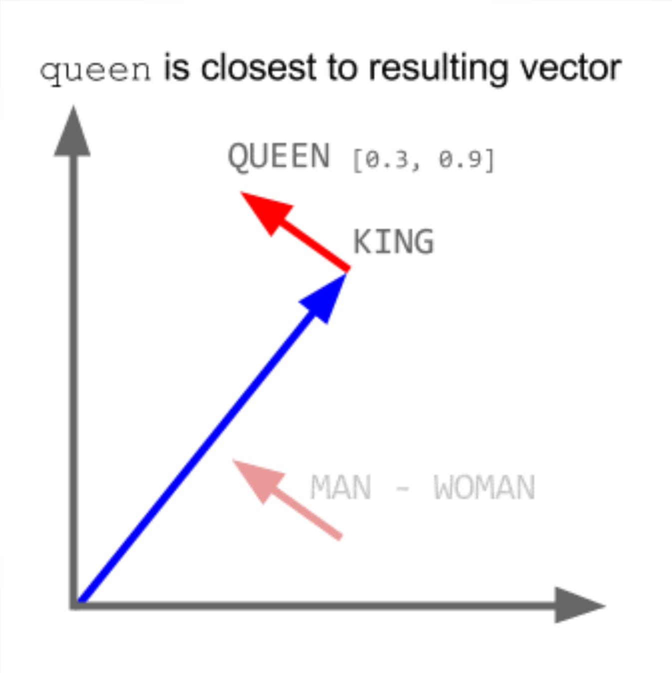
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



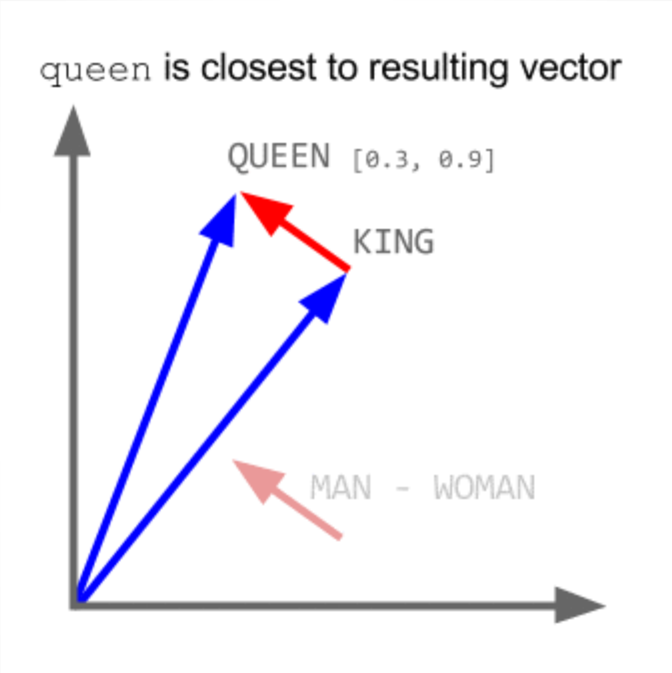
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



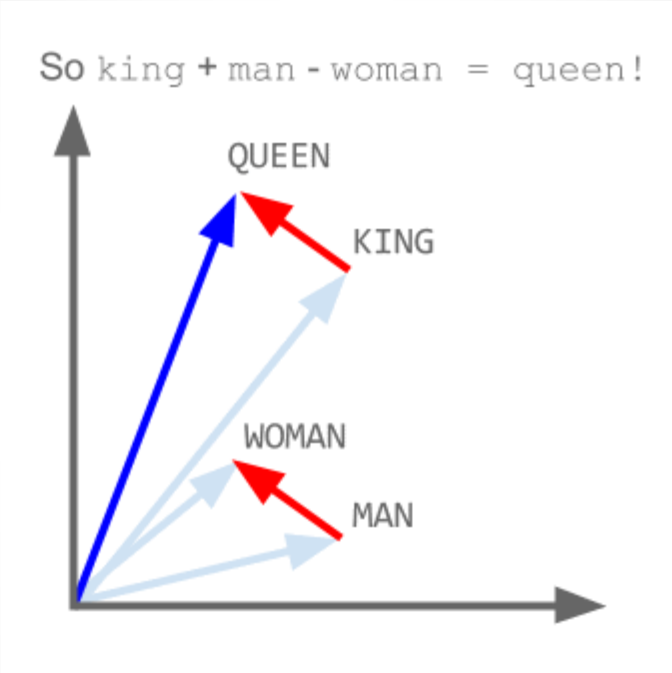
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



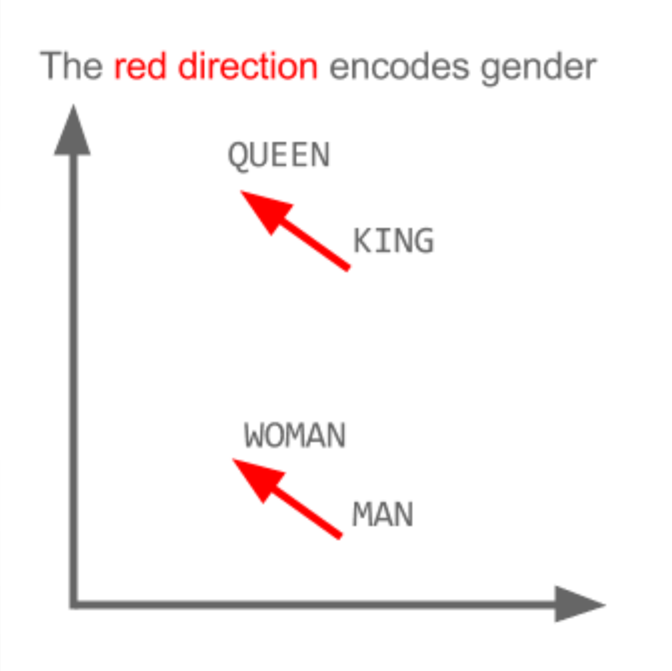
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



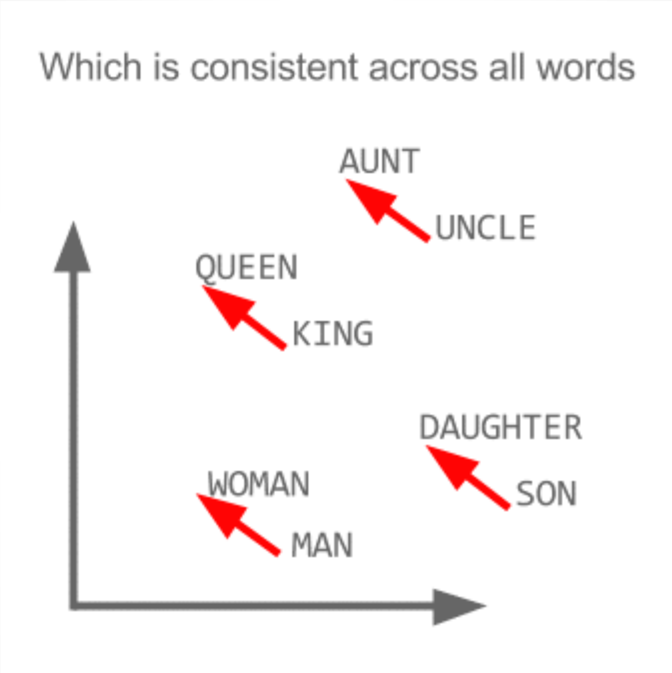
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



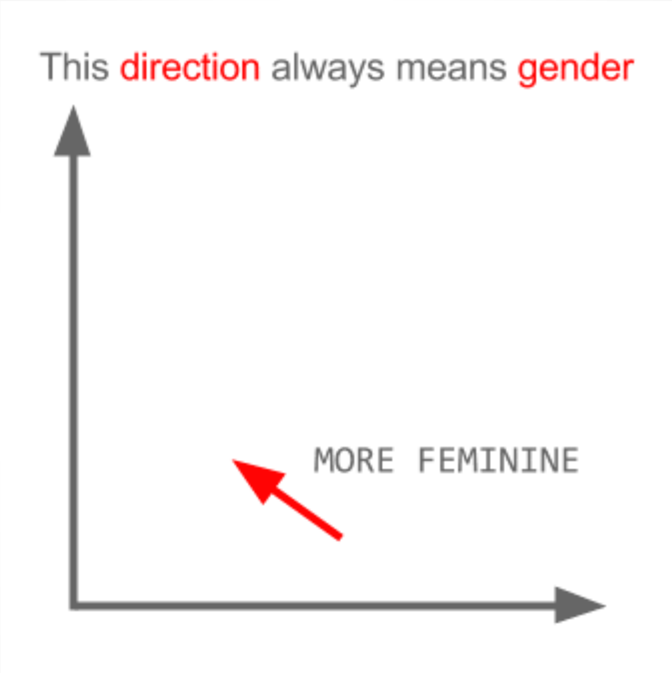
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



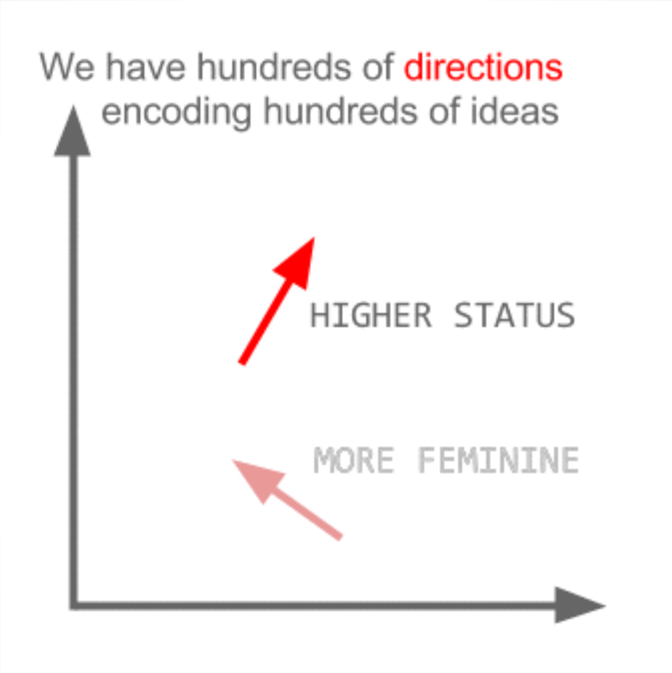
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



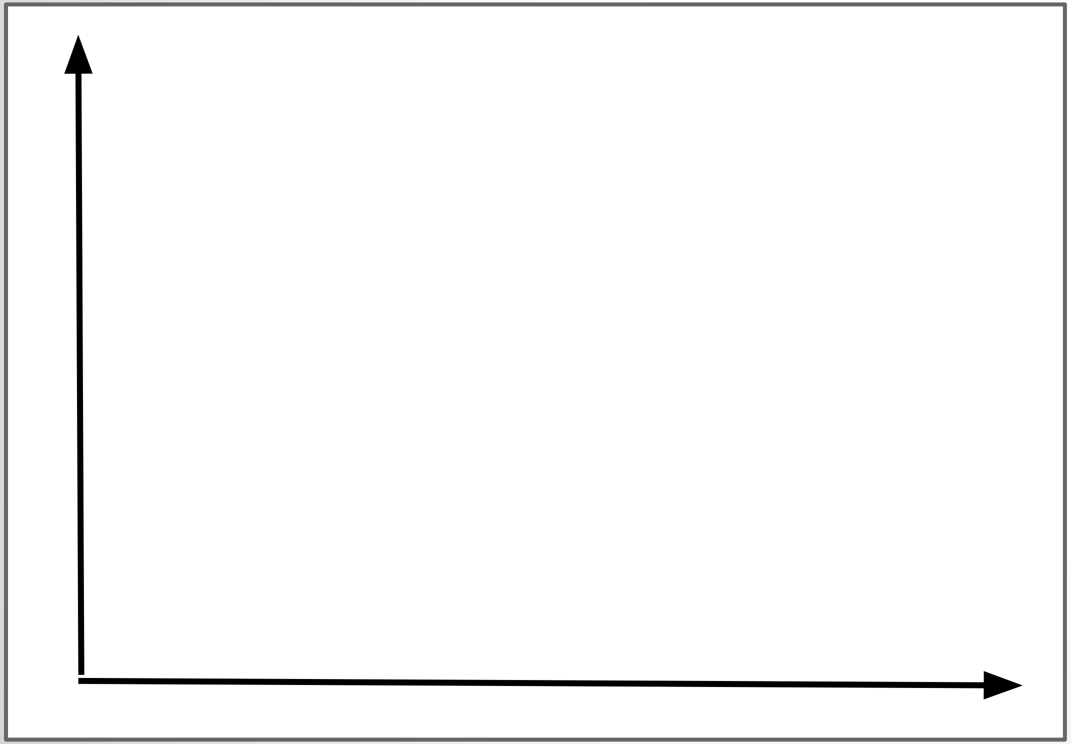
http://technology.stitchfix.com/assets/images/blog/vectors.gif

W_{w2v} Matrix Captures Conceptual Relations:



http://technology.stitchfix.com/assets/images/blog/vectors.gif

What is Doc2Vec?



… best of times it was

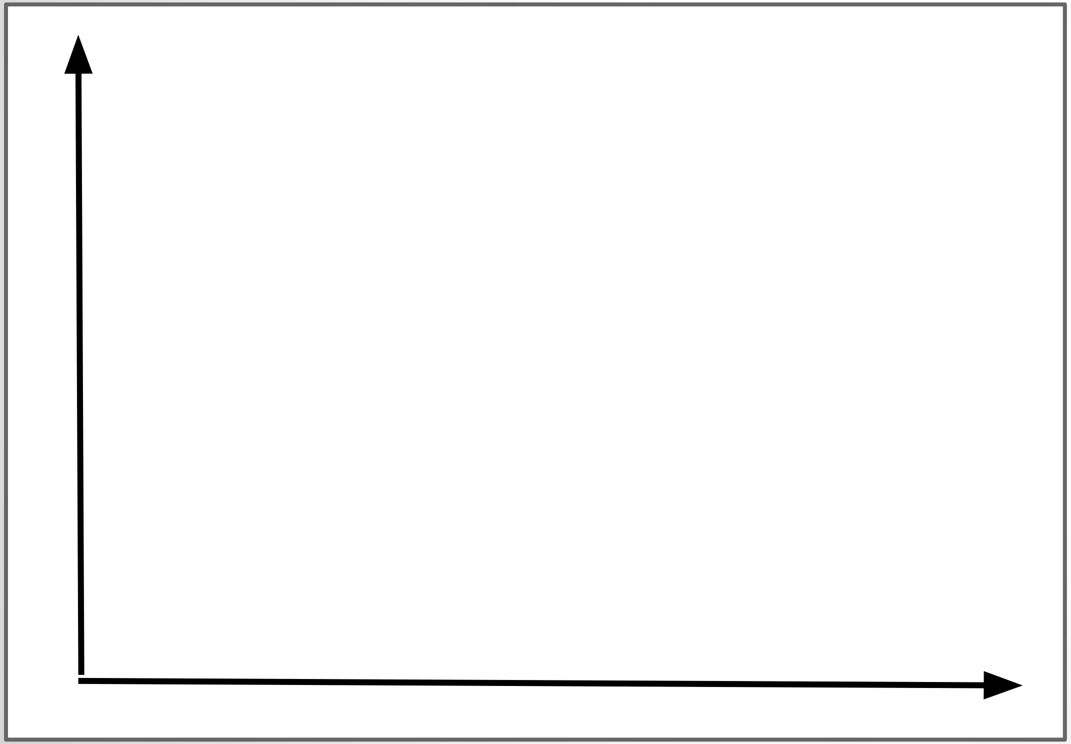
the worst of times, it

was the age of

wisdom, it was the age

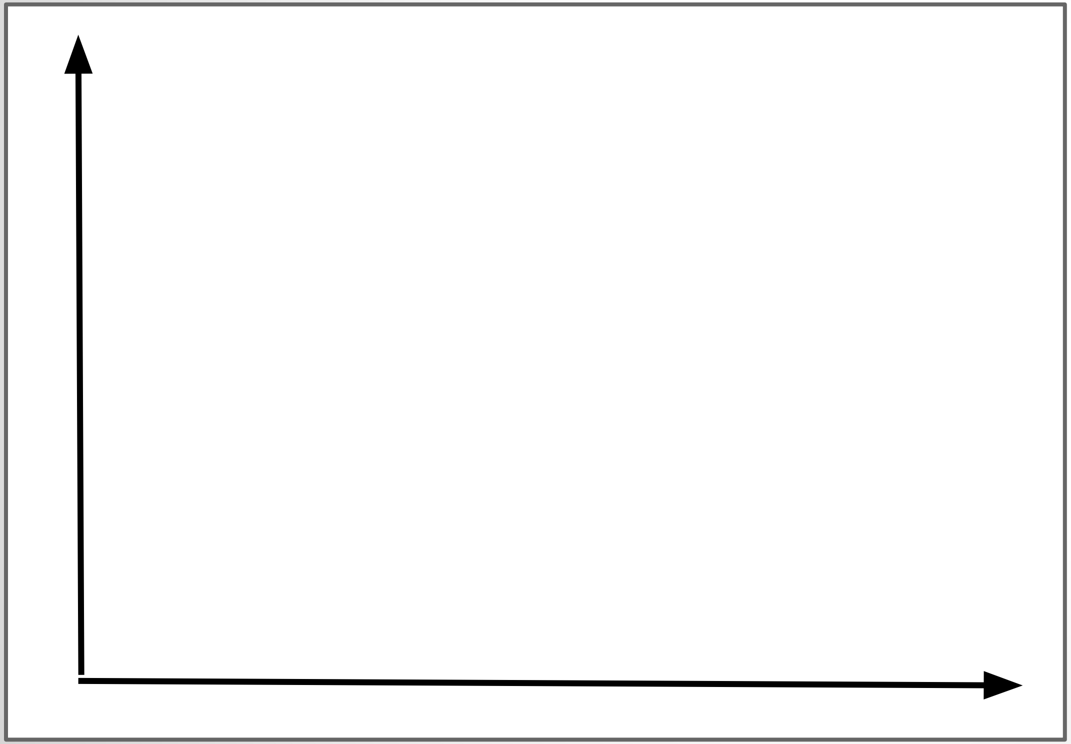
of foolishness ...

What is Doc2Vec?



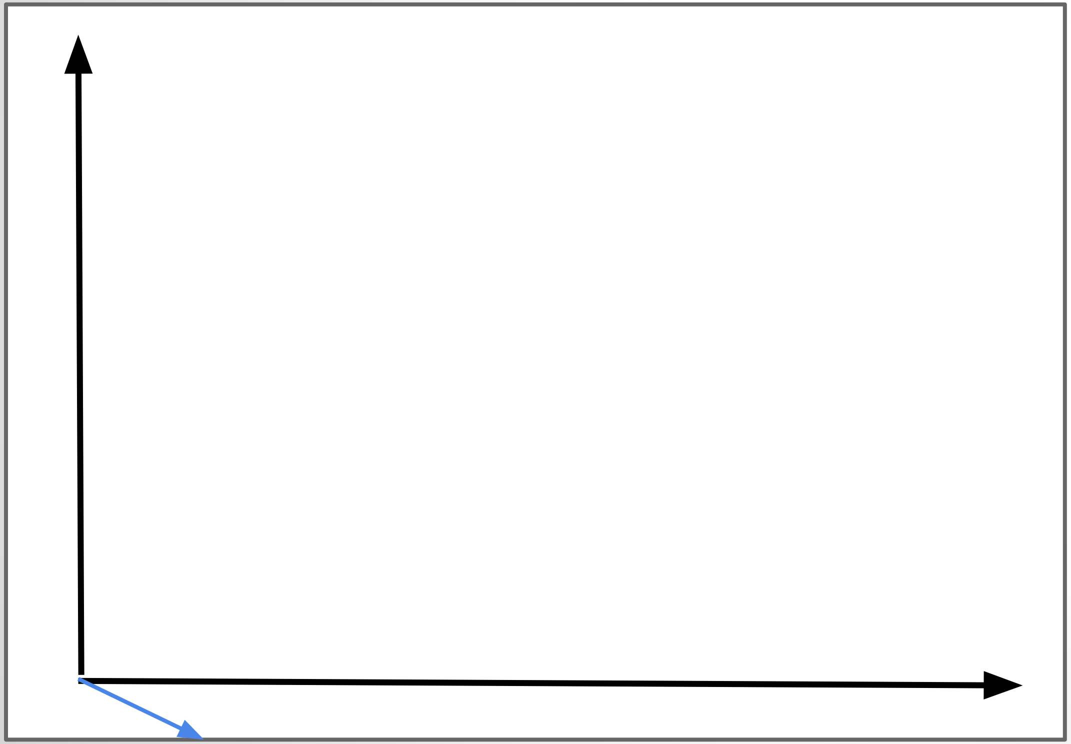
… best of times it was the worst of times, it was the age of wisdom, it was the age of foolishness …

What is Doc2Vec?



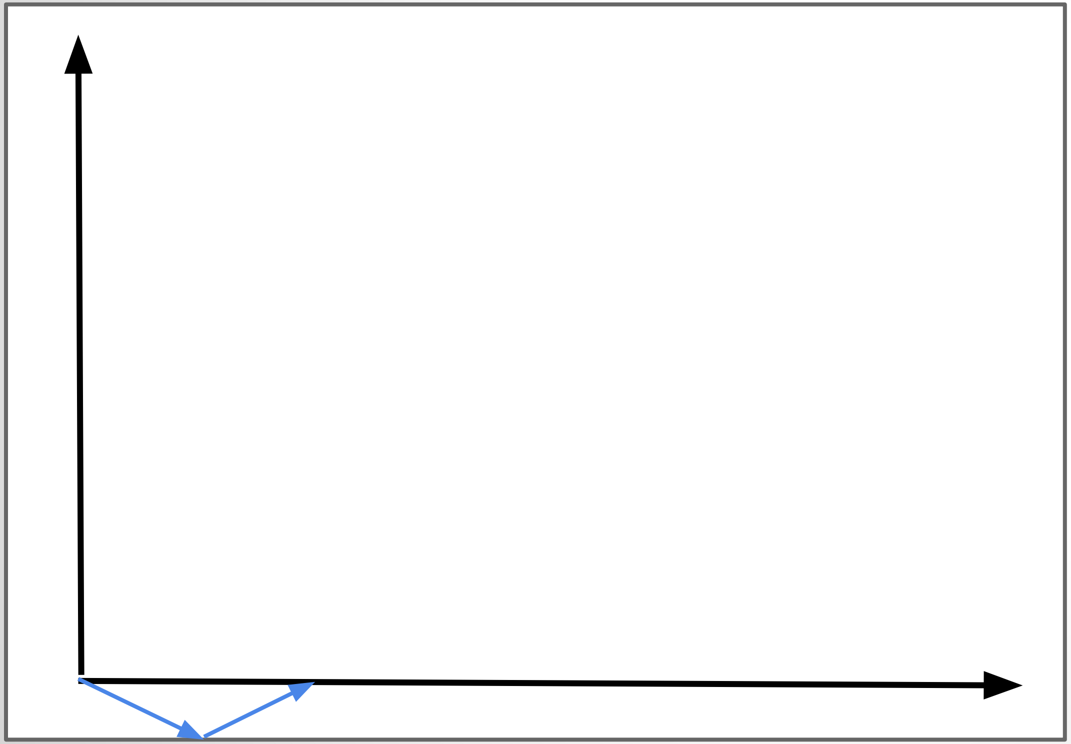
… best of times it was
the worst of times, it
was the age of
wisdom, it was the age
of foolishness ...

What is Doc2Vec?



… best of times it was
the worst of times, it
was the age of
wisdom, it was the age
of foolishness ...

What is Doc2Vec?



… best of times it was

the worst of times, it

was the age of

wisdom, it was the age

of foolishness ...

What is Doc2Vec?

… best of times it was

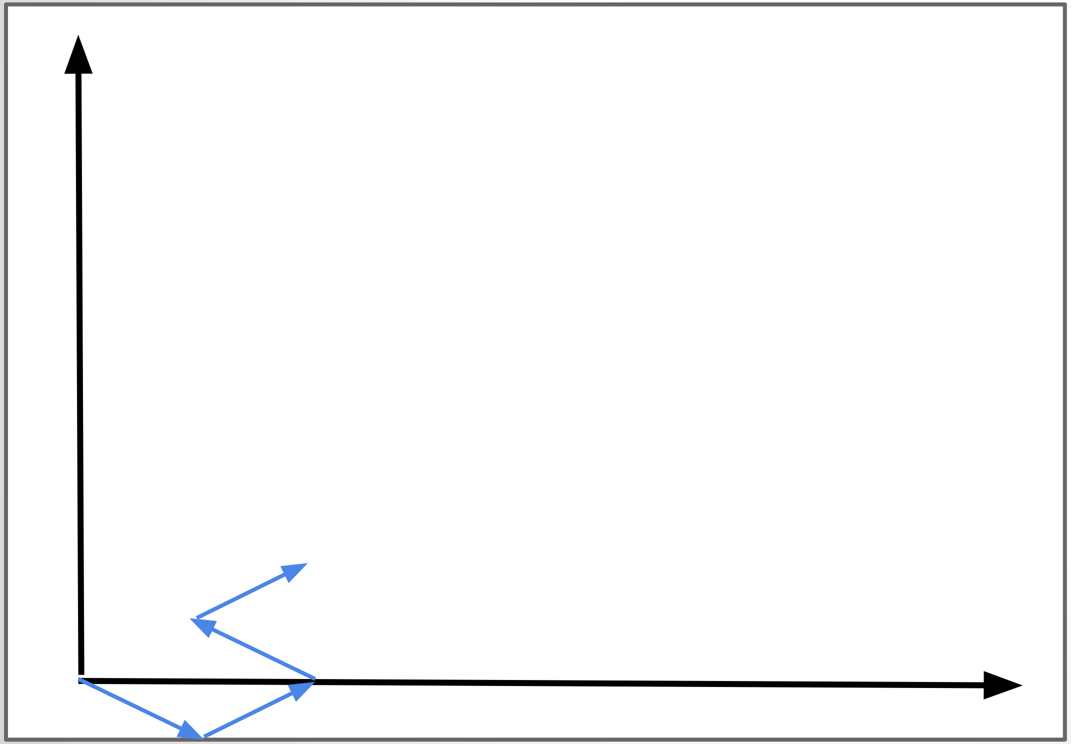
the worst of times, it

was the age of

wisdom, it was the age

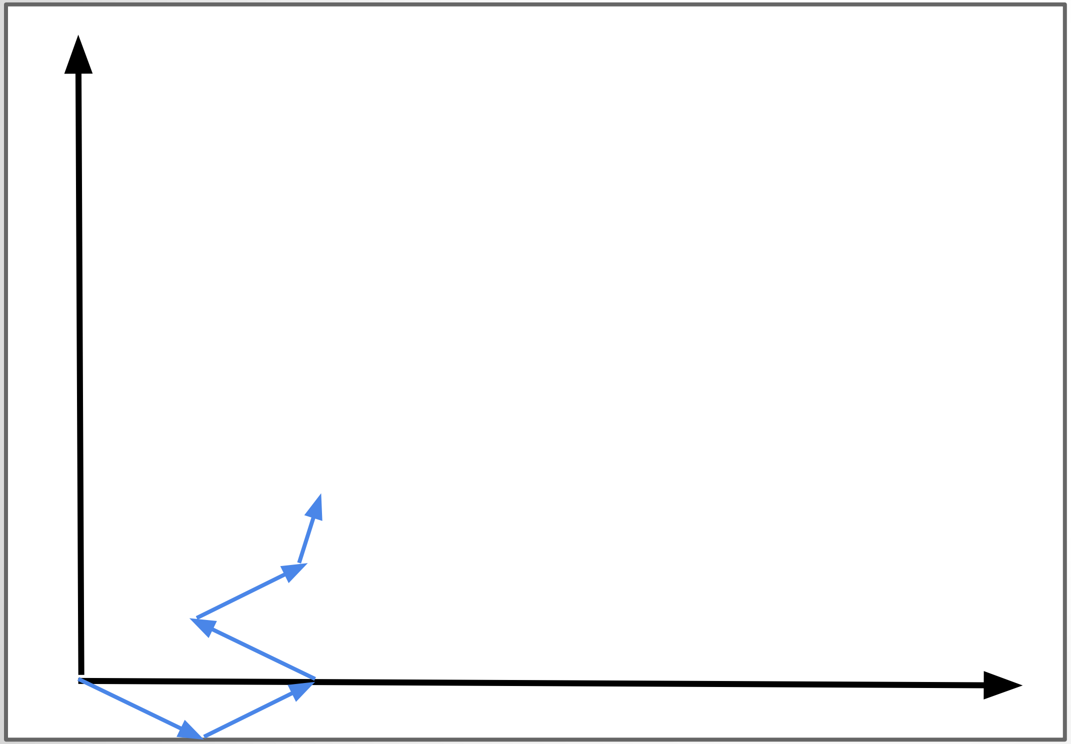
of foolishness …

What is Doc2Vec?



… best of times it was
the worst of times, it
was the age of
wisdom, it was the age
of foolishness …

What is Doc2Vec?



… best of times it was

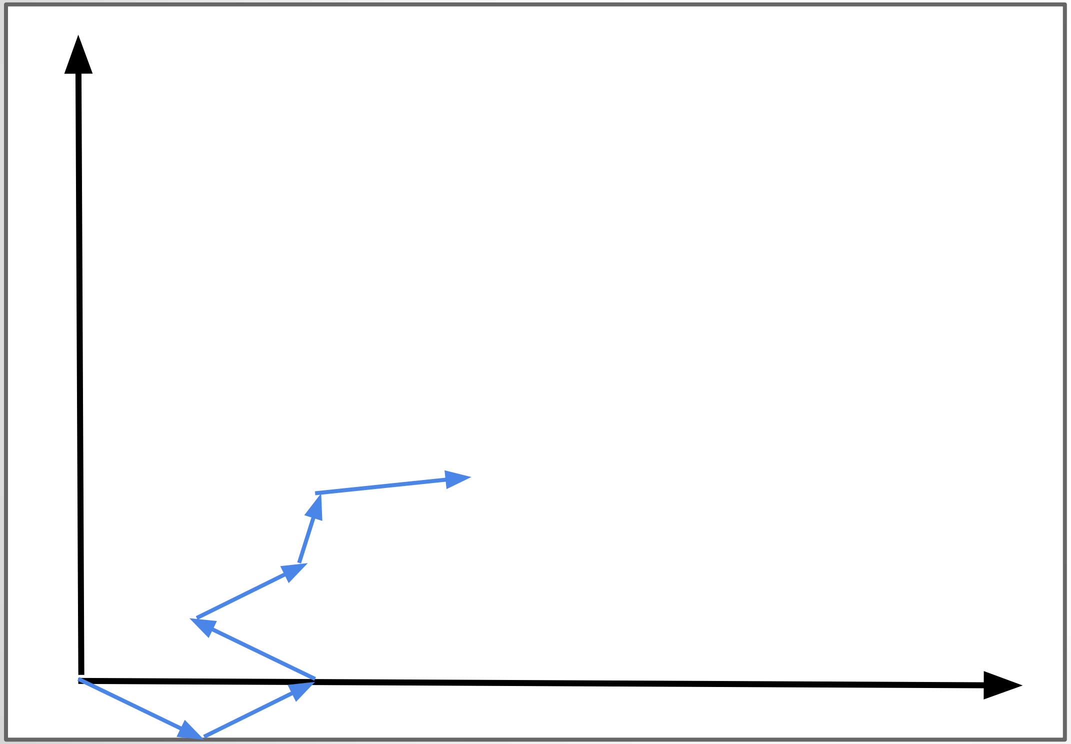
the worst of times, it

was the age of

wisdom, it was the age

of foolishness ...

What is Doc2Vec?



… best of times it was

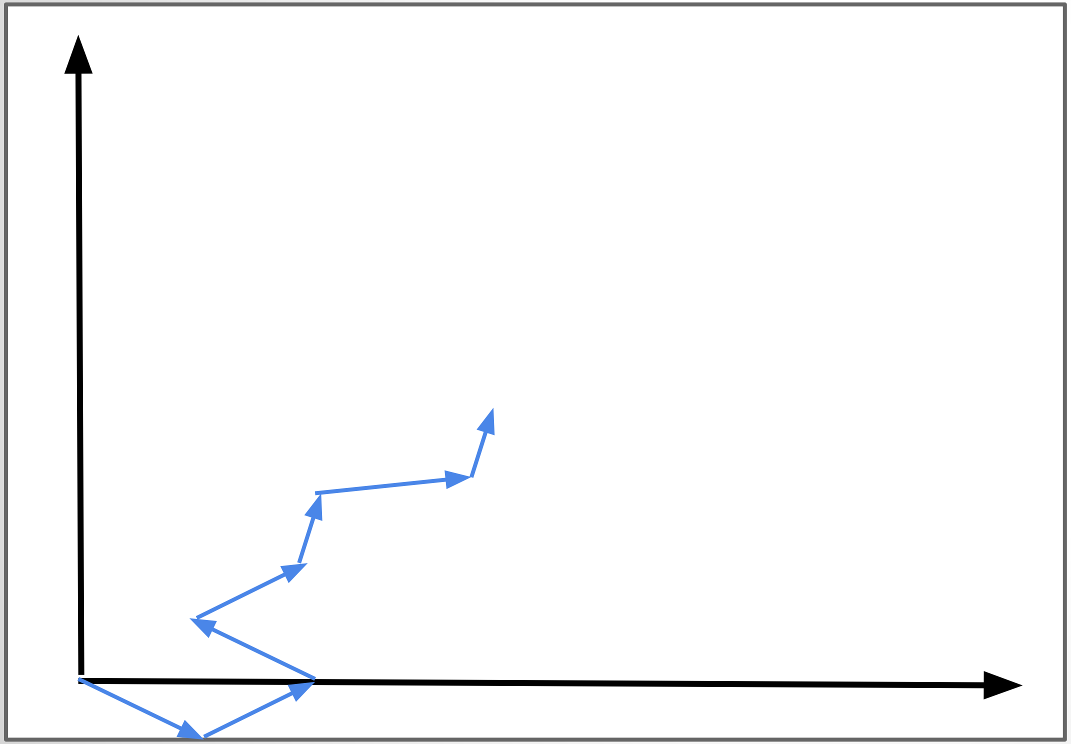
the worst of times, it

was the age of

wisdom, it was the age

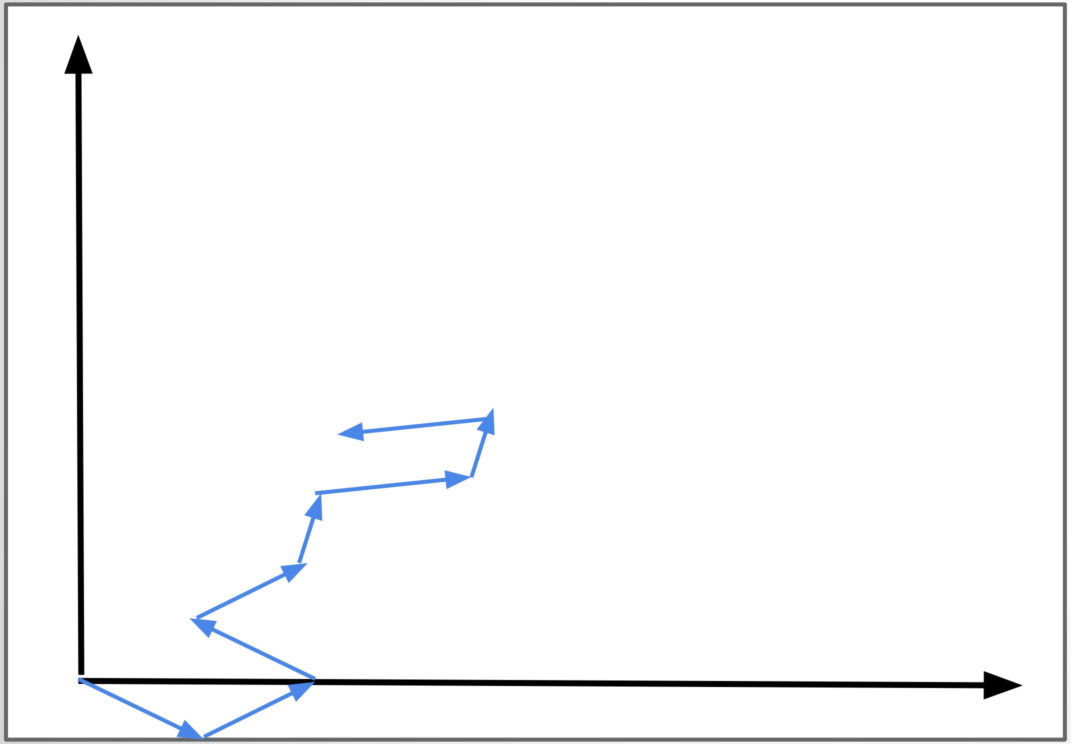
of foolishness …

What is Doc2Vec?



… best of times it was the worst of times, it was the age of wisdom, it was the age of foolishness …

What is Doc2Vec?



… best of times it was

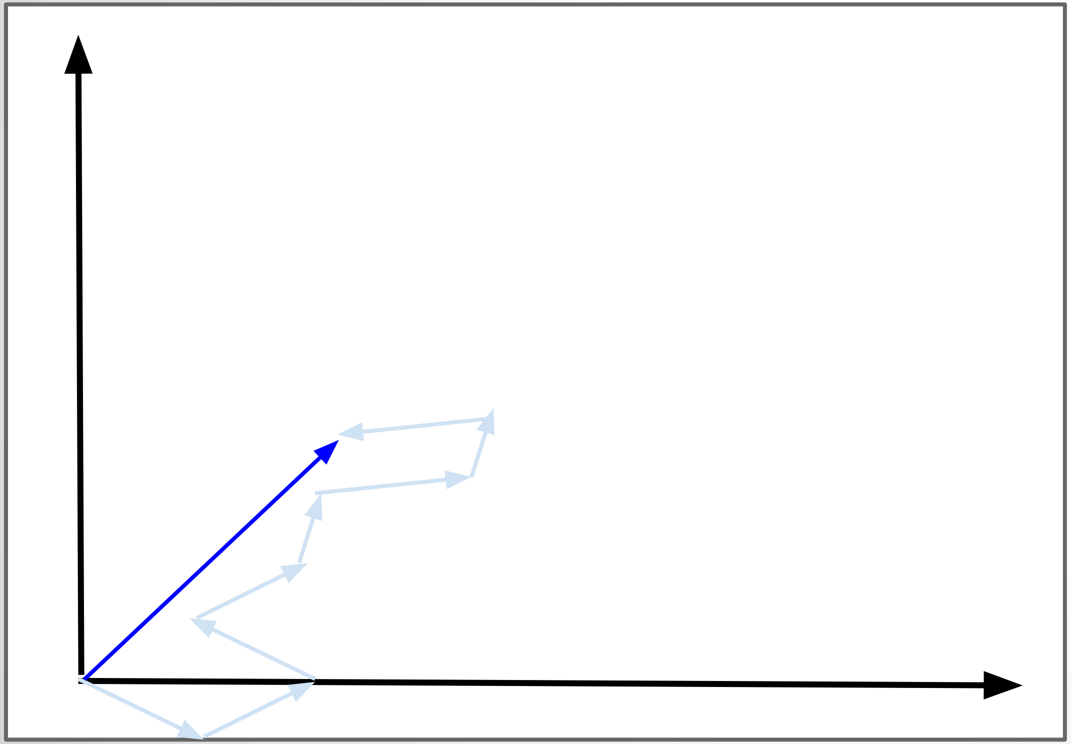
the worst of times, it

was the age of

wisdom, it was the age

of foolishness ...

What is Doc2Vec?



… best of times it was

the worst of times, it

was the age of

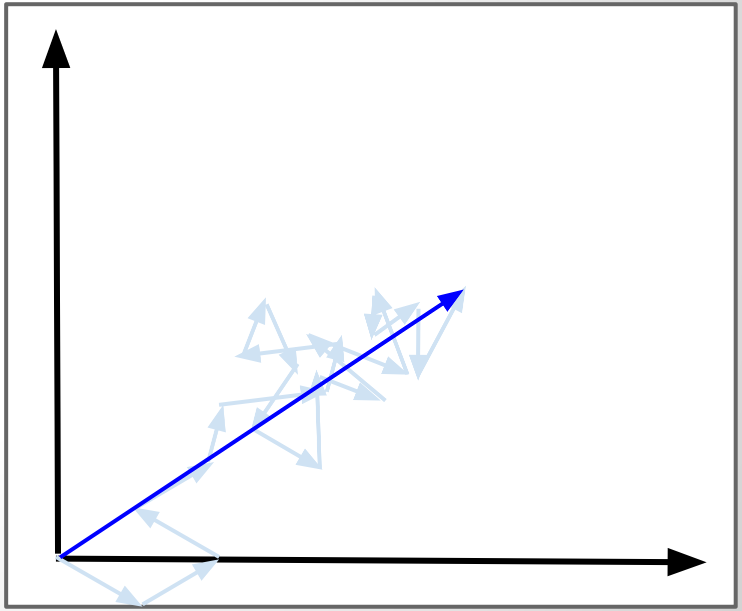
wisdom, it was the age

of foolishness ...

What is Doc2Vec?

Taking the linear combination of every term in the document creates a **random walk** with **bias** process in the w2v space.

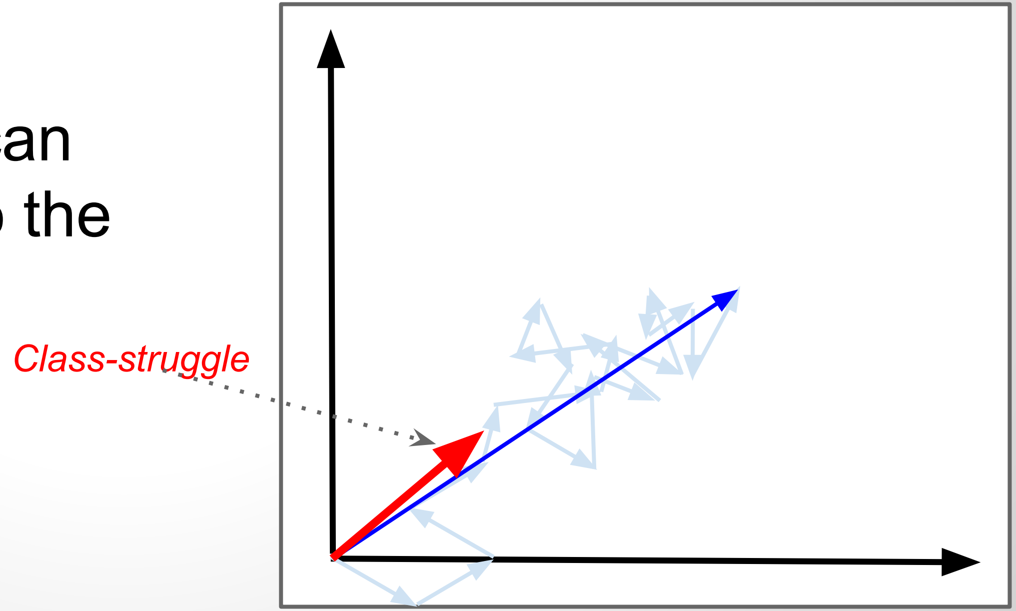
- In aggregate, the sum vector drifts in the direction of the aggregate topic of the document.



What is Doc2Vec?

Taking the linear combination of every term in the document creates a **random walk** with **bias** process in the w2v space.

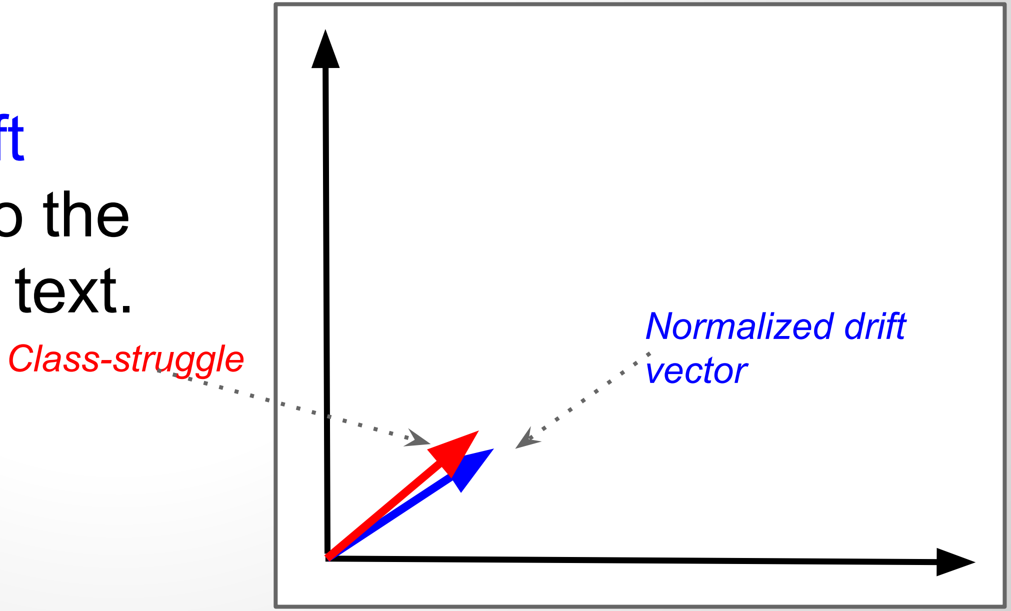
- And taxonomy topics can also be embedded into the w2v space.



What is Doc2Vec?

Taking the linear combination of every term in the document creates a **random walk** with **bias** process in the w2v space.

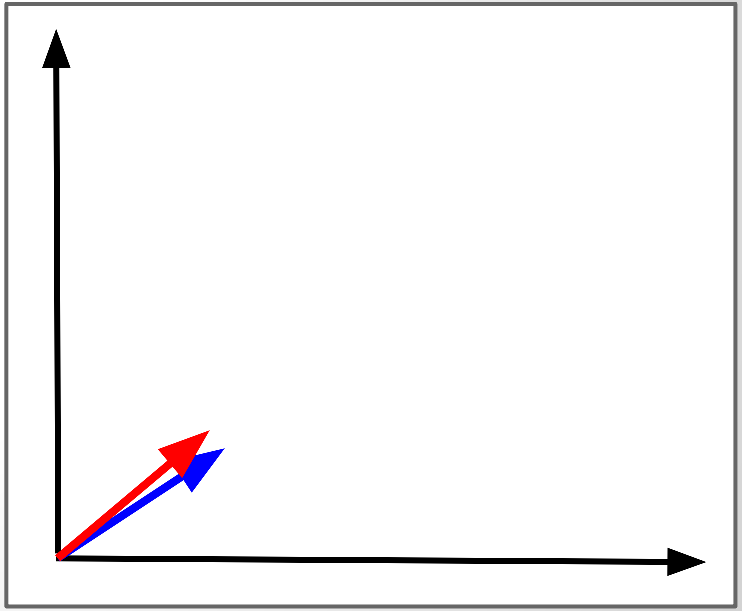
- The direction of the drift vector tends to rotate to the direction of topic of the text.



What is Doc2Vec?

Taking the linear combination of every term in the document creates a **random walk** with **bias** process in the w2v space.

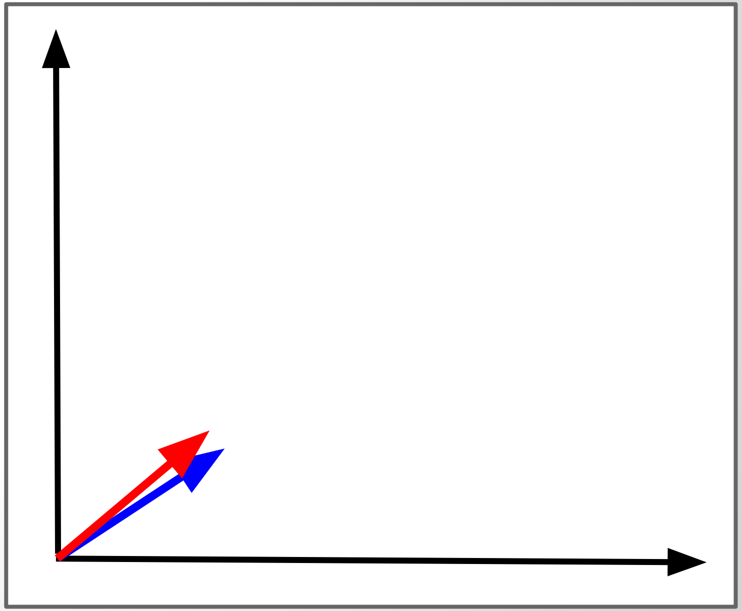
- The angle of the drift vector can then be used as a topic feature for the vector



What is Doc2Vec?

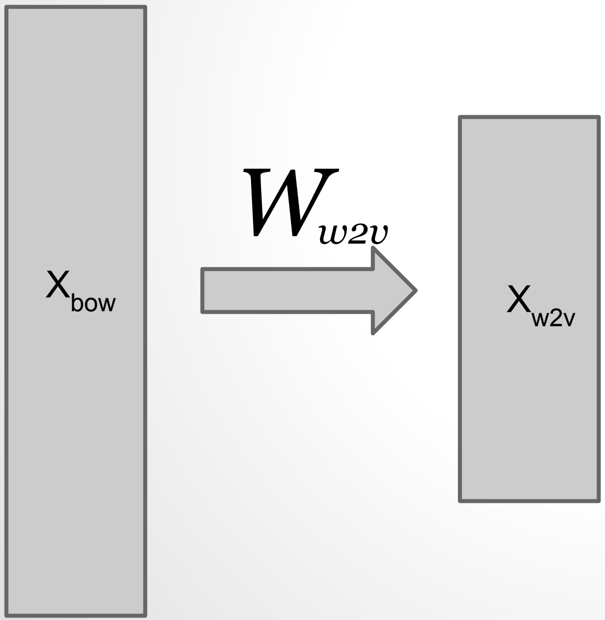
Taking the linear combination of every term in the document creates a **random walk** with **bias** process in the w2v space.

- The angle of the drift vector can then be used as a topic feature for the vector
- Distance (cos, L_1, L_2, etc) are effective doc features applied to text classification.



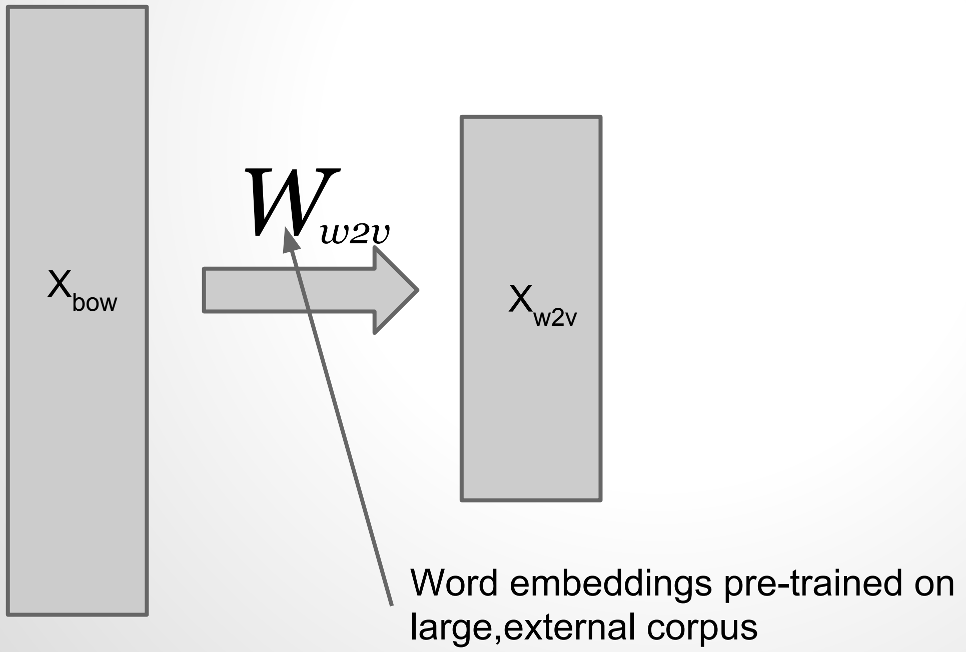
w2v as Feature Compression

Minimize prediction error J = Loss(out,label)



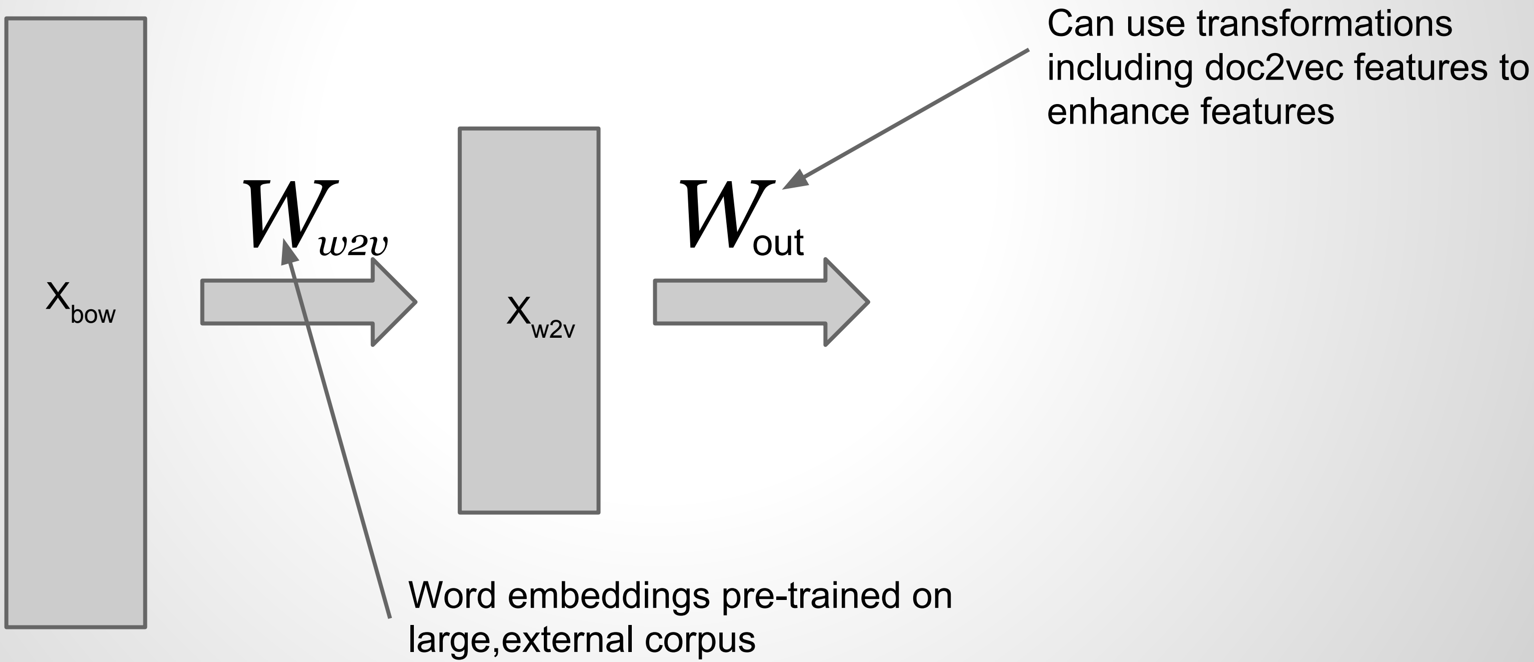
w2v as Feature Compression

Minimize prediction error J = Loss(out,label)



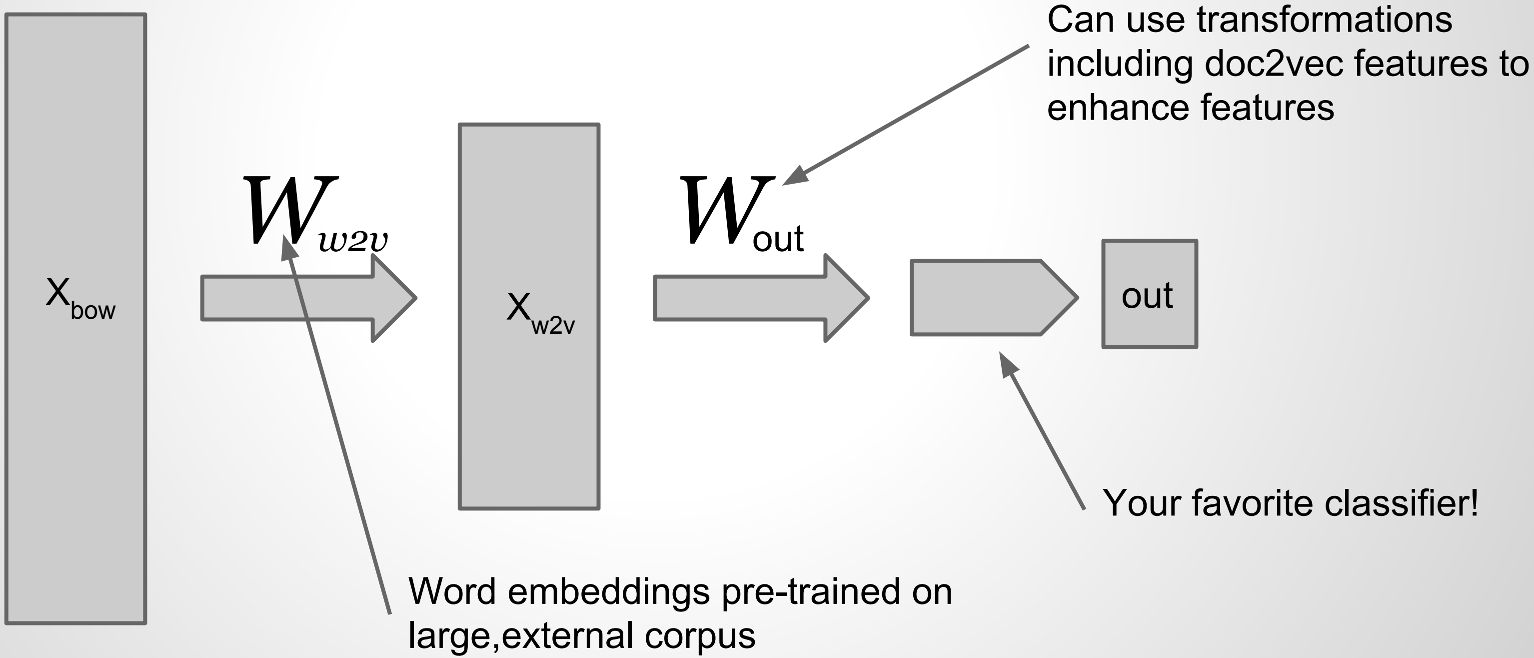
w2v as Feature Compression

Minimize prediction error J = Loss(out,label)



w2v as Feature Compression

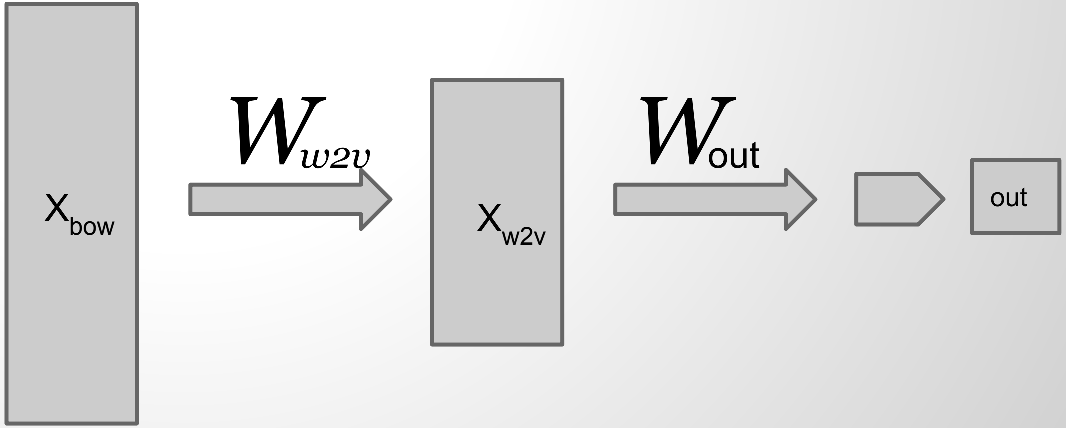
Minimize prediction error J = Loss(out,label)



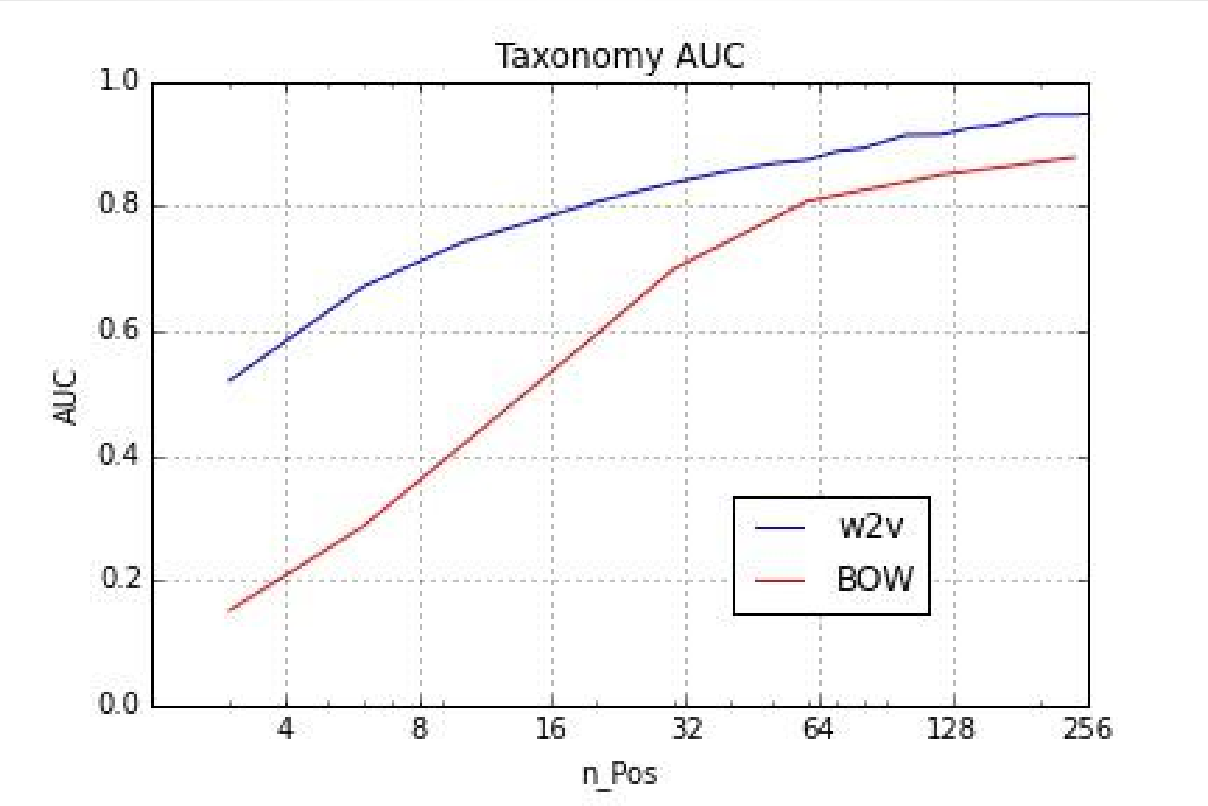
w2v as Feature Compression

Benefits:

- Sparse vectors made dense
- Training time restricted to output layer
- No expensive hyperparameter search
- More effective usage of sparse labels



w2v with Sparse Labels



Imbalance in Text

- Text classification problems are typically very imbalanced.

Imbalance in Text

- Text classification problems are typically very imbalanced.
  - Small number of (+)s vs (-)s

Imbalance in Text

- Text classification problems are typically very imbalanced.
  - Small number of (+)s vs (-)s
- Because of imbalance, real model performance can be far worse than estimated by balanced testing.

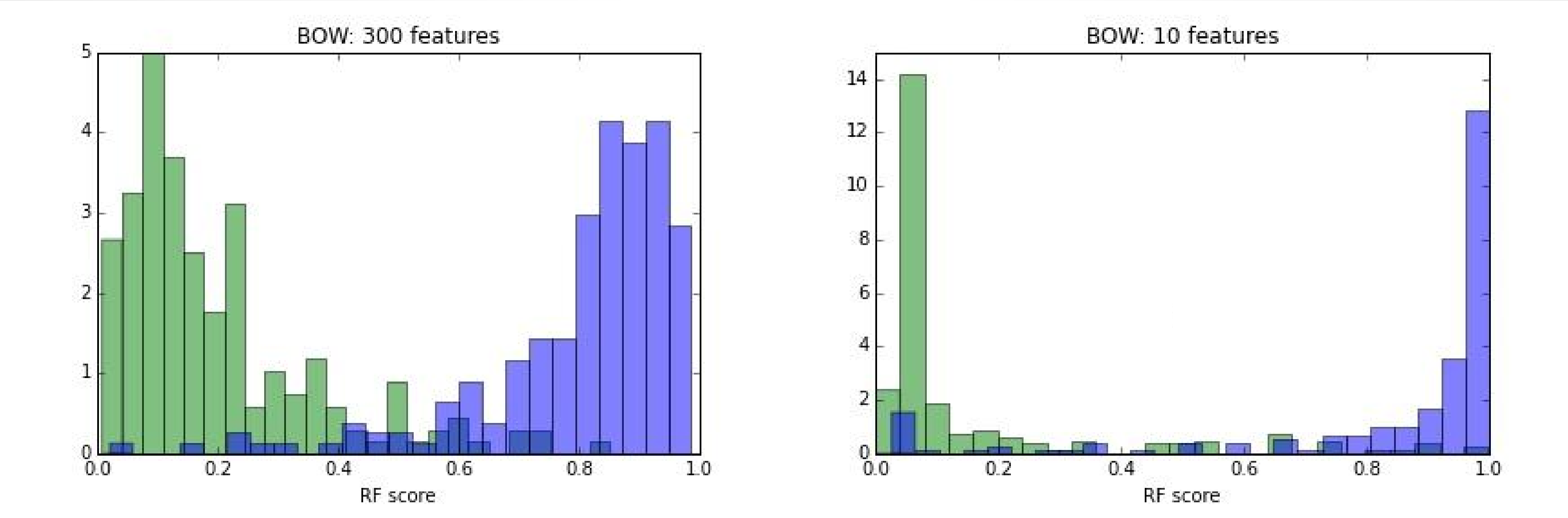
Imbalance in Text

- Text classification problems are typically very imbalanced.
  - Small number of (+)s vs (-)s
- Because of imbalance, real model performance can be far worse than estimated by balanced testing.
- Re-thresholding can help models perform well even under imbalanced conditions.

Imbalance in Text

- Text classification problems are typically very imbalanced.
    - Small number of (+)s vs (-)s
- Because of imbalance, real model performance can be far worse than estimated by balanced testing.
- Re-thresholding can help models perform well even under imbalanced conditions.
- Using feature selection to make classes well separated is essential to successful thresholding.

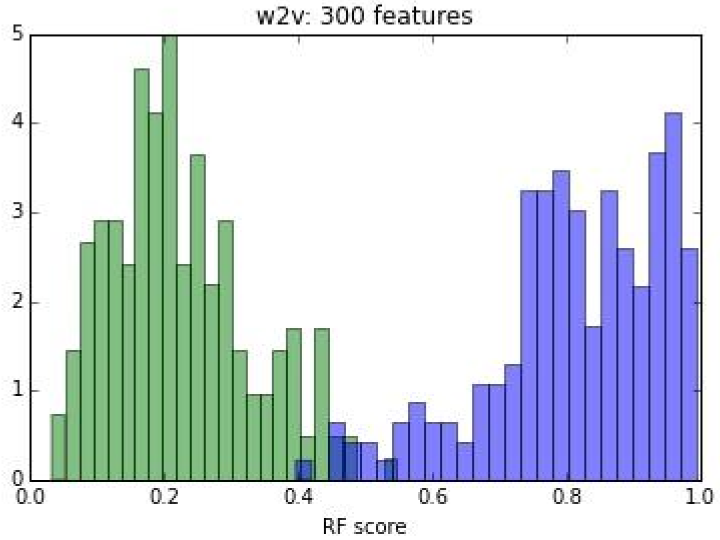
Comparing w2v and BOW

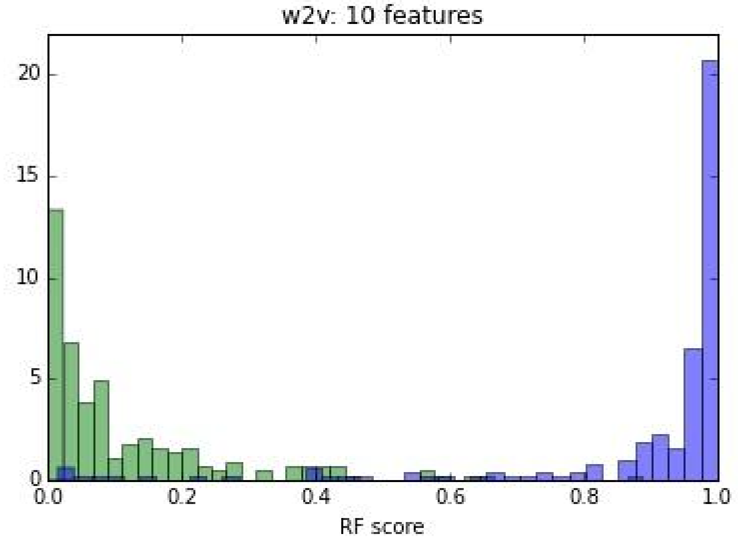


Significant loss of F_1 is incurred in achieving well separated class distributions

- bow 300: $F_1 = .933$
- bow 10: $F_1 = .885$

Comparing w2v and BOW

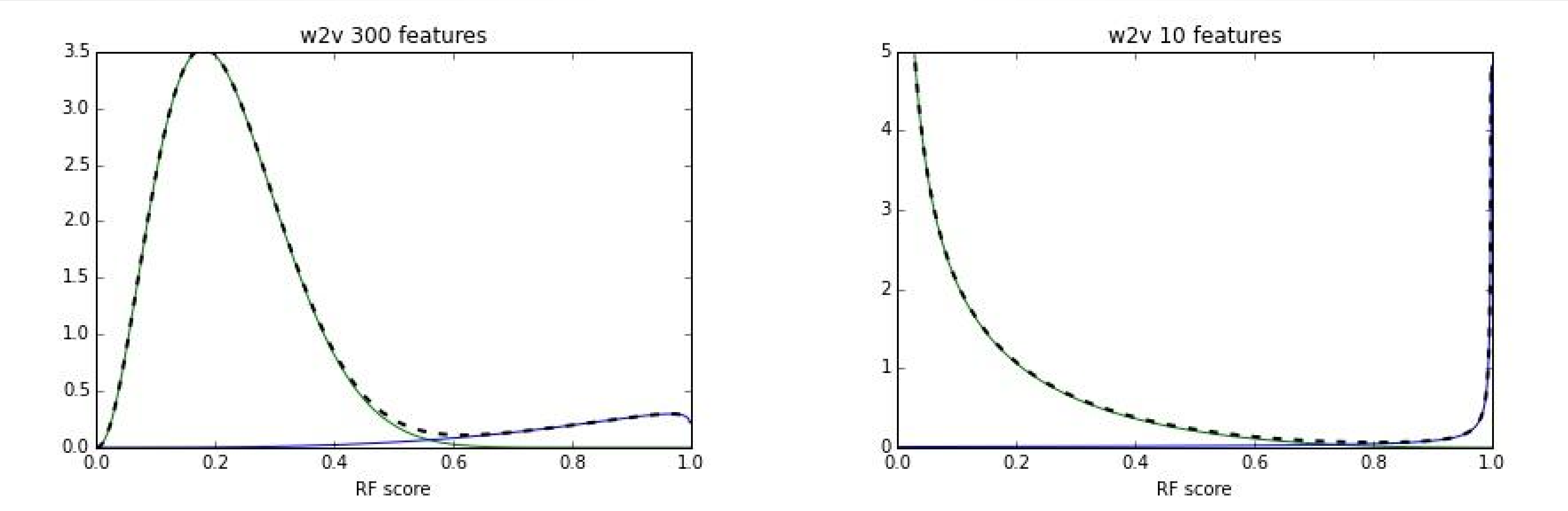




With doc2vec feature engineering, F_1 is higher overall and we achieve well separated class distributions with smaller loss in precision and recall

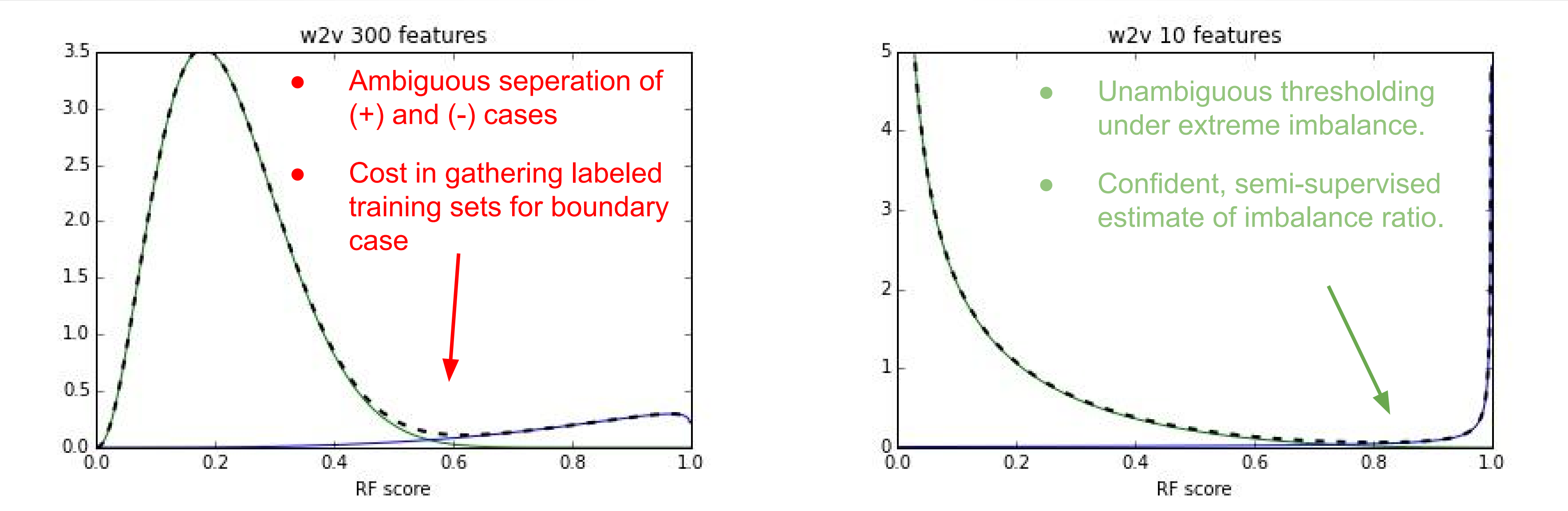
- w2v 300: F_1 = .964
- w2v 10: F_1 = .946

Better Label Imbalance Management

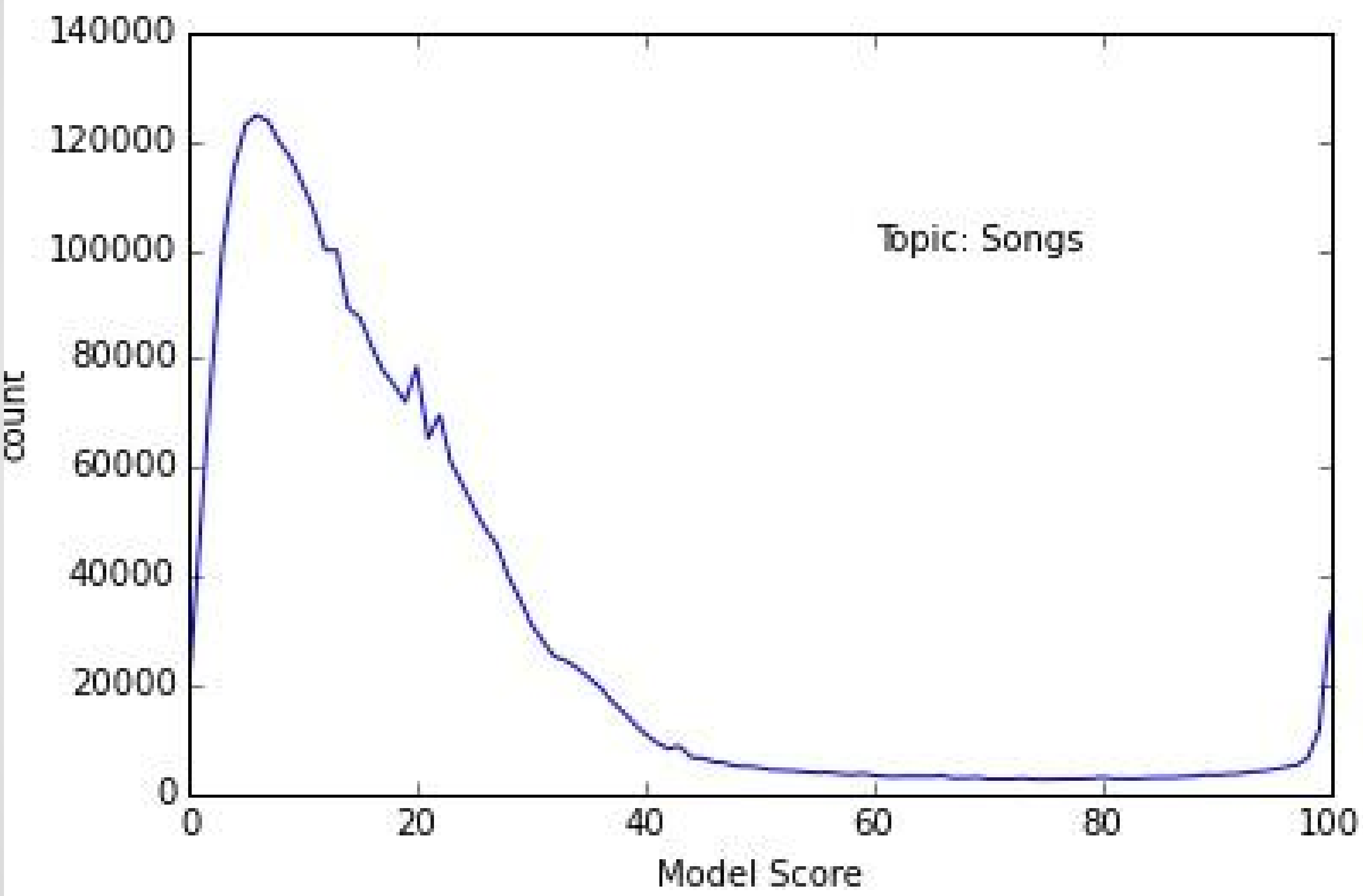


Modest Imbalance ratio 10:1
Full population score distributions predicted from fits on the w2v class distributions
Without proper feature selection even high performing classifier will fail in imbalanced context

Better Label Imbalance Management



Modest Imbalance ratio 10:1
Full population score distributions predicted from fits on the w2v class distributions
Without proper feature selection even high performing classifier will fail in imbalanced context


Topic: Songs
count
Model Score
0
20000
40000
60000
80000
100000
120000
140000
0
20
40
60
80
100

Conclusions

- Pretrained w2v provides a low investment entry to ‘deep’ text classification by circumventing pre-training phase (dAE,RBM)
- Results are competitive in F_1 for highly optimized BOW, and dominate for cases with small training sets
- Ensemble of expert trees helps deal with precision problem at extreme imbalance.
    - Feature selection and well-engineered w2v features avoids washout effects of imbalanced populations
    - Requires far less investment in training examples of boundary cases
    - Enables more efficient scaling for larger space of text class taxonomy

personagraph
an Intertrust Company
Daniel Hansen Ph.D.

galvanize
Mike Tamir Ph.D.
mtamir@galvanize.com

Thank You