Using Doc2Vec for Automated Content Curation

Presented by Chuong Ngo

Thirteenth Labour of Heracles

- Thermodynamics Research Center
  - Provides source data to researchers/industry
  - Accelerates background research; greater productivity
- Manual review of literature
  - Determine relevance & extract key ideas
  - Very labor intensive
- Literature sources growing rapidly

The Road Just Traveled

- 2 topic models
  - LDA & LSI
- 2 corpus sizes
  - 728 & 2357
- 10 classifiers

Table1. Classifiers Evaluated

1. OneR
2. Decision Stump
3. J48
4. Naïve Bayes
5. AdaBoost M1 with Decision Stump
6. AdaBoost M1 with J48
7. AdaBoost M1 with Naïve Bayes
8. LogitBoost with Decision Stump
9. Logistic
10. Simple Logistic

The Road Just Traveled

Table 2. Mean Classifier Results

Classifier	Mean F Score		
	Overall	LDA	LSI
SimpleLogistic	81.37%	81.50%	81.24%
Logistic	80.38%	81.07%	79.70%
AdaBoostM1 with J48	80.03%	80.18%	79.88%
J48	78.48%	79.02%	77.93%
LogitBoost with Decision Stump	78.37%	68.28%	76.58%
AdaBoostM1 with Naïve Bayes	78.29%	78.55%	78.02%
NaiveBayes	77.49%	78.59%	76.39%
AdaBoostM1 with Decision Stump	74.89%	72.66%	77.12%
Decision Stump	73.76%	70.23%	77.29%
OneR	66.49%	59.20%	73.77%

Table 3. Top 5% Classifier Results

Size	TM	Topics	Attrib.	Classifier	Corr, %	F
2357	LDA	100	No	AB/J48	86.3	0.864
2357	LDA	500	No	Si. Log.	85.9	0.860
2357	LSI	100	No	Logistic	85.6	0.858
728	LSI	100	No	Si. Log.	85.3	0.856
2357	LSI	100	No	Si. Log.	85.0	0.852
2357	LSI	500	No	Si. Log.	85.1	0.851
728	LSI	500	Yes	Logistic	84.8	0.851
2357	LDA	100	No	Logistic	84.3	0.849
728	LDA	500	No	Si. Log.	84.9	0.848
2357	LDA	100	No	Si. Log.	84.3	0.848
728	LSI	500	Yes	Si. Log.	84.3	0.846
728	LDA	500	No	AB/J48	84.3	0.844
2357	LDA	500	No	AB/J48	84.2	0.844
728	LSI	500	Yes	AB/J48	84.1	0.843
2357	LDA	500	No	Logistic	83.6	0.840
2357	LSI	500	Yes	Si. Log.	83.9	0.838

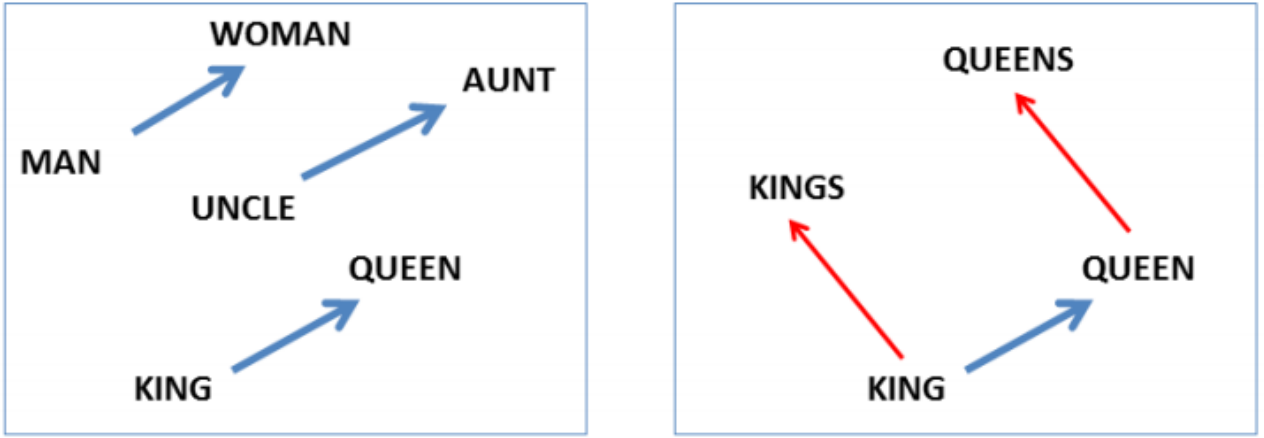
Size = Corpus Size, Attrib = Attribute Filtering, Corr = Correct
AB/J48 = AdaBoost M1/J48, Si. Log. = Simple Logistic

Background

Word2Vec

- Produces vector representations of corpus
- Two-layer neural network: shallow learning
  - Skip-gram with negative sampling (SGNS)
  - Continuous bag of words (CBoW)
- Semantic/syntactic relationships represented by distance

Word2Vec



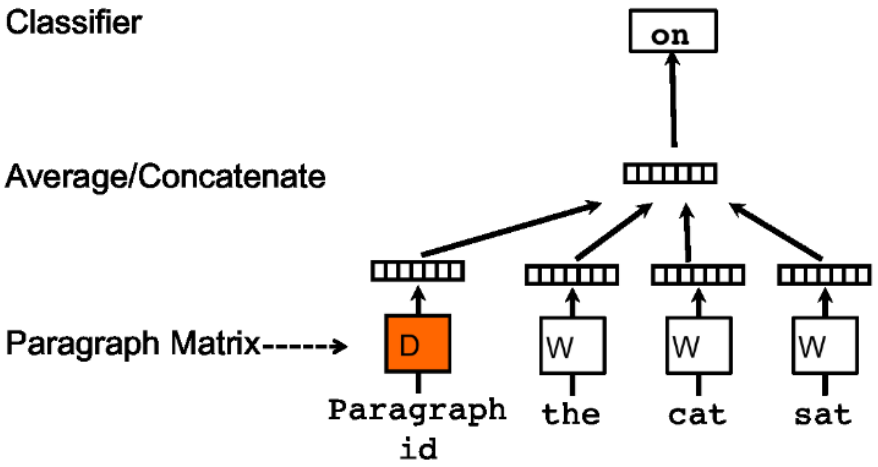
(Mikolov et al., NAACL HLT, 2013)

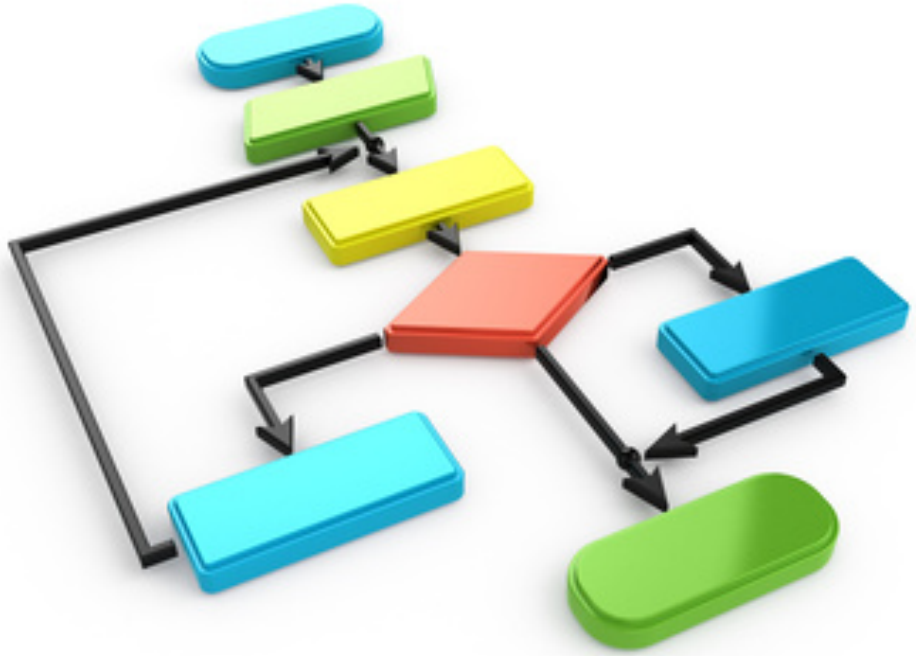
$b \quad a \quad a^* \quad b^*$

$$\text{king} - \text{man} + \text{woman} = \text{queen}$$

Doc2Vec

- Word2Vec w/ labeled token groups
- Relationships built b/t tokens and labels





Methodology

Methodology Overview

- Corpus pre-processing
- Train doc2vec model
- Extract document vectors from model
  - Tag w/ ground truth, interesting or not interesting)
- Train classifier with extracted vectors
- Test the accuracy of the classifiers

Corpus Pre-processing

- NLTK-based pre-processor
  - RegEx tokenizer
  - Porter stemmer
  - Stopword removal
- Alternative techniques
  - Lemmatizer instead of stemmer
  - Removal of irrelevant identifiers (e.g. email addresses)
  - Removal of numbers

Train Model & Label Doc Vectors

- Gensim Implementation
  - Number of features: 300 & 500
  - Negative samples: 0, 5, 10, 20
  - Minimum count: 1
- Retrieve vector for each document in corpus from model
  - Label with ground truth data

Train Classifiers and Test

- Use WEKA implementations
  - 11 Classifiers: 10 previously seen and Random Forest
- Tests using cross-validation and prediction

Train Classifiers and Test

- Use WEKA implementations
  - 11 Classifiers: 10 previously seen and Random Forest
- Test classifiers for accuracies
  - Cross-validation & prediction
  - Prediction uses average word vectors of all tokens and unique tokens

Results

Cross-validation Accuracies

Method	Corpus Size: 728		Corpus Size: 2357	
	300 Model	500 Model	300 Model	500 Model
OneR	56.04%	55.77%	61.43%	59.86%
Naive Bayes	79.12%	78.85%	79.21%	78.96%
Naive Bayes w/ AdaBoost	76.37%	75.27%	80.14%	78.87%
Decision Stump	63.87%	57.97%	62.58%	62.96%
Decision Stump w/ AdaBoost	69.09%	71.98%	70.13%	70.59%
Decision Stump w/ LogitBoost	70.33%	68.13%	72.17%	72.93%
J48	50.27%	65.52%	67.42%	65.29%
J48 w/ AdaBoost	73.35%	73.35%	72.63%	72.72%
Logistic Regression	68.41%	59.48%	79.76%	76.28%
Simple Logistic	76.51%	76.65%	80.36%	79.17%
Random Forest	79.81%	77.75%	80.31%	78.83%

Cross-validation w/ Negative Samples

Method	Accuracy (%)		
	5 Neg. Samples	10 Neg. Samples	20 Neg. Samples
OneR	62.07	57.79	58.29
Naive Bayes	78.96	79.38	79.34
Naive Bayes w/ AdaBoost	78.45	78.36	78.40
Decision Stump	66.74	63.00	64.53
Decision Stump w/ AdaBoost	72.25	71.87	70.51
Decision Stump w/ LogitBoost	72.59	71.96	72.68
J48	65.97	68.18	66.31
J48 w/ AdaBoost	72.30	75.10	73.48
Logistic Regression	77.98	79.00	78.19
Simple Logistic	80.02	79.97	79.47
Random Forest	80.27	80.40	79.97

Cross-validation w/ Negative Samples

Method	Accuracy difference (±%)		
	5 Neg. Samples	10 Neg. Samples	20 Neg. Samples
OneR	+1.04	−5.93	−5.11
Naive Bayes	−0.32	+0.21	+0.16
Naive Bayes w/ AdaBoost	−2.11	−2.22	−2.17
Decision Stump	+6.65	+0.67	+3.12
Decision Stump w/ AdaBoost	+3.02	+2.48	+0.54
Decision Stump w/ LogitBoost	+0.58	−0.29	+0.71
J48	−2.15	+1.13	−1.65
J48 w/ AdaBoost	−0.45	+3.40	+1.17
Logistic Regression	−2.23	−0.95	−1.97
Simple Logistic	−0.42	−0.49	−1.11
Random Forest	−0.05	+0.11	−0.42

Prediction Accuracies

Method	Average Accuracy w/ All Words		Average Accuracy w/ Unique Words Only	
	728 Corpus	2357 Corpus	728 Corpus	2357 Corpus
OneR	50.19%	48.4%	55.45%	44.12%
Naive Bayes	49.72%	49.72%	49.72%	49.72%
Naive Bayes w/ AdaBoost	50.28%	50.28%	50.28%	50.28%
Decision Stump	49.72%	49.72%	49.94%	49.72%
Decision Stump w/ AdaBoost	49.94%	59.65%	50.15%	49.72%
Decision Stump w/ LogitBoost	50.28%	63.13%	50.53%	57.32%
J48	50.28%	46.80%	50.23%	50.02%
J48 w/ AdaBoost	50.28%	50.28%	50.28%	50.11%
Logistic Regression	50.06%	50.28%	49.85%	50.28%
Simple Logistic	50.28%	50.28%	50.28%	50.28%
Random Forest	50.28%	50.28%	50.28%	49.68%

Summary

- Utility of Doc2Vec questionable
- Impact of negative sampling inconclusive
- Doc2Vec presents a challenge when dealing with new documents

Questions?