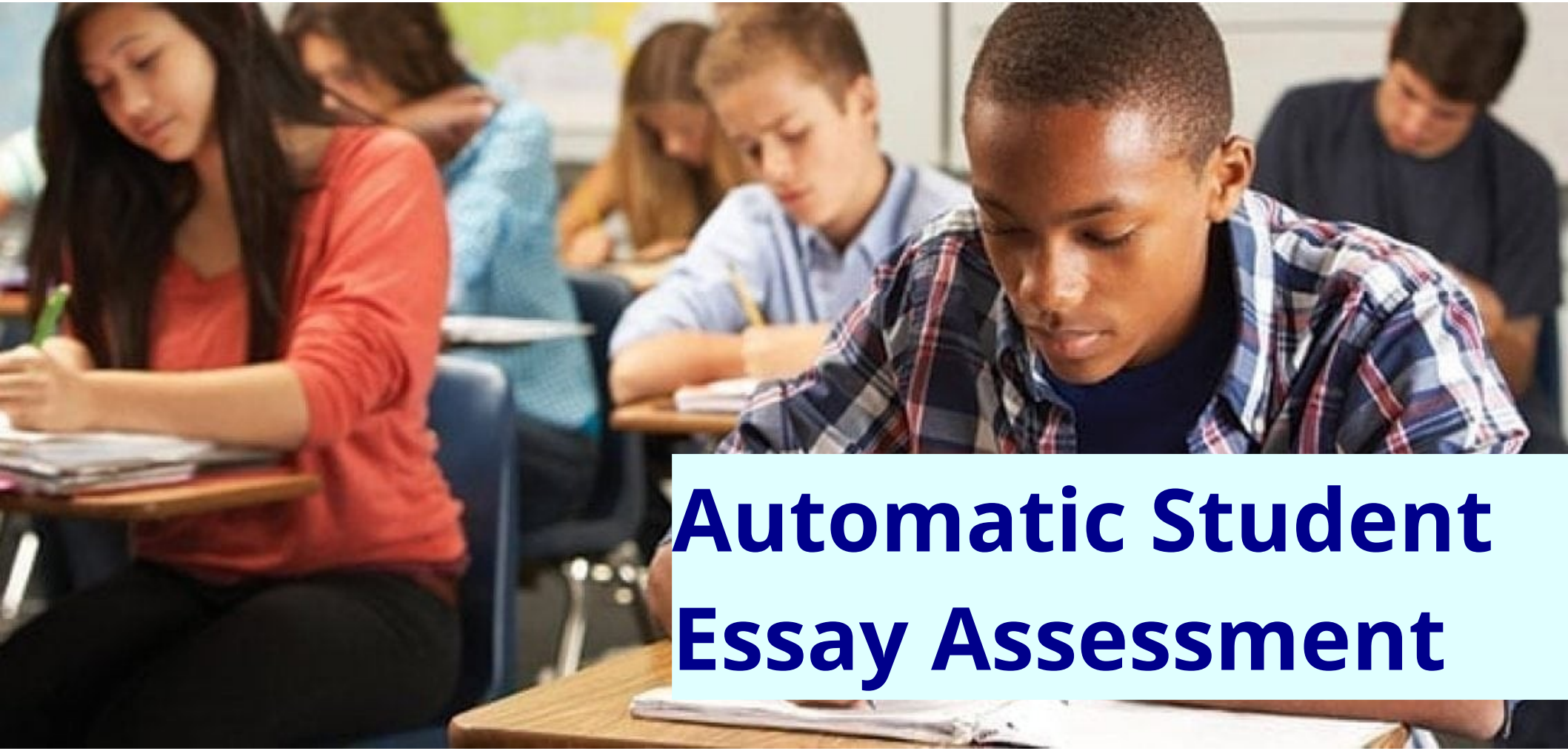


Automatic Student Essay Assessment

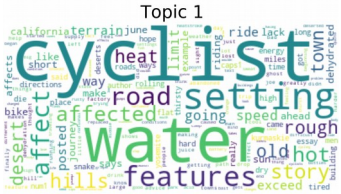
Jeffrey Nicolich, Ph.D.

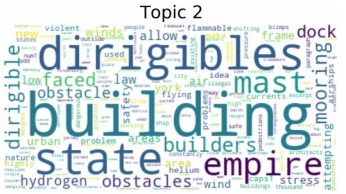
Outline

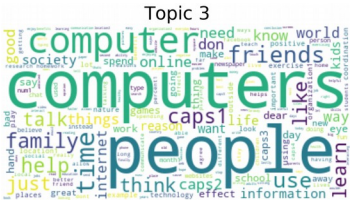
A) Intro/Motivation
B) Grammar and spelling correction
C) Three approaches:
    1. Topic modeling
    2. “Classic” machine learning
    3. Neural networks with word embeddings
D) “Kappa” evaluation metric
E) Summary/Outlook

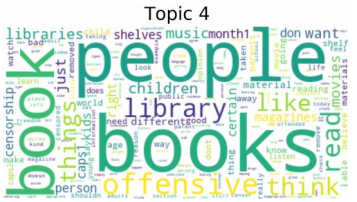
Introduction

- Originally subject of a Kaggle competition in 2012
- Have algorithms/methods improved since then?
- Similar problems have commercial applications today.

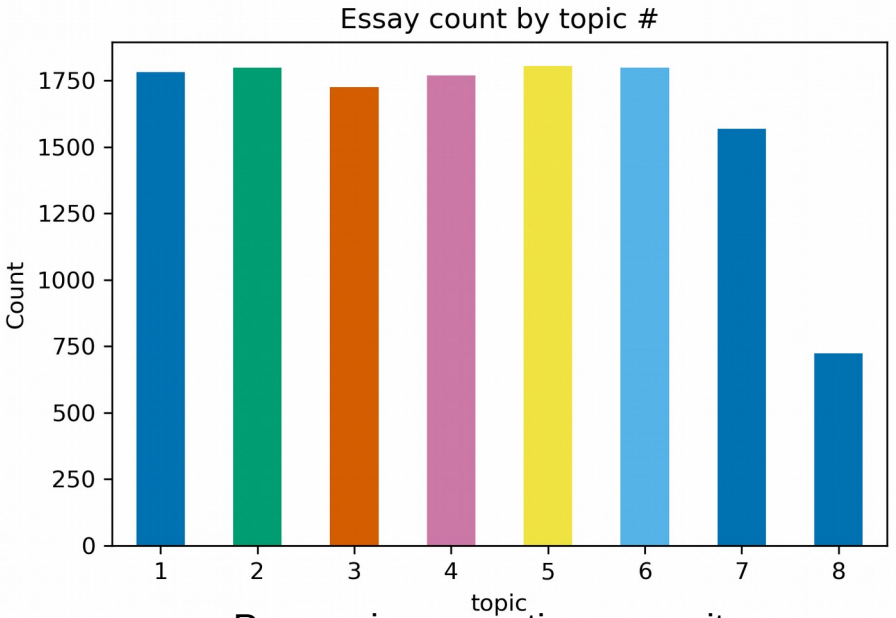




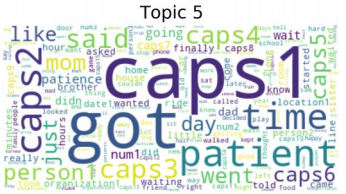


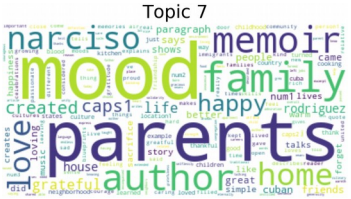


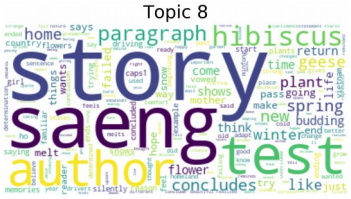
8 Topics
12976 Essays



Persuasive, narrative, expository or source dependent

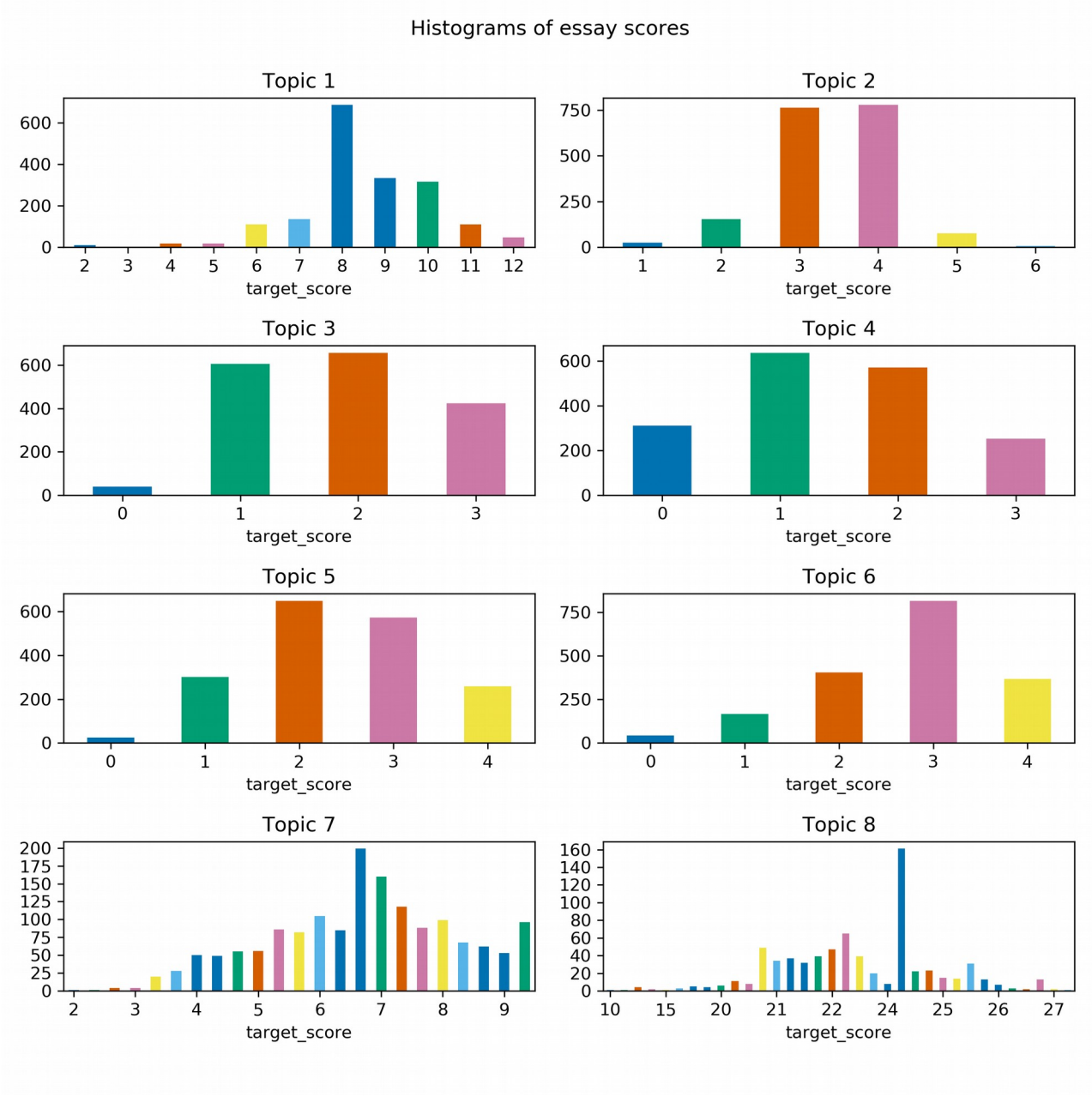






Introduction

Score ranges and distributions are topic dependent



Grammar and Spelling Corrections

- **Why?**
  - Number of errors can be used as feature
  - Better input for NLP processing
- **How?**
  - LanguageTool / language_check python wrapper

Before

'I do think that there should be a censorship in not just in libaries, but everywhare. Personlly, I think that the way that the libraries have the books are appropite and if the parents do not want thier children going any where that is not privy to them keep a hand lenght away As for the parents, the parents know the aera that intrest them ,therefor the parents should go there'

After

'I do think that there should be a censorship in not just in libraries, but everywhere. Personally, I think that the way that the libraries have the books are appropriate and if the parents do not want their children going anywhere that is not privy to them keep a hand length away As for the parents, the parents know the area that interest them,therefor the parents should go there'

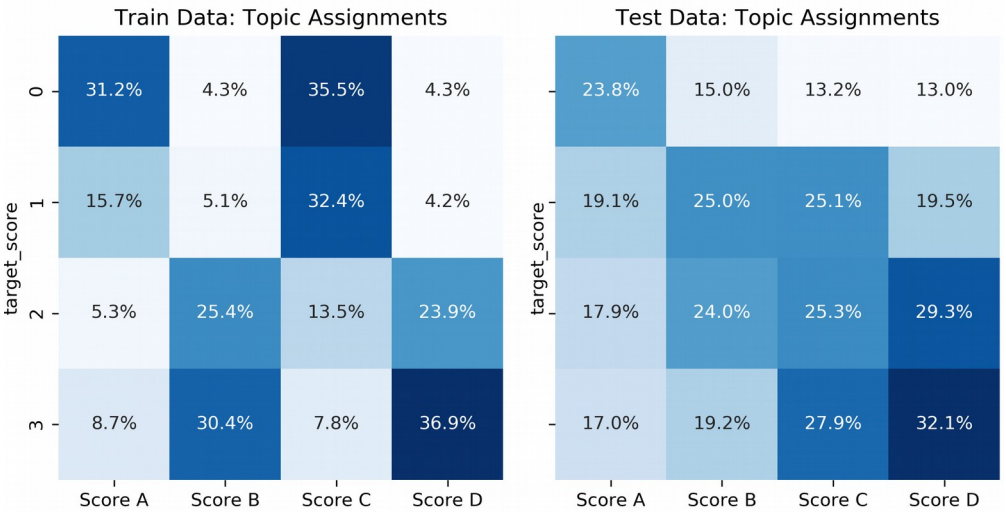
A real example (!!!):

'I aegre waf the evansmant ov tnachnolage. The evansmant ov tnachnolige is being to halp fined a kohar froi alnsas. Tnanchnolage waf ont ot we wod not go to the moon. Tnachnologe evans as we maech at. The people are in tnacholege to the frchr fror the good ov live. Famas invanyor ues tnacholage leki lena orde dvanse and his fling mashine. Tnachologe is the grat'

Topic Modeling

Latent Dirichlet Allocation (LDA) uses word-in-document and word-in-topic probabilities to assign a topic to a given document.

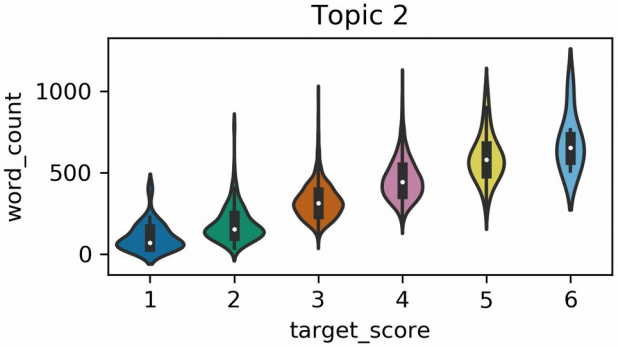
- **Can an essay score be thought of as a topic?**
  - Test on topic 4: Four target scores with reasonably balanced distribution.
  - Training data suggests LDA derived topic “C” is the lowest score and topic “D” is highest score.
  - “A” and “B” are again assigned to lowest and highest scores instead of intermediate scores
  - Test data fails to confirm the training data



Not quite a confusion matrix. True positives are not along the diagonal.

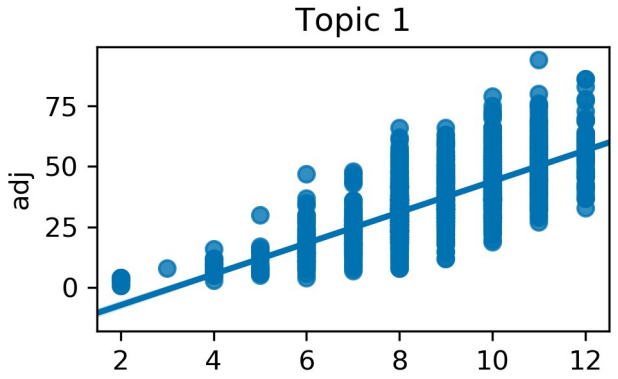
Look for similar color patterns.

Natural Language Processing



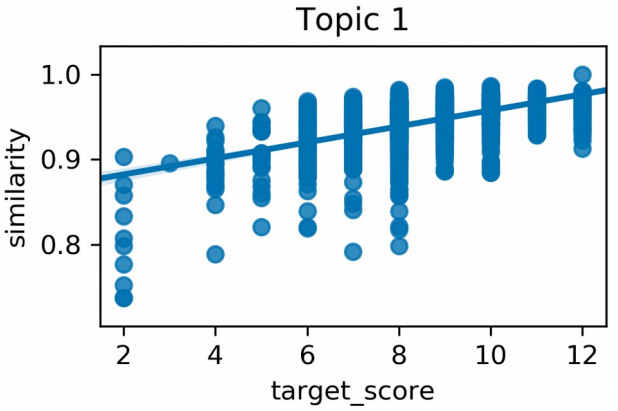
Length-based features:

'word_count', 'token_count', 'unique_token_count', 'nostop_count', 'sent_count'



Parts-of-speech features:

'comma', 'question', 'exclamation', 'quotation', 'organization', 'caps', 'person', 'location', 'money', 'time', 'date', 'percent', 'noun', 'adj', 'pron', 'verb', 'cconj', 'adv', 'det', 'propn', 'num', 'intj'



Other features:

'corrections', 'similarity', 'ner_count'

Natural Language Processing

- Using SpaCy's "small" English model.
- Sample output :

lemma	pos	ner
[dear, @caps1, @caps2, -PRON-, feel, shocked, to, see, that, there, be, people, that, believe, t...	[ADJ, PROPN, PUNCT, PRON, VERB, ADJ, PART, VERB, ADP, ADV, VERB, NOUN, ADJ, VERB, ADP, NOUN, VER...	[Facebook, MySpace, our every day]
[-PRON-, think, the, author, conclude, the, story, with, this, paragraph, because, in, the, spri...	[PRON, VERB, DET, NOUN, VERB, DET, NOUN, ADP, DET, NOUN, ADP, ADP, DET, NOUN, ADJ, ADJ, CCONJ, A...	[winter, the spring time]
[the, obstacle, the, builder, of, the, empire, state, building, face, in, attempt, to, allow, di...	[DET, NOUN, DET, NOUN, ADP, DET, PROPN, PROPN, PROPN, NOUN, ADP, VERB, PART, VERB, NOUN, PART, V...	[the Empire State Building, Second, New York City]

Generate vectorized features

- Univariate feature selection shows little difference between topics

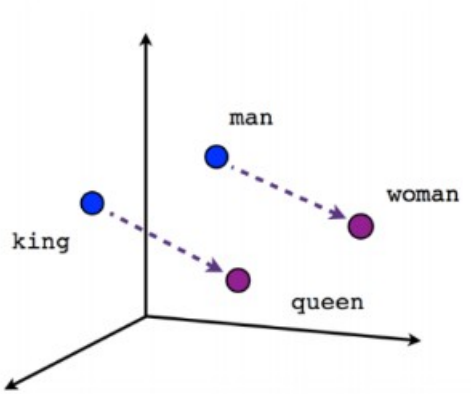
Top 10 features

Topic	1	2	3	4	5	6	7	8
0	similarity	unique_token_count	similarity	unique_token_count	similarity	similarity	similarity	similarity
1	unique_token_count	sent_count	unique_token_count	sent_count	unique_token_count	unique_token_count	unique_token_count	unique_token_count
2	sent_count	comma	sent_count	ner_count	sent_count	sent_count	sent_count	sent_count
3	comma	noun	comma	comma	comma	ner_count	ner_count	comma
4	noun	adj	noun	noun	noun	comma	noun	noun
5	adj	verb	adj	adj	adj	noun	adj	adj
6	verb	cconj	verb	verb	verb	adj	pron	verb
7	cconj	adv	cconj	cconj	cconj	verb	verb	adv
8	adv	det	adv	adv	adv	adv	adv	det
9	det	part	det	det	det	det	det	part

Word Embeddings/Neural Networks

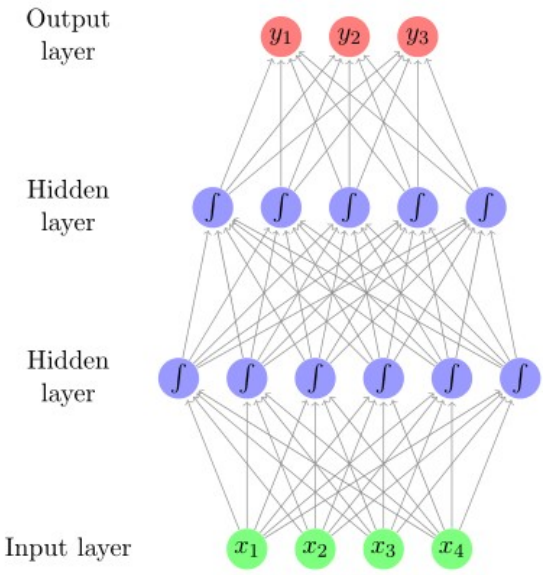
Why word embeddings?

- Efficient representation of words and their context
- Student essays use unique vocabulary and similar context → custom model
- Additional essay sets available to generate custom model (but no target scores)



Why neural networks?

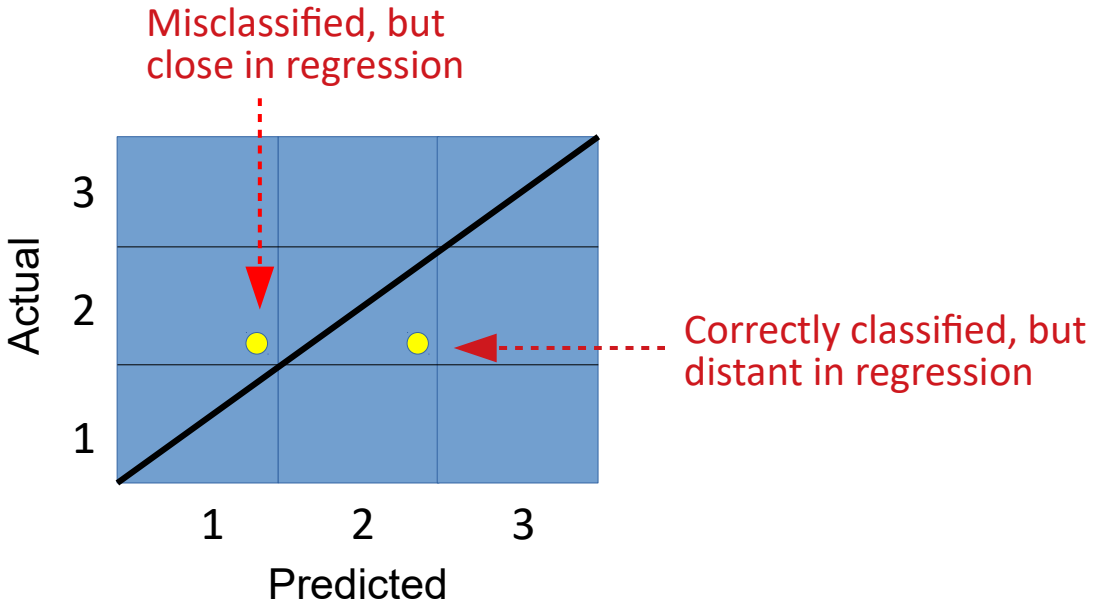
- Neural networks scale well with data dimensions
- Feed-forward network (“MLP”) gave best results
- Convolutional (“CNN”) and recurrent networks (“RNN”/”LSTM”) are able to model context and word sequences and also performed well



Y. Goldberg, JAIR 57 (2016) 345–420

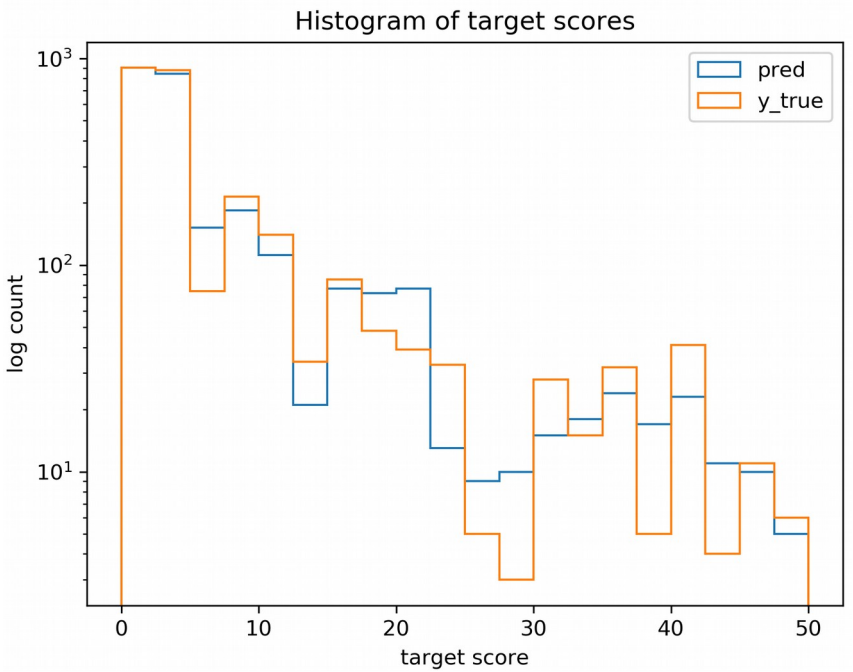
Classification or Regression?

- Primarily a classification problem
- Scores are linearly progressive, therefore regression is possible
- Both classification and regression were attempted
- Regression results were typically better
- Up to 60 imbalanced classes
- Re-balancing not a requirement with kappa metric
- Kappa metric converts classification into quasi-regression



Evaluation Metric: Kappa

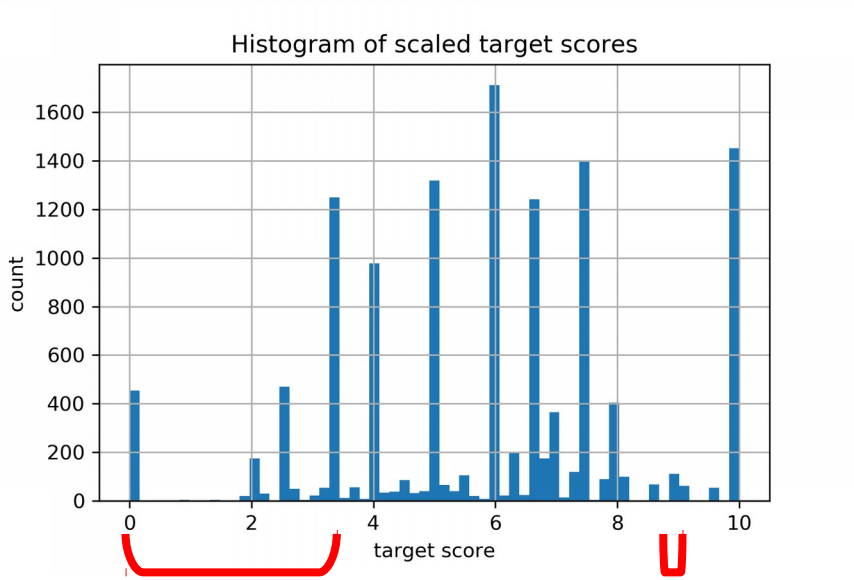
- Kappa score measures the agreement between two raters
- **Why kappa?**
  - Independent of model (i.e. topic)
  - Distance between random choice and perfect agreement
- **What does it mean?**
  - Two human raters achieve kappa of 0.76
  - 76% of the way from random choice to perfect agreement



Kappa = 0.67

Evaluation Metric: Kappa

- Comparison can be misleading:
  - Only original Kaggle competitors had full training data set and validation set available
  - Some kappas are reported on subset of topics
  - Some kappas are calculated on scaled scores
  - Ambiguity about exact calculation method



4 unique scores: Large distance, low kappa

50 scores: Small distance, high kappa

My results:

Best Kappa (not scaled): 0.7844

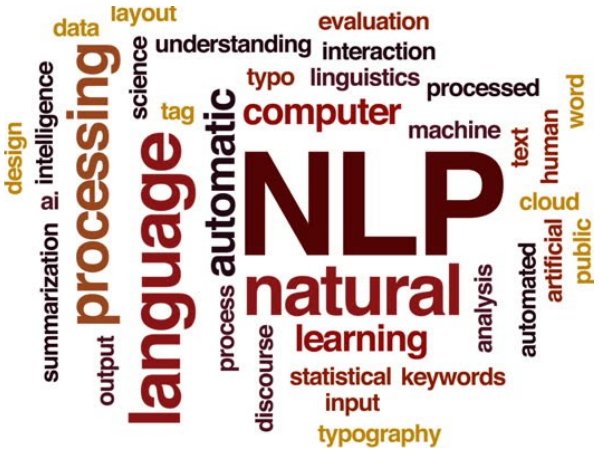
Best Kappa (scaled): 0.9784

Scaled kappas are higher because topics with many scores are given equal weight to topics with few scores

Neural networks outperformed machine learning algorithms

Conclusions/Outlook

- Achieved highest kappa scores (comparing apples to apples)
- Strategy to create better word vectors:
  - Use all available essays (even without scores)
  - Apply grammar and spelling corrections
  - Use Gensim's Word2Vec instead of SpaCy's small model



- Similar problems with commercial impact:
  - Given a set of financial documents, which one should a manager read first?
  - Sentiment analysis on a graded scale, *e.g. very upset - upset - satisfied - happy - very happy.*
  - Optimized brand strategy based on users social media postings.
  - Classification of fake news vs real news.

References

A selection of published work on the Kaggle ASAP data:

- https://www.kaggle.com/c/asap-aes
- https://nlp.stanford.edu/courses/cs224n/2013/reports/song.pdf
- http://aclweb.org/anthology/D/D16/D16-1193.pdf
- https://github.com/vasu5235/Kaggle-Automated-Essay-Checking-System/blob/master/Capstone%20report/capstone_report.pdf
- https://github.com/m-chanakya/AutoEssayGrading/blob/master/papers/paper1.pdf
- http://dspace.bracu.ac.bd/xmlui/bitstream/handle/10361/5399/12101114.pdf?sequence=1&isAllowed=y