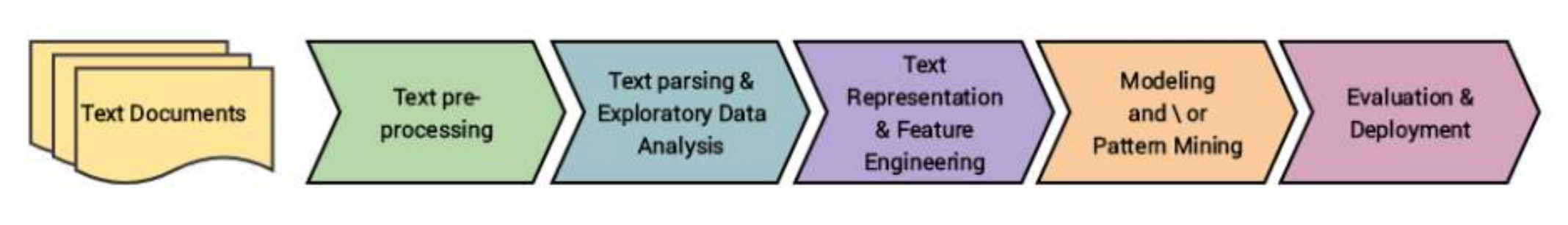
NATURAL LANGUAGE PROCESSING

ANNA NYULUND

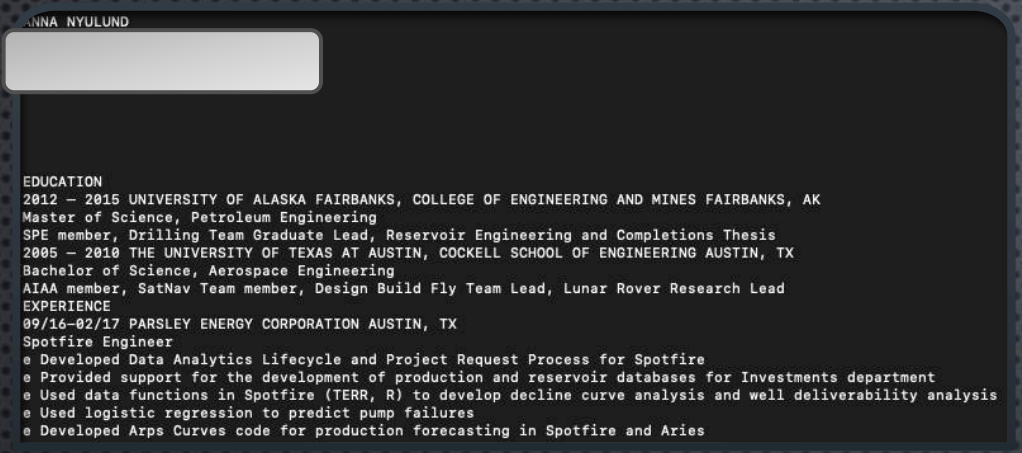
NLP PIPELINE

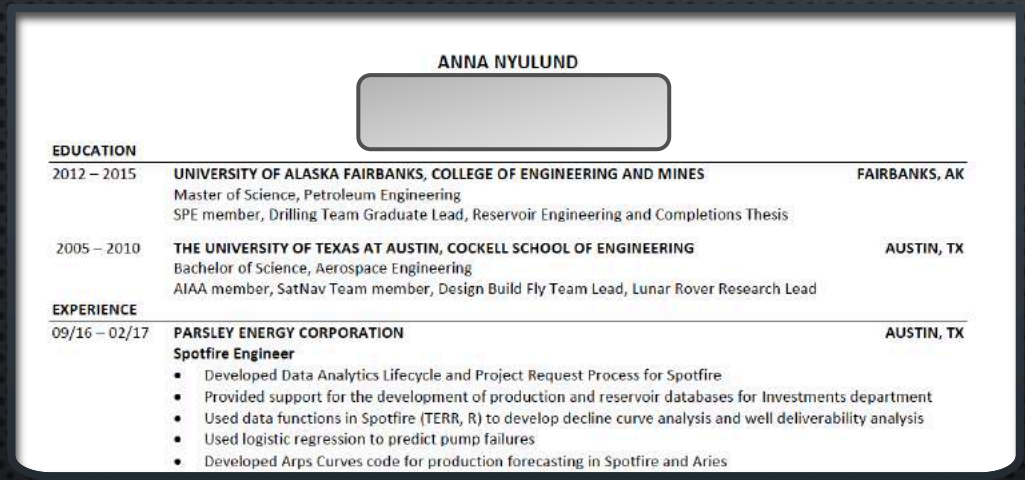


- BASED ON CRISP-DM (DATA SCIENCE WORKFLOWS) 

NLP FRAMEWORKS

NLTK	spacy	gensim	textblob	Flair	Torchtext
Text preprocessing Text classification	Industrial-strength NLP processing Large scale extraction tasks Provides pretrained word vectors Fasted in the world	Topic modeling Similarity analysis Feature engineering	Text processing Phrase extraction Text classification POS tagging Text translation Sentiment analysis	NER Works with Torch	Train/Val/Test Split File Loading Tokenization Vocab Numericalize/Indexify Word Vector Batching Embedding Lookup





TEXT EXTRACTION: PYTHON PACKAGES

- TEXTRACT
- IMAGES: TESSERACT
- PDF: PDF2IMAGE, TESSERACT
- WORD: ANTIWORD, DOCX2TXT
- HTML: BEAUTIFULSOUP

TEXT PRE-PROCESSING

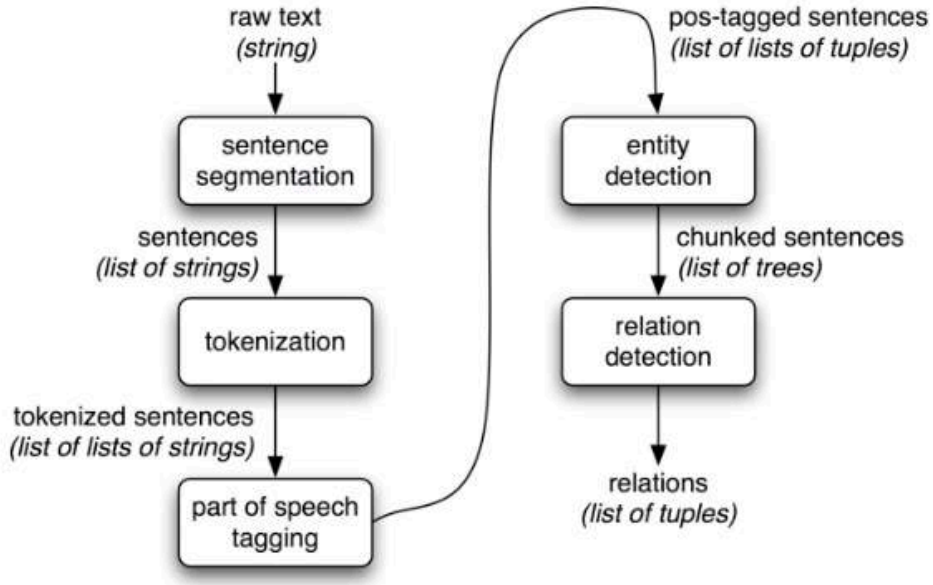
TEXT NORMALIZATION

- REMOVE ACCENTED CHARACTERS (UNICODEDATA)
- EXPAND CONTRACTIONS (CONTRACTIONS.PY)
- REMOVE SPECIAL CHARACTERS
- STEMMING (NLTK)
- LEMMATIZATION (NLTK, SPACY)
- REMOVE STOPWORDS (NLTK)
- CORRECT SPELLINGS (PYENCHANT, ASPELL-PYTHON)

TEXT PARSING AND EXPLORATORY DATA ANALYSIS

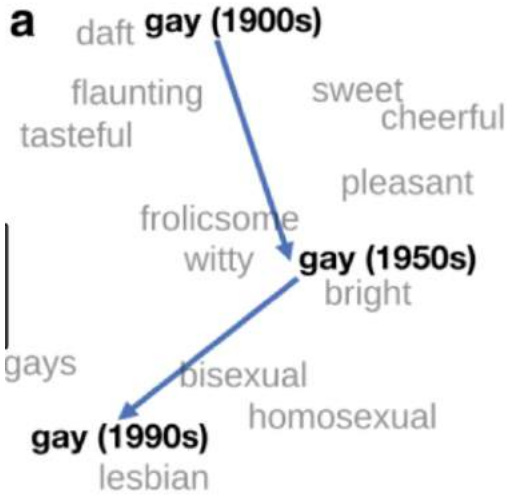
LANGUAGE SYNTAX AND STRUCTURE

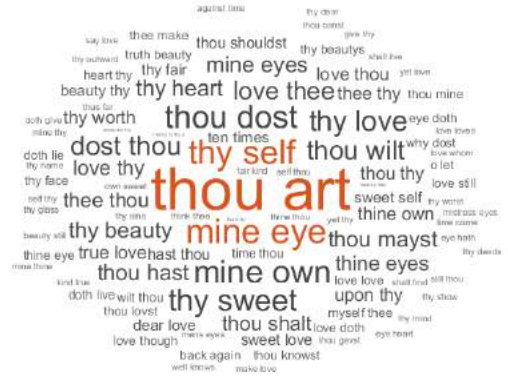
- **PART OF SPEECH (POS) TAGGING (NLTK, SPACY)**
  - NOUNS, VERBS, ADJECTIVES
- **SHALLOW PARSING OR CHUNKING (NLTK, OPEN NLP)**
  - ANALYZES SENTENCE TO IDENTIFY THE CONSTITUENTS
  - DOES NOT SPECIFY THEIR INTERNAL STRUCTURE, NOR THEIR ROLE IN THE SENTENCE
  - GOOD FOR CLASSIFIERS AND TOPIC MODELING
- **CONSTITUENCY PARSING (NLTK)**
  - TREE OF PHRASE STRUCTURE GRAMMAR
  - GOOD IF INTERESTED IN SUBPHRASES
- **DEPENDENCY PARSING (SPACY)**
  - CONNECTS WORDS ACCORDING TO THEIR RELATIONSHIP
  - GOOD TO DETERMINE DEPENDENCY RELATIONSHIPS BETWEEN WORDS

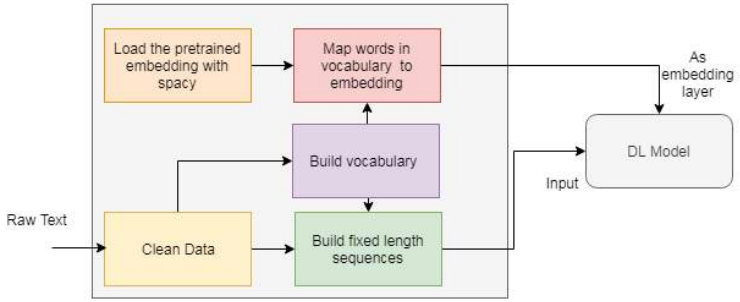


TEXT REPRESENTATION AND FEATURE ANALYSIS

- BAG OF WORDS (SKLEARN)
- BAG OF N-GRAMS (NLTK, ZIP FUNCTION)
- TF-IDF: TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (SKLEARN)
- SIMILARITY FEATURES (SKLEARN)
- TOPIC MODELS (GENSIM)
- WORD2VEC – "PREDICTIVE" (CBOW, SKIP-GRAM) (SPACY, GENSIM, FLAIR)
- GLOVE - COUNT BASED (SPACY, GENSIM, FLAIR)
- FASTTEXT (SPACY, GENSIM, FLAIR)

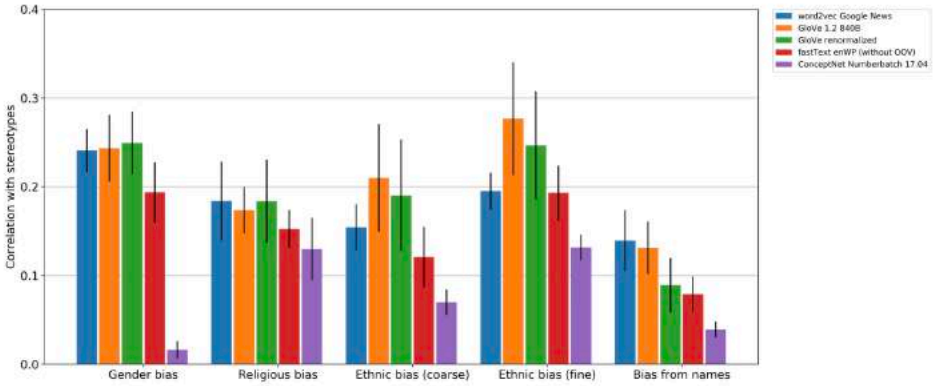




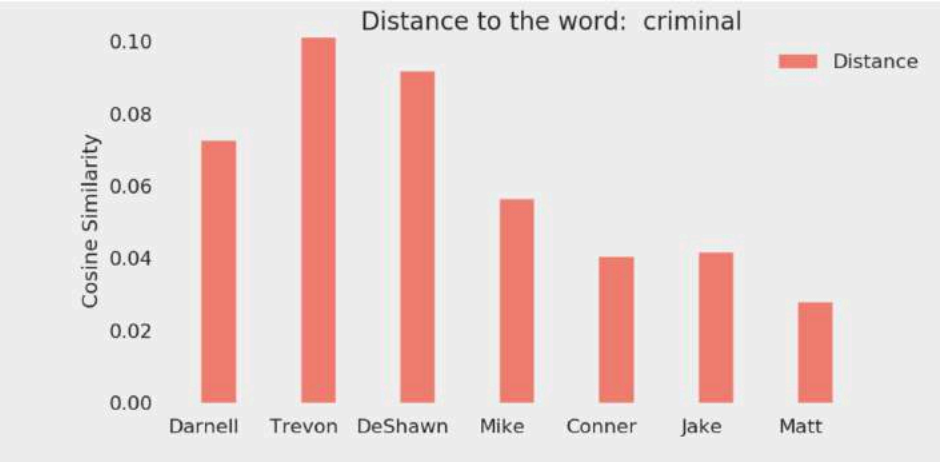


WORD EMBEDDINGS – A WORD OF CAUTION

- GLOVE – TRAINED ON WIKIPEDIA
- WORD2VEC – TRAINED ON GOOGLE NEWS.
- BOTH REFLECT SOCIETAL BIAS
- (KING – MAN) +WOMAN = QUEEN
- (PARIS – FRANCE) + ENGLAND = LONDON
- (PROGRAMMER – MAN) +WOMAN = HOMEMAKER



ConceptNet Bias Graph Comparisons

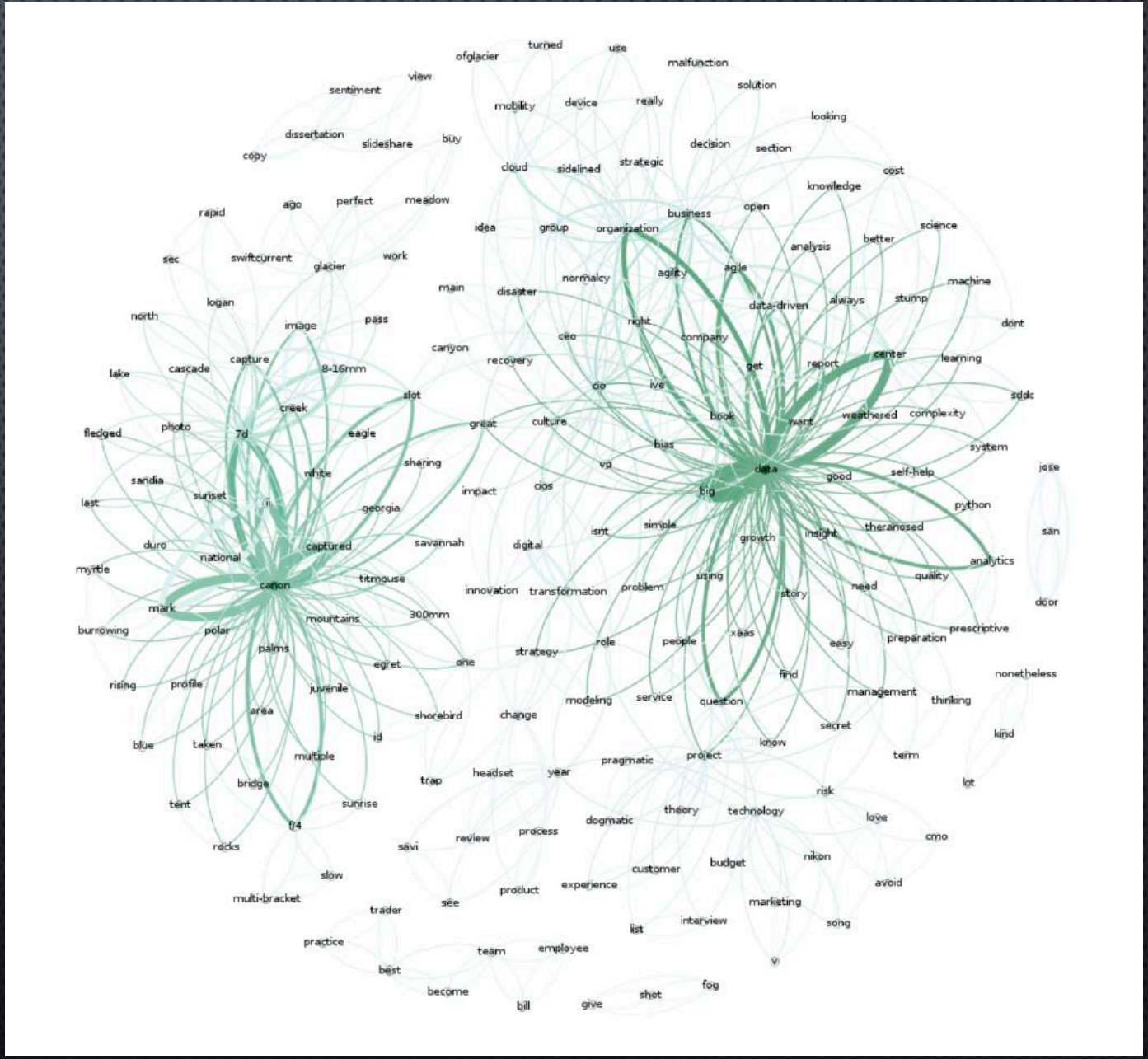


MITIGATING ALGORITHMIC BIAS

- Redirect word embeddings (Bolukbasi, 2016)
- Use different word embeddings
- Synthetic data/ different data
- AI Fairness 360 Open Source Toolkit
  - State-of-arts bias mitigation algorithms
  - Metrics that measure individual and group fairness
  - Slack
- Publish tutorials

TEXT ANALYSIS

- TEXT CLASSIFICATION
- TEXT CLUSTERING
- TEXT SUMMARIZATION
- SENTIMENT ANALYSIS
- ENTITY EXTRACTION AND RECOGNITION
- SIMILARITY ANALYSIS AND RELATION MODELING
- RULE-BASED VS DEEP LEARNING (DEPENDS ON THE DOMAIN!)



NAMED ENTITY RECOGNITION (NER)

- Conditional Random Fields (CRF)
- Spacy, Flair (Torch), NLTK
- Train custom NER on top of Spacy
- Human labeling
  - Dataturks
  - Amazon Mechanical Turk

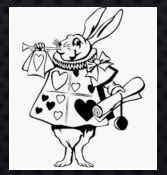
```
Name: [ANNA NYULUND]
Organizaton: []
Location: [Austin, Texas]
Phone or email:
```

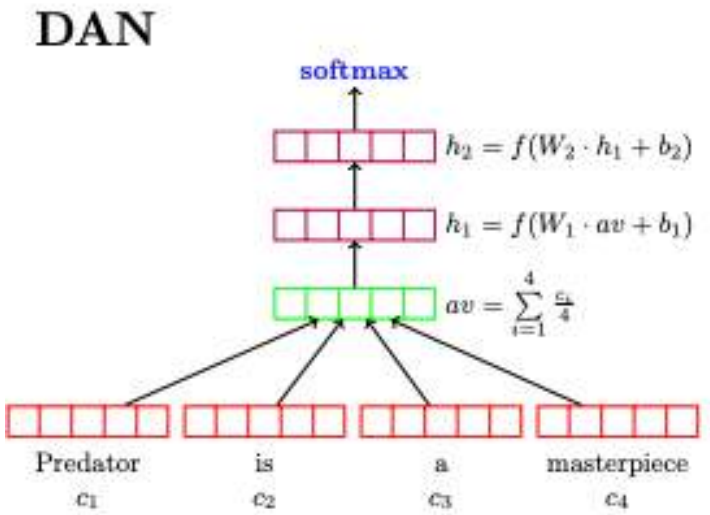
All Names:	0	1
0	ANNA	NYULUND
1	Spotfire	Engineer
2	Iron	Python
3	OKLAHOMA	CITY
4	Data	Scientist
5	Reservoir	Engineer
6	Planit	None
7	Spotfire	None
8	Fekete	RTA
9	Gohfer	None
10	Austin	None

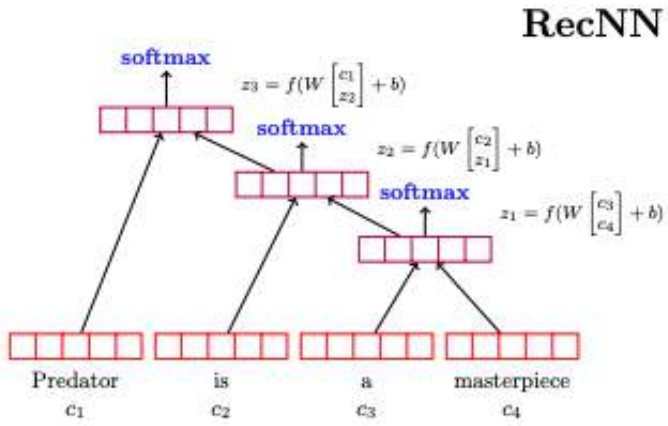
SENTIMENT ANALYSIS DEEP AVERAGING NETWORK (DAN)

- Provides state-of-art accuracies on a variety of sentence and document-level tasks with just minutes of training time and an average laptop computer
- Takes the vector average of the embeddings associated with an input sequence of tokens
- Passes the average through one or more feed forward layers
- Performs classification on the final layer's representation
- Reference: Deep Unordered Composition Rivals Syntactic Methods for Text Classification

- Softmax: takes an input of K real numbers, and normalizes it into a probability distribution consisting of K probabilities.







DEEP AVERAGING NETWORK (DAN) EVALUATION

```
1795 / 224 / 225 train/dev/test examples
Read in 17615 vectors of size 300
=====Train Accuracy=====
Accuracy: 1465 / 1795 = 0.816156
Precision: 1083 / 1308 = 0.827982
Recall: 1083 / 1188 = 0.911616
F1: 0.867788
=====Dev Accuracy=====
Accuracy: 180 / 224 = 0.803571
Precision: 136 / 166 = 0.819277
Recall: 136 / 150 = 0.906667
F1: 0.860759
Time for training and evaluation: 9.98 seconds
```

- Accuracy = TP + TN/TP+FP+FN+TN
- Precision = TP/TP + FP
- Recall = TP/TP + FN
- F1 Score = 2*(Recall*Precision)/(Recall + Precision)

DEEP AVERAGING NETWORK (DAN) RESULTS

- Pretrained on Netflix Movie Reviews
- Used PyTorch
- Ran on oil-and-gas contractor reviews
- Challenge: oil-and-gas and work skill review language: hardworking, problem solver, consistent, driller, etc. and algorithmic bias.
- Label and retrain
- Regularize and redirect word embeddings
- Select different model (BERT – Bidirectional Encoder Representations from Transformers)
- Experiment with different word embeddings

1	Learns fast, great attention to detail. Performs well in high stress situations.
1	Very smart guy , Good on his feet hard worker, Great Employee , Good people , Always make it to work very dependable
0	Consistent performer
0	[redacted] is in my opinion, a go to guy for problems on workovers.
0	[redacted] is one of the hardest workers I've ever had the pleasure to work with.

QUESTIONS? – HINT: YOU CAN GOOGLE IT!!!