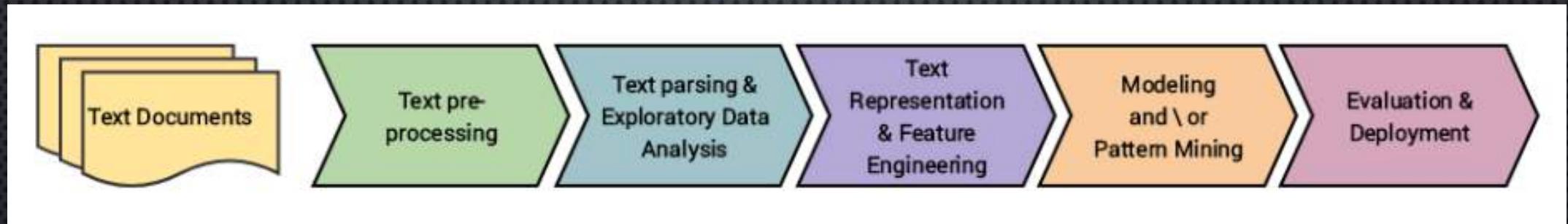


NATURAL LANGUAGE PROCESSING

ANNA NYULUND

NLP PIPELINE



- BASED ON CRISP-DM (DATA SCIENCE WORKFLOWS)



NLP FRAMEWORKS

<u>NLTK</u>	<u>spacy</u>	<u>gensim</u>	<u>textblob</u>	<u>Flair</u>	<u>Torchtext</u>
Text preprocessing Text classification	Industrial-strength NLP processing Large scale extraction tasks Provides pretrained word vectors Fastest 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/Inde xify Word Vector Batching Embedding Lookup

TEXT EXTRACTION: PYTHON PACKAGES

ANNA NYULUND

EDUCATION
2012 – 2015 UNIVERSITY OF ALASKA FAIRBANKS, COLLEGE OF ENGINEERING AND MINES FAIRBANKS, AK
Master of Science, Petroleum Engineering
SPE member, Drilling Team Graduate Lead, Reservoir Engineering and Completions Thesis
2005 – 2010 THE UNIVERSITY OF TEXAS AT AUSTIN, COCKELL SCHOOL OF ENGINEERING AUSTIN, TX
Bachelor of Science, Aerospace Engineering
AIAA member, SatNav Team member, Design Build Fly Team Lead, Lunar Rover Research Lead
EXPERIENCE
09/16–02/17 PARSLEY ENERGY CORPORATION AUSTIN, TX
Spotfire Engineer
• Developed Data Analytics Lifecycle and Project Request Process for Spotfire
• Provided support for the development of production and reservoir databases for Investments department
• Used data functions in Spotfire (TERR, R) to develop decline curve analysis and well deliverability analysis
• Used logistic regression to predict pump failures
• Developed Arps Curves code for production forecasting in Spotfire and Aries

ANNA NYULUND

EDUCATION

2012 – 2015	UNIVERSITY OF ALASKA FAIRBANKS, COLLEGE OF ENGINEERING AND MINES	FAIRBANKS, AK
	Master of Science, Petroleum Engineering SPE member, Drilling Team Graduate Lead, Reservoir Engineering and Completions Thesis	
2005 – 2010	THE UNIVERSITY OF TEXAS AT AUSTIN, COCKELL SCHOOL OF ENGINEERING	AUSTIN, TX
	Bachelor of Science, Aerospace Engineering AIAA member, SatNav Team member, Design Build Fly Team Lead, Lunar Rover Research Lead	

EXPERIENCE

09/16 – 02/17	PARSLEY ENERGY CORPORATION	AUSTIN, TX
	Spotfire Engineer	
	• Developed Data Analytics Lifecycle and Project Request Process for Spotfire • Provided support for the development of production and reservoir databases for Investments department • Used data functions in Spotfire (TERR, R) to develop decline curve analysis and well deliverability analysis • Used logistic regression to predict pump failures • Developed Arps Curves code for production forecasting in Spotfire and Aries	

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

TEXT PRE-PROCESSING

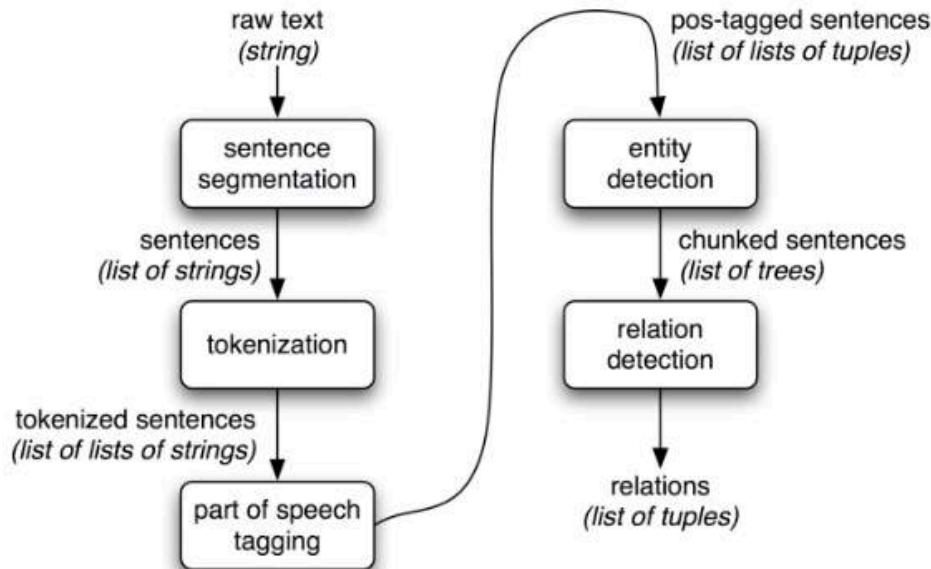
TEXT NORMALIZATION

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

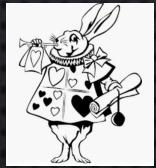
TEXT PARSING AND EXPLORATORY DATA ANALYSIS

LANGUAGE SYNTAX AND STRUCTURE

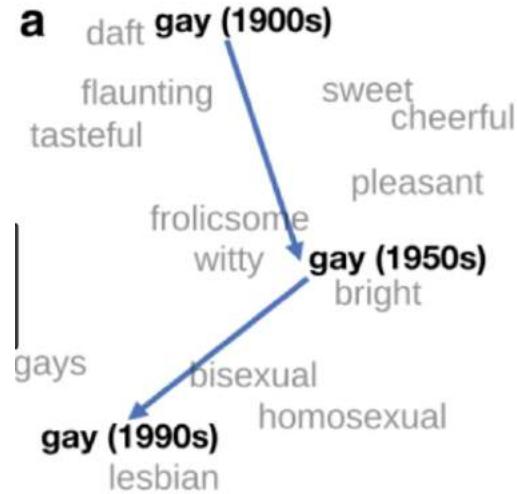
- **PART OF SPEECH (POS) TAGGING (NLTK, SPACY)**
 - NOUNS, VERBS, ADJECTIVES
- **SHALLOW PARSING OR CHUNKING (NLTK, OPEN NLP)**
 - ANALYZES SENTENCE TO IDENTIFY THE CONSTITUENTS
 - DOES NOT SPECIFY THEIR INTERNAL STRUCTURE, NOR THEIR ROLE IN THE SENTENCE
 - GOOD FOR CLASSIFIERS AND TOPIC MODELING
- **CONSTITUENCY 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)**

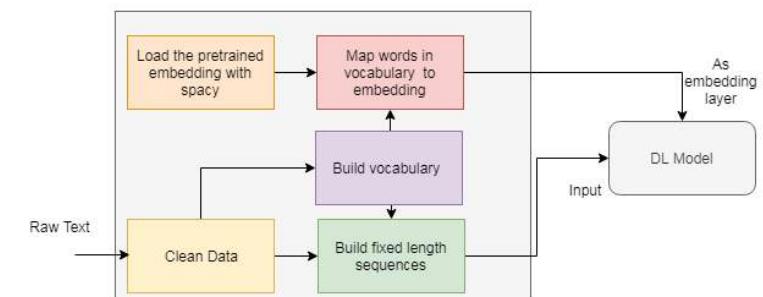


```

the 0.416 0.24988 -0.41242 0.127 0.34527 -0.044557 -0.49688 -0.17862 -0.08966
... 0.813441 0.23882 -0.16890 0.40953 0.33812 0.47796 -0.42852 -0.59641 -0.364
... 0.15164 0.30177 -0.16763 0.16848 0.31719 0.33973 -0.34347 -0.31998 -0.44999
et 0.70953 0.57888 -0.4718 0.16081 0.54861 0.72808 0.16379 -0.51236 0.93031 -0
to 0.60503 0.40601 0.31851 0.27793 0.15902 0.62202 0.16288 -0.47808 -0.3280
and 0.46653 0.13634 0.37071 0.07327 0.11384 0.60139 -0.51312 -0.47168 -0.328
us 0.31842 0.24998 -0.48074 0.10923 0.09372 0.151 -0.55603 -0.47423 0.09238
e 0.21785 0.46215 -0.46757 0.10032 0.0125 0.74445 -0.51104 -0.62626 0.10812 0
s 0.21222 0.48678 -0.20547 0.58846 0.65533 0.22867 -0.41364 -0.23236 0.27428
for 0.19272 0.36161 -0.22568 0.064951 0.13039 0.37075 -0.75874 -0.44722 0.2294
... -0.10704 1.2151 0.45915 0.20538 -0.4285 -0.23312 -0.52222 -1.3357 0.16690 0
that 0.85187 -0.10194 0.13565 0.098686 0.5122 0.49138 -0.47155 -0.30740 0.618
can 0.30945 0.35808 -0.16809 0.1923 0.02692 -0.076486 -0.01393 -0.1074 -0.053
cos 0.61285 0.64274 -0.40552 0.3757 0.74038 0.53739 0.8822339 -0.68577 0.26488 0
was 0.06688 -0.18418 0.24257 -0.33391 0.56731 0.39783 -0.97889 0.01359 +0.6
exists 0.38953 0.22108 0.51887 0.89136 0.17396 -0.27784 -0.84525 -0.25333 0.17386
with 0.72986 0.40968 0.20898 0.26089 0.16071 0.04949 -0.31895 -0.07177 0.017
... 0.42023 -0.06271 0.13715 0.5444 0.5765 0.1671 -0.200 0.16208 0.05264
etc 0.20782 0.37715 -0.10198 -0.23122 0.38175 0.33184 -0.52778 -0.44042 -0.4846
by 0.01381 -0.22072 -0.05088 -0.052967 0.50004 0.34046 -0.33558 -0.19192 -0.05
say 0.39235 -0.15662 0.25798 -0.16811 -0.20738 0.63596 -1.01229 -0.45596 -0.4074
et 0.77722 0.88409 0.064104 0.40013 -1.1687 -0.08522 0.1427 -0.57345
... -0.24978 1.0476 0.21602 0.32778 0.12371 0.2761 0.51184 -1.36 -0.6802 -0.6667
... 0.28534 1.0028 0.14748 0.22267 0.087898 0.23184 0.57782 -0.1787 -0.72415
from 0.42037 0.11242 0.051224 -0.33032 -0.12912 0.22247 -0.9494 -0.18003 -0.26
... 0.83357 0.47532 -0.69706 -0.73861 0.84028 0.63304 -0.75457 0.61247 -0.54

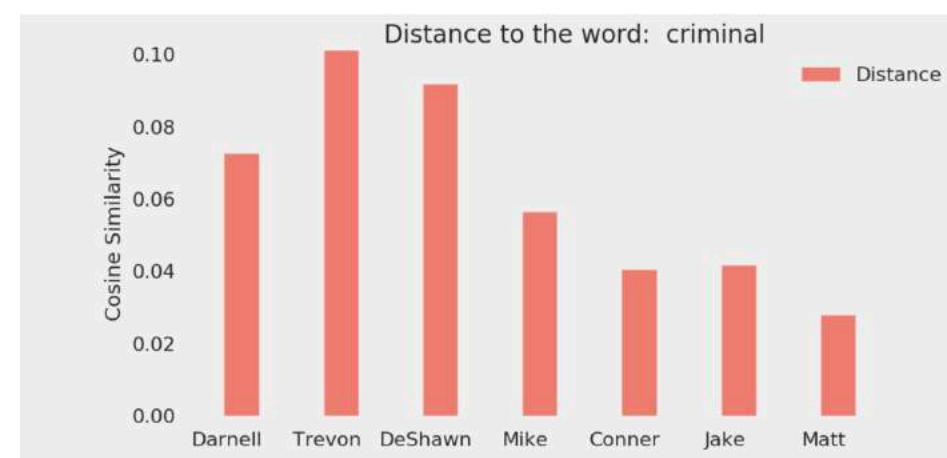
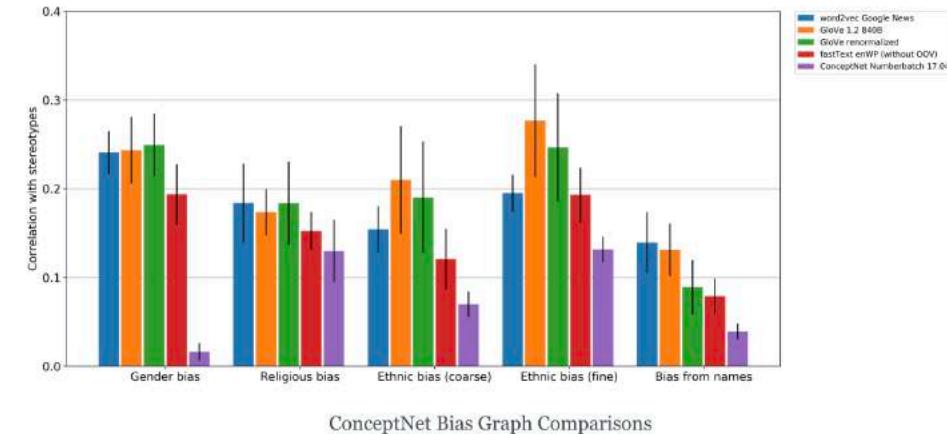
```

against time
my dear
thou const
give thy
thy beautys
skillfull
beauty thy thy heart love theeth thee thy mine
thou
soh grif thy worth
mine the
say love
thee make
thou shouldst
thy beautys
say love
thy heart
thy fair
beauty thy thy heart love theeth thee thy mine
thou
dost thou
ten times
dost thou
thou art
thou dost thy love
eye dath
love dath
thy self
sweet self
thy werm
modest eyes
time name
thine own
thou mayst
eye hath
thine eyes
true love hast thou
time thou
thou hast mine own
love love studed still now
upon thy thy shew
thy sweet
myself thee
dear love
thou shalt
love doth
eye roent
back again
thou knowst
well known
make love



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**



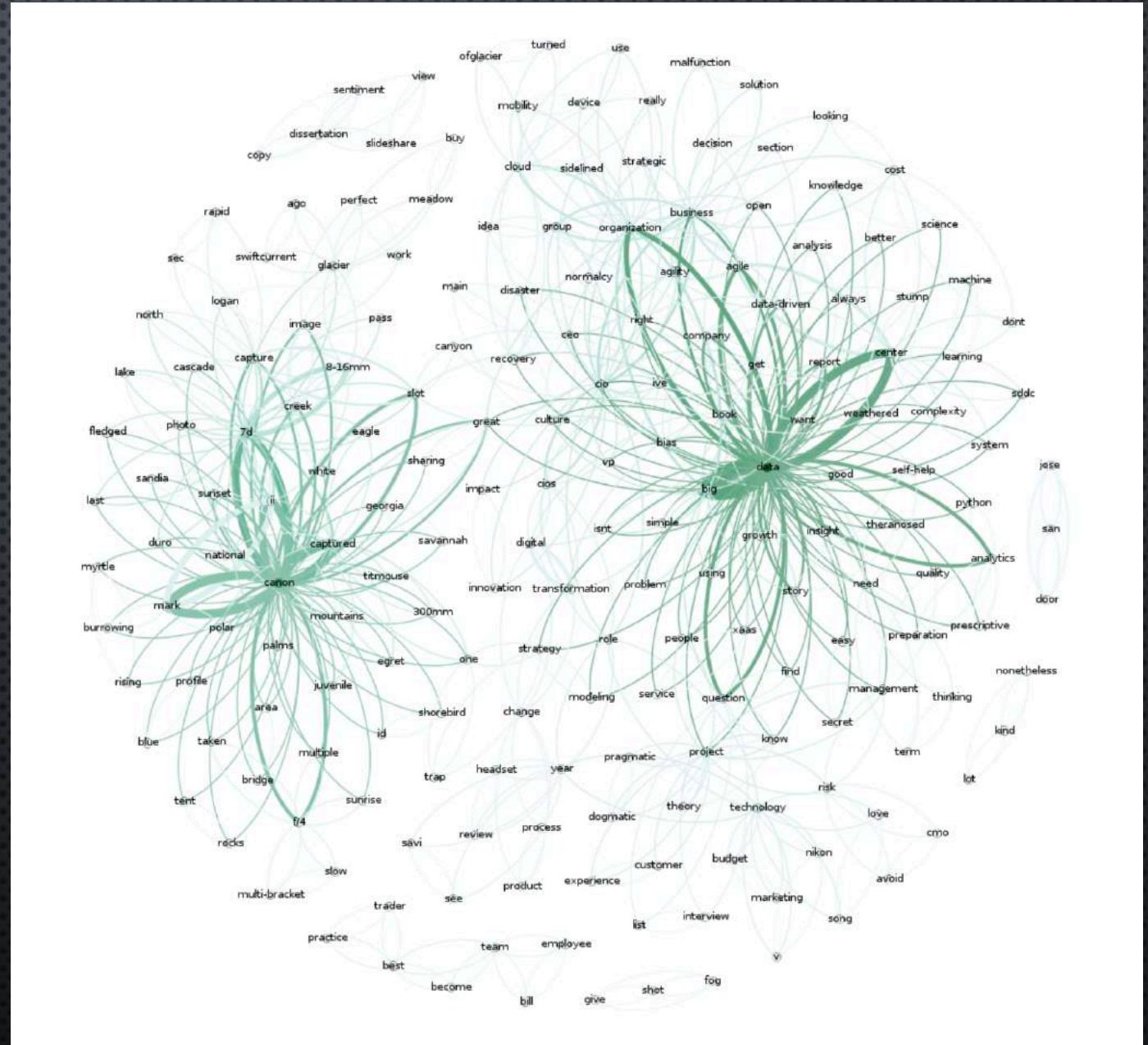


MITIGATING ALGORITHMIC BIAS

- **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:

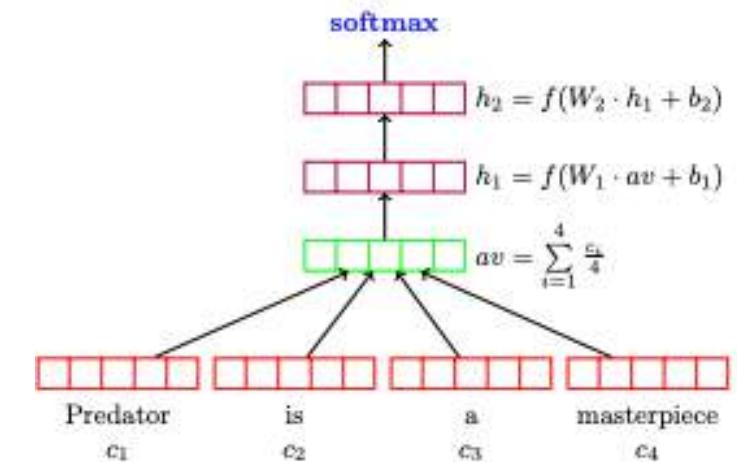
	All Names:	0	1
0	ANNA	NYULUND	
1	Spotfire	Engineer	
2	Iron	Python	
3	OKLAHOMA	CITY	
4	Data	Scientist	
5	Reservoir	Engineer	
6	Planit	None	
7	Spotfire	None	
8	Fekete	RTA	
9	Gohfer	None	
10	Austin	None	

SENTIMENT ANALYSIS DEEP AVERAGING NETWORK (DAN)

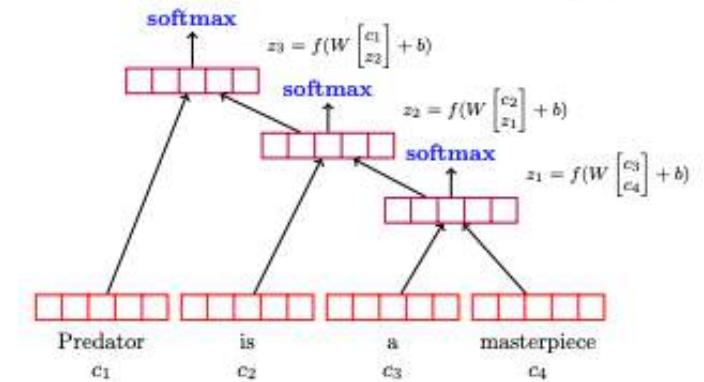
- PROVIDES STATE-OF-ART ACCURACIES ON A VARIETY OF SENTENCE AND DOCUMENT-LEVEL TASKS WITH JUST MINUTES OF TRAINING TIME AND AN AVERAGE LAPTOP COMPUTER
- TAKES THE VECTOR AVERAGE OF THE EMBEDDINGS ASSOCIATED WITH AN INPUT SEQUENCE OF TOKENS
- PASSES THE AVERAGE THROUGH ONE OR MORE FEED FORWARD LAYERS
- 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.



DAN



RecNN



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 = $\frac{TP + TN}{TP+FP+FN+TN}$**
- **PRECISION = $\frac{TP}{TP + FP}$**
- **RECALL = $\frac{TP}{TP + FN}$**
- **F1 SCORE =
 $2 * (\text{RECALL} * \text{PRECISION}) / (\text{RECALL} + \text{PRECISION})$**

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!!!!