COMP 3225

Natural Language Processing

4. Training, Evaluation & Linguistic Resources

Les Carr

lac@soton.ac.uk

University of Southampton

In This Lecture You Will …

- Learn how we train and evaluate language models
- Learn how to train a model for the language context in which it needs to work.
- Learn how to use a variety of metrics to judge the effectiveness of our models against different criteria
- Discover a range of online resources to support a variety of NLP challenges

The material in this lecture is based on various material from Jurafsky and Martin, Speech and Language Processing, 3rd edition (online)

Evaluations

- **Extrinsic evaluation** evaluates the performance of an NLP component (tokenizer, language model *etc*) by embedding it in an application and measuring how much the whole application improves
  - An extrinsic evaluation of two *language models* for *speech recognition* would mean running the speech recognizer twice, once with each language model, and seeing which gives the *more accurate transcription*.
- This is often very expensive and impractical
- **Intrinsic evaluation** measures the quality of an NLP component independent of any application.

Data Splits

- Training Dataset
  - The sample of data used to fit the model. There is no universal guidance for the size of the training data set Frequently 80% training, 10% development, and 10% test
- Validation Dataset (or dev dataset)
  - The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.
- Test Dataset
  - The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.
  - It should be independent of the training set (it should not overlap)
  - It should be unseen, in that none of its data should have been used to develop the algorithm or train its operation

Data Splits

- Classic AI **develops linguistic theories** about how texts that are based on
  - patterns, prescriptive grammars, symbolic rules
- Modern AI **infers statistical patterns** and rules at scale from examining enormous quantities of text
  - e.g. Google 1 Trillion 5-gram corpus
- You can't just know about a text by looking at that text
  - need a sophisticated understanding of the context in which this text appears
  - You have to TRAIN your algorithm on some representative data
  - You can then USE your algorithm to analyse your target data
  - But before that, you must to TEST your algorithm's validity and accuracy on an intermediate set of data, perhaps to make a selection from a variety of alternative algorithms, or a variety of parameters to the same algorithm.
- The training process aims to uncover dependencies and patterns in the data
  - Distinguish between noise and signal
  - Is this appearance of particular words / phrases / sequences / concepts significant
  - Therefore, the training and test data set must be a representative sample of the target data.

Vocabulary Size

Corpus	Tokens = N	Types = $\|V\|$
Shakespeare	884 thousand	31 thousand
Brown corpus	1 million	38 thousand
Switchboard telephone conversations	2.4 million	20 thousand
COCA	440 million	2 million
Google N-grams	1 trillion	13 million

Figure 2.11 Rough numbers of types and tokens for some English language corpora. The largest, the Google N-grams corpus, contains 13 million types, but this count only includes types appearing 40 or more times, so the true number would be much larger.

- Bigger corpora == more varied language == better training data == more word types
  - Relationship between the size of the vocabulary (number of types) and N, the number of tokens, is given by the expression $|V| = kN^{\beta}$ where k and β are +ve constants and $0 < \beta < 1$.
  - β depends on the corpus size and the genre, and ranges from .67 to .75 in fig 2.11.
  - Also significant is number of lemmas *vs* wordform types. The OED lists 615K entries.
- Vocab size is restricted to fit in GPU memory, so |V| usually limited to 30k - 50k
  - large corpus can have V>200k
- Closed vocabulary (all V are in training data) *vs* open vocabulary (unknown words may appear)
  - High out of vocabulary (OOV) rate
  - Map all unknown words (or rare words) to the <UNK> token and proceed
  - Can lead to overfitting problems (limited vocab ignores important patterns with UNK)

Cross-validation

- training / dev / test partitions of available data limits amount of data for representative test
- **k-fold cross-validation** allows data to be used for training and test

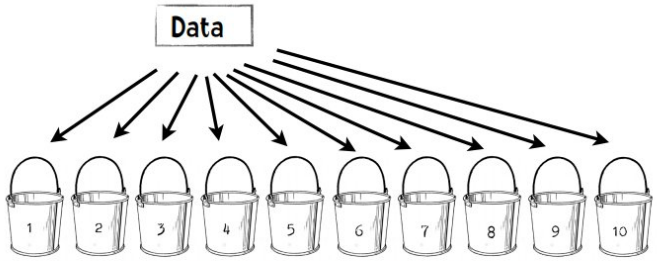


Diagram sourced from KDNuggets.com

- Divide data into k buckets (*k is commonly 10)*
- Use all but one of the buckets for training, and the remaining bucket for testing
  - Repeat this *k* times, but each time choose a different test bucket
  - Calculate average error rate
- **Alternative** (see Fig 4.9, p68) is to use above process for internal dev testing and keep a separate (11^{th}) bucket for final evaluation.
- **Variation** "leave one out cross-validation" when k=n. Each data sample is used in turn for testing

pros:
less biased error measure compared to a single test set

cons:
can be a time-consuming process to use when *n* is large.
can be computationally expensive.

Cross-validation

- **random subsampling** e.g. for large corpus
  - Similar to k-fold: repeat model training/testing k times
  - Each time, randomly choose a proportion of dataset to be test set
- Pros: proportion of the train-test split is not dependent on the number of iterations. Randomisation may also be more robust to selection bias.
- Con: some points may never be selected, or may be selected multiple times.

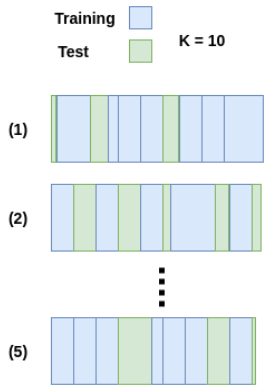


Diagram sourced from towardsdatascience.com

- maximum variance sampling (human choice of samples to represent all classes of text)

Metrics - ROUGE

- used to evaluate text summarization
  - a good machine summary is one that includes many of the same sequences of words as a human-generated (reference) summary
  - ROUGE-n is the proportion of the significant word sequences (n-grams, see lecture 5) in the machine summaries matching those in the reference summaries
    - Variations for N=1, 2, longest common subsequence, etc
  - recall-oriented
    - depends on quantity of material that matches

- ROUGE: A Package for Automatic Evaluation of Summaries, by Chin-Yew Lin (ISI)
  - https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/was2004.pdf

Metrics - BLEU

see chapter 11.8.2

- used to evaluate machine translations
  - a good machine translation is one that includes many of the same sequences of words as a human-generated (reference) translation

- Precision-based
  - how many n-word sequences from the machine-generated set also appear in the reference set for n=1,2,3,4
  - Proportion of all common sequences matching those in the reference set, compared to the total number of sequences

- BLEU ignores recall-based factors (how much of the material did it translate) and focuses only on precision (how much of the material that it translated did it translate well).
  - One solution is to combine both kinds of metric (see F1 score)
  - Instead, BLEU penalizes translations that are shorter than the reference translations

Source

la verdad, cuya madre es la historia, émula del tiempo, depósito de las acciones, testigo de lo pasado, ejemplo y aviso de lo presente, advertencia de lo por venir.

Reference

truth, whose mother is history, rival of time, storehouse of deeds, witness for the past, example and counsel for the present, and warning for the future.

Candidate 1

truth, whose mother is history, voice of time, deposit of actions, witness for the past, example and warning for the present, and warning for the future

Candidate 2

the truth, which mother is the history, émula of the time, deposition of the shares, witness of the past, example and notice of the present, warning of it for coming

Figure 11.16 Intuition for BLEU: One of two candidate translations of a Spanish sentence shares more n-grams, and especially longer n-grams, with the reference human translation.

$$\text{prec}_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} \text{Count}_{\text{match}}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} \text{Count}(n\text{-}gram')}$$

Equation 11.23

Metrics - PERPLEXITY

- used to evaluate language models
  - how good a vocabulary, or a list of word sequences (n-grams), is at "predicting" a target text
  - based on the probability of all the words in the text appearing in that order
    - inverted and normalized by the number of words
- minimizing perplexity is equivalent to maximizing the test set probability according to the language model.

$$\begin{aligned} PP(W) &= \sqrt[N]{P(w_1 w_2 \ldots w_N)^{-1}} \\ &= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N)}} \end{aligned}$$

Equation 3.14

	Unigram	Bigram	Trigram
Perplexity	962	170	109

We trained unigram, bigram, and trigram grammars on 38 million words (including start-of-sentence tokens) from the Wall Street Journal, using a 19,979 word vocabulary. We then computed **the perplexity** of each of these models on a test set of 1.5 million words. The blue table above shows the perplexity of a 1.5 million word WSJ test set according to each of these grammars. (see page 37)

Metrics – Precision, Recall & F1

- used in text classification, searching etc
- **Recall** = proportion of relevant items that were actually selected from the set of all relevant items
  - Items that were classified compared to all the items that should have been classified
- **Precision** = proportion of genuinely relevant items that were selected compared to all the items that were selected
  - Items that were correctly classified compared to all the items that were classified
- Trade-off: easy to have 100% precision but 1% recall or 100% recall and 1% precision

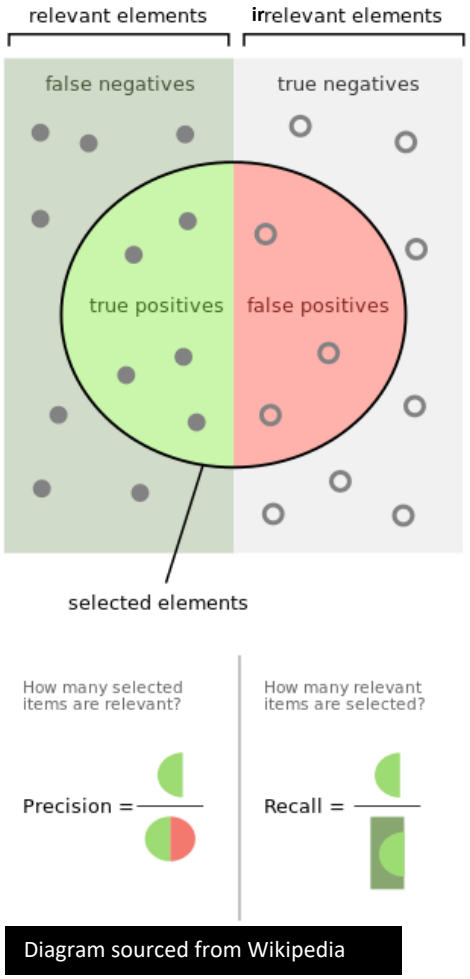


Diagram sourced from Wikipedia

Metrics – Precision, Recall & F1

- **F-measure** is the weighted harmonic mean of precision and recall (see equation 4.18)
  - it can be tuned towards varying balances between precision & recall
  - weighted more heavily to the minimum of the two (conservative measure)
- F_1 balances both factors equally
  - $F_1 = 2PR / (P + R)$

Metrics – Precision, Recall & F1

- Where multiple classes exist, P, R and F scores can be
  - calculated separately for each class and combined (macroaveraging)
  - or caculated once based on pooled data from each class (microaveraging)
- Microaveraging is dominated by frequent classes
- Macroaveraging is a more balanced overview

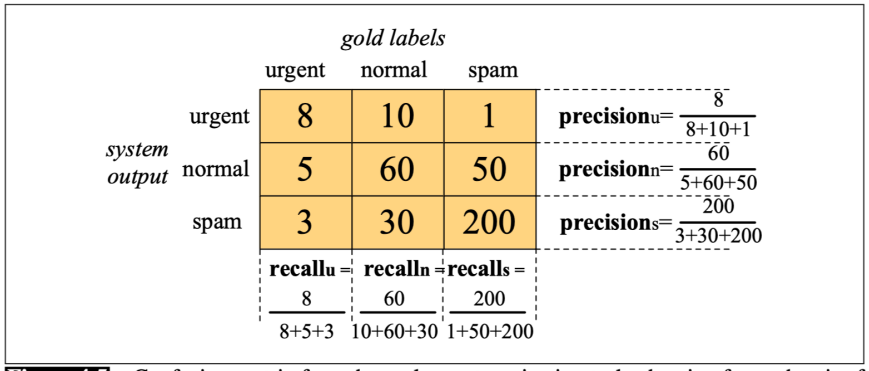


Figure 4.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (in)correctly assigned to c_2

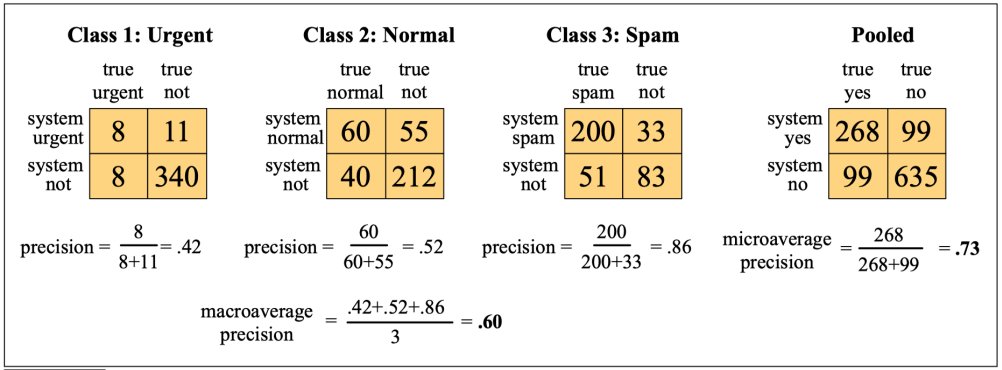


Figure 4.6 Separate confusion matrices for the 3 classes from the previous figure, showing the pooled confusion matrix and the microaveraged and macroaveraged precision.

Break

- Have a browse around the resources listed later in these slides to make yourself familiar with the range

Linguistic Resources (1/7)

https://www.ldc.upenn.edu/

UoS (Stuart) has licensed many NLP text-related LDC datasets available at http://edshare.soton.ac.uk/20520/

- Domain-Specific Hyponym Relations
- TIMIT Acoustic-Phonetic Continuous Speech Corpus
- 2015-2016 CoNLL Shared Task
- Multilingual ATIS
- BOLT Chinese-English Word Alignment and Tagging -- SMS/Chat Training
- BOLT English Treebank - Discussion Forum
- TimeBank 1.2
- Unified Linguistic Annotation Text Collection
- FactBank 1.0
- SemEval-2010 Task 1 OntoNotes English: Coreference Resolution in Multiple Languages
- OntoNotes Release 5.0
- Penn Discourse Treebank Version 3.0
- Treebank-3 (Three "map" files are available in a compressed file)
- 2007 CoNLL Shared Task - Arabic & English
- DEFT Chinese Committed Belief Annotation
- 2009 CoNLL Shared Task Part 1
- Machine Reading Phase 1 NFL Scoring Training Data
- Treebank-3
- 2006 CoNLL Shared Task - Ten Languages
- Phrase Detectives Corpus Version 2
- DEFT English Committed Belief Annotation
- Treebank-3 (Coordination Annotation for the Penn Treebank)
- 2006 CoNLL Shared Task - Arabic & Czech
- DEFT Spanish Committed Belief Annotation
- 2007 CoNLL Shared Task - Basque, Catalan, Czech & Turkish
- 2007 CoNLL Shared Task - Greek, Hungarian & Italian
- 2009 CoNLL Shared Task Part 2
- TIPSTER Complete
- TAC Relation Extraction Dataset
- TAC KBP Entity Discovery and Linking - Comprehensive Evaluation Data 2016-2017

Linguistic Resources (2/7)

- NLTK corpora (free to download via nltk)
  - http://www.nltk.org/howto/corpus.html

Natural Language Analysis with Python NLTK

ACE Named Entity Chunker (Maximum entropy)
Alpino Dutch Treebank
Australian Broadcasting Commission 2006
Averaged Perceptron Tagger
BLLIP Parser: WSJ Model
Brown Corpus
Brown Corpus (TEI XML Version)
C-Span Inaugural Address Corpus
C-Span State of the Union Address Corpus
CESS-CAT Treebank
CESS-ESP Treebank
CONLL 2000 Chunking Corpus
CONLL 2002 Named Entity Recognition Corpus
Chat-80 Data Files
City Database
ComTrans Corpus Sample
Comparative Sentence Dataset
Cross-Framework and Cross-Domain Parser Evaluation Shared Task
Crubadan Corpus
Dependency Parsed Treebank
Dolch Word List
Evaluation data from WMT15
Experimental Data for Question Classification
FrameNet 1.7
Gazeteer Lists
Genesis Corpus
Grammars from NLTK Book
JEITA Public Morphologically Tagged Corpus (in ChaSen format)
KNB Corpus (Annotated blog corpus)
Lin's Dependency Thesaurus
MASC Tagged Corpus
MULTEXT-East 1984 annotated corpus 4.0
Mappings to the Universal Part-of-Speech Tagset
Moses Sample Models
NIST IE-ER DATA SAMPLE
NPS Chat
Names Corpus, Version 1.3 (1994-03-29)
NomBank Corpus 1.0
Non-Breaking Prefixes (Moses Decoder)
Open Multilingual Wordnet
Opinion Lexicon
PC-KIMMO Data Files
PanLex Swadesh Corpora
Paradigm Corpus
Penn Treebank
Penn Treebank Sample
Porter Stemmer Test Files
Prepositional Phrase Attachment Corpus
Problem Report Corpus
Product Reviews (5 Products)
Product Reviews (9 Products)
Project Gutenberg Selections
Proposition Bank Corpus 1.0
Pros and Cons
Punkt Tokenizer Models
RSLP Stemmer (Removedor de Sufixos da Lingua Portuguesa)
SENSEVAL 2 Corpus: Sense Tagged Text
SMULTRON Corpus Sample
Sample European Parliament Proceedings Parallel Corpus
Sample Grammars
SemCor 3.0
Sentence Polarity Dataset v1.0
SentiWordNet
Sentiment Polarity Dataset Version 2.0
Shakespeare XML Corpus Sample
Sinica Treebank Corpus Sample
Snowball Data
Stopwords Corpus
Subjectivity Dataset v1.0
Swadesh Wordlists
Switchboard Corpus Sample
TIMIT Corpus Sample
The Carnegie Mellon Pronouncing Dictionary (0.6)
The Patient Information Leaflet (PIL) Corpus
The Reuters-21578 benchmark corpus, ApteMod version
The monolingual word aligner (Sultan et al. 2015) subset of the Paraphrase Database.
Toolbox Sample Files
Treebank Part of Speech Tagger (Maximum entropy)
Twitter Samples
Unicode Samples
Universal Declaration of Human Rights Corpus
Universal Treebanks Version 2.0
VADER Sentiment Lexicon
VerbNet Lexicon, Version 2.1
VerbNet Lexicon, Version 3.3
Web Text Corpus
Word Lists
Word2Vec Sample
WordNet
WordNet-InfoContent
York-Toronto-Helsinki Parsed Corpus of Old English Prose

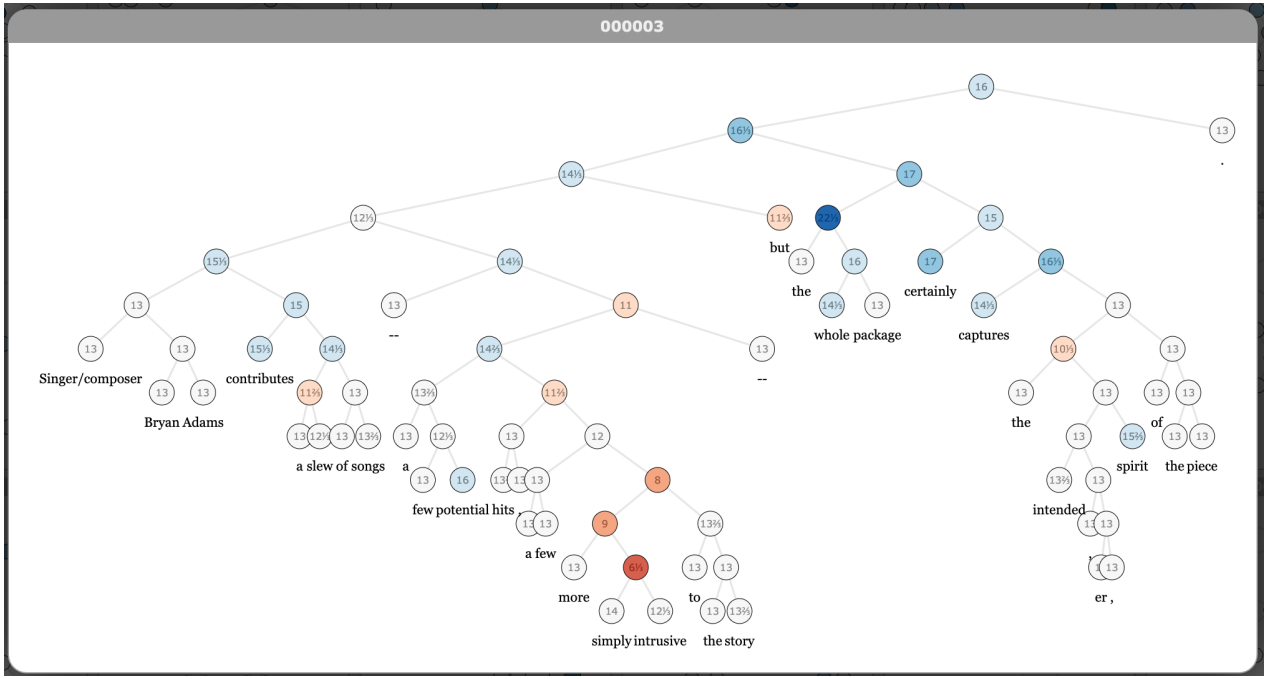
Linguistic Resources (3/7)

- The SIGNLL Conference on Computational Natural Language Learning
  - CoNNL shared task datasets -> (many years covering a lot of topics) https://www.conll.org/previous-tasks

2019	Cross-Framework Meaning Representation Parsing
2017/8	Universal Morphological Reinflection
2017/8	Multilingual Parsing from Raw Text to Universal Dependencies
2016	Multilingual Shallow Discourse Parsing
2015	Shallow Discourse Parsing
2013/4	Grammatical Error Correction
2011/2	Modelling Multilingual Unrestricted Coreference in OntoNotes
2010	Hedge Detection
2009	Syntactic and Semantic Dependencies in Multiple Languages
2008	Joint Parsing of Syntactic and Semantic Dependencies
2007	Dependency Parsing: Multilingual & Domain Adaptation
2006	Multi-Lingual Dependency Parsing
2004/5	Semantic Role Labeling
2002/3	Language-Independent Named Entity Recognition
2001	Clause Identification
2000	Chunking
1999	NP Bracketing

Linguistic Resources (4/7)

- Sentiment analysis datasets (there are many)
  - e.g. Stanford Sentiment Treebank
    https://nlp.stanford.edu/sentiment/treebank.html



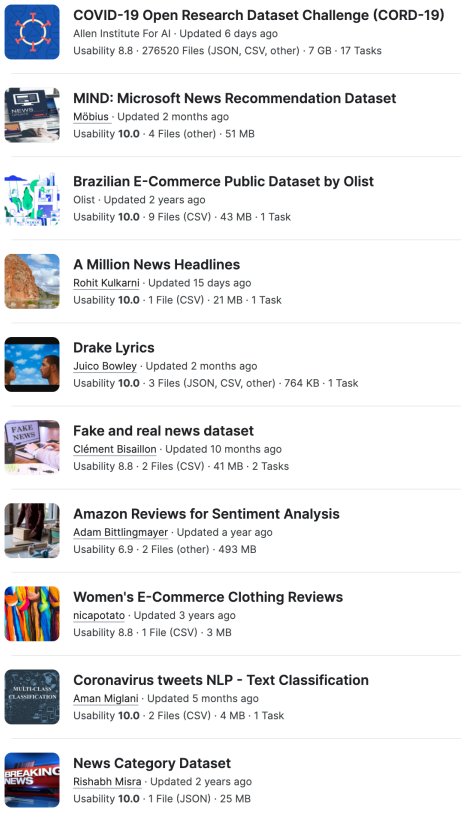
Linguistic Resources (5/7)

- NLU datasets - (there are some)
  - Stanford NLI https://nlp.stanford.edu/projects/snli/

The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels *entailment*, *contradiction*, and *neutral*, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE). We aim for it to serve both as a benchmark for evaluating representational systems for text, especially including those induced by representation learning methods, as well as a resource for developing NLP models of any kind.

Linguistic Resources (6/7)

- Kaggle NPL Datasets (350)
  - https://www.kaggle.com/datasets?tags=13204-NLP



Linguistic Resources (7/7)

- DocNow Catalogue of 132 Twitter ethical datasets totalling > 5bn tweets
  - https://catalog.docnow.io/

Fortune 100 response to 2020 BLM Protests
2020 US Presidential Election
Diego Maradona
SolarWinds
Jessica Krug aka Jess La Bombalera
Twitter Corpus of the #BlackLivesMatter Movement And Counter Protests: 2013 to 2020
Trump Early Primary Tweets
RBG Tweets
Hydroxychloroquine Twitter Dataset
Tropical Storm Imelda Twitter Dataset
Hurricane Dorian Twitter Dataset
John Lewis Twitter Dataset
#metoo Digital Media Collection
WWG1WGA
GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information
Twitter Historical Dataset
116th Congress
Tyendinaga tweet ids
Wet'suwet'en tweet ids
ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks
A Twitter Dataset of 100+ million tweets related to COVID-19
Coronavirus (COVID-19) Tweets
Coronavirus Tweets
#COVID-19
…

Photo by Edwin Andrade on Unsplash

End of Lecture Questions

宏观和微观之间有存在关系，在构建算法时，应该以哪一个为主？
为什么Reducing the overall error rate for an application
thus involves two antagonistic efforts
是对立的？

- Panopto Quiz - 1 minute brainstorm for interactive questions

Please spend **1 minute** using Panopto quiz to wrie down two or three questions that you would like to have answered at the next interactive session.

Do it **right now** while its fresh.

Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.