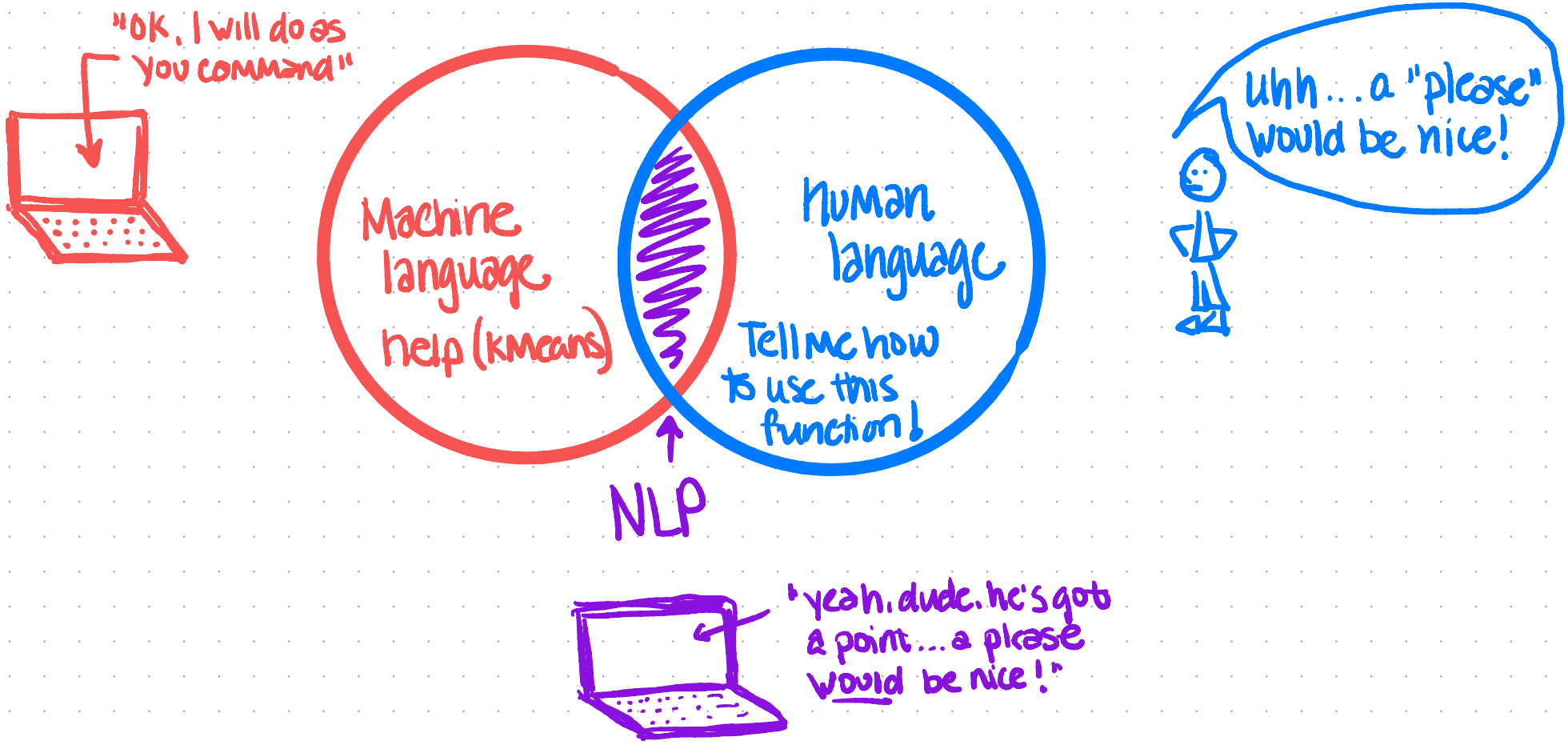
NLP

Natural Language Processing

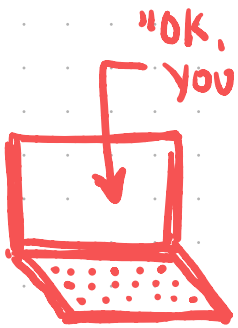
Processing & analyzing large amounts of natural language data via programming.



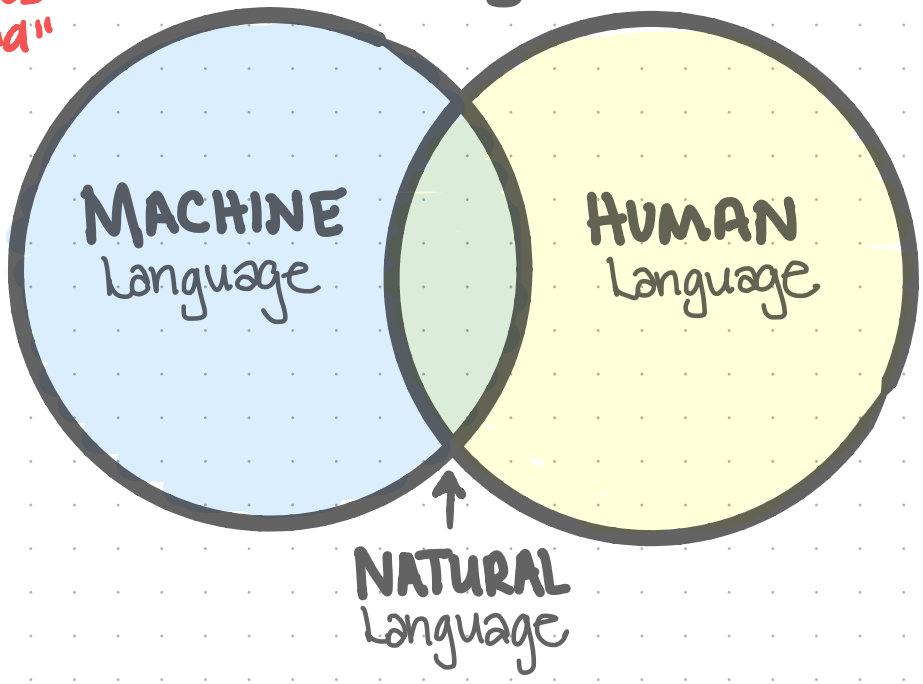
NLP

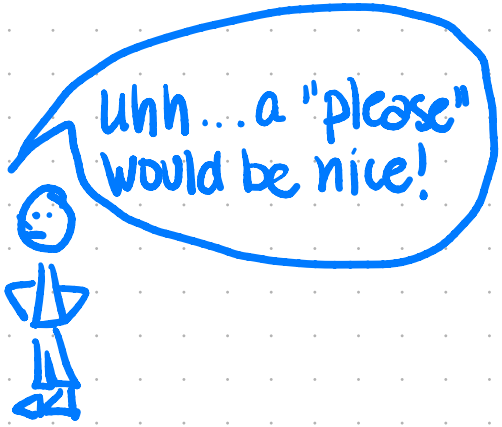
Natural Language Processing

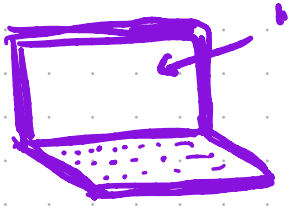
{ Processing & analyzing large amounts of natural language data via programming }



The Venn Diagram of Language







USE CASES for NLP

Voice of the Customer — analysis of surveys, tweets, other feedback
What drives +/- responses? What is being asked for?
What drives engagement?

Semantic Search — a search engine that searches using intent & context rather than just key words.

Knowledge Management & Discovery — tag documents, blog posts, etc. based on the topics for ease of recommending, finding, discovering.

Customer support — Route support requests based on needs / topic identified through the text contained in the request.
Chat bots to solve the common issues before routing to a human.

Security — Phishing emails. national security / terrorism on Social Media, dark web
Cyber bully

Healthcare — New cells of viruses/flu through SM monitoring, Mental health via SM

Virtual Assistants — Siri, Alexa, google home

Translation, Speech to text. competitive intel. Spam

Spell check & Autocorrect, Auto completion,

Types of NLP Programs

supervised | unsupervised

Sentiment Analysis

You're the s**t!
You're full of s**t!
My dog takes a s**t on our walks.

Topic Modeling

"Data Science is fun"

- No label - extract topics (LDA)
- tag = DataScience LABEL!

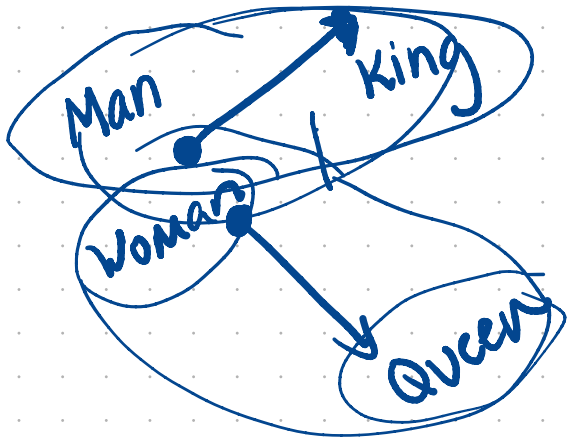
Text Classification

Google's Crowdsource App → Get labeled data
Also: some Rule based using dictionaries

Translation

Word Embeddings → Document Embeddings

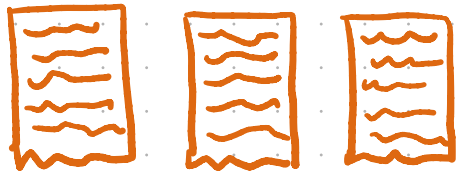
find similar articles, forum question duplication



THE NLP PIPELINE - WRANGLE

Acquire your **CORPUS** of documents

Convert your corpus to individual **DOCUMENTS**



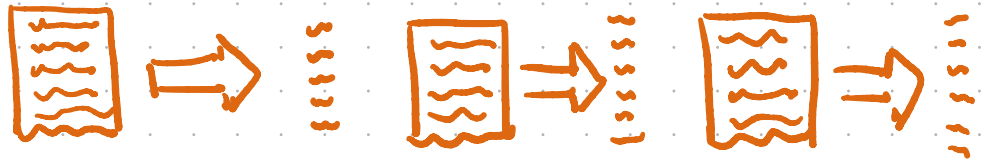
NORMALIZE text

He went to the store to buy 4 rolls of toilet paper. He was really naive to think there would be toilet paper.

↓

he went to the store to buy rolls of toilet paper he was really nave to think there would bee toilet paper

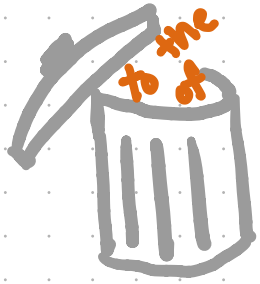
Break documents into **TOKENS**



went → go rolls → roll was → is

Create **STEMS** or **LEMMAS**

Remove **STOPWORDS**



NLP Methods

	PROS	CONS
Rule-based	Explicit Logic, Human Readable	Time to dev, Maintain, doesn't Scale well.
Statistical (e.g. Bayesian)	Automatable, Easy to develop	Doesn't generalize well & not flexible. doesn't capture semantics
Vector Methods (e.g. word2vec)	Uses common and easily accessible ML algorithms	Doesn't capture document structure, or the "Big Picture" well.
Deep Learning / Neural Nets	Good results, little Manual effort.	Black box, Risk of discrimination, biased decisions Made as result, unexplainable, high computational cost.

SOURCE: textician.com

NLP Vocabulary

ENTITY -

CORPUS - Body of docs, entire sample using to analyze

DOCUMENT - single observation

STEM - removing prefixes, suffixes. to the core string.

LEMMA - the base word.

TOKEN - single linguistic unit

CORPUS → DOCUMENT → TOKEN

NLP Stemming vs. Lemmatization

Stemming:

rapply → apply
helpful → help
asked → ask
playing → play
toys → toy

Stemming is fast & simple to use. BUT:

news → New
caring → car
tried → tri

Lemmatization: Uses dictionaries to reduce a word to its base word in order to group all words with the same base together. It requires more computational resources than stemming.

NLP Modeling Techniques

LDA - Latent Dirichlet Allocation - Topic Modeling, Unsupervised

Data Science is a growing field. — 100% Data Science - (Topic A)

Web developers build applications — 100% Web Dev (Topic B)

Codeup offers a web development and data science program.

50% Data Science (Topic A)
50% Web Dev (Topic B)