What Is Language?

- It’s what we humans use to share information

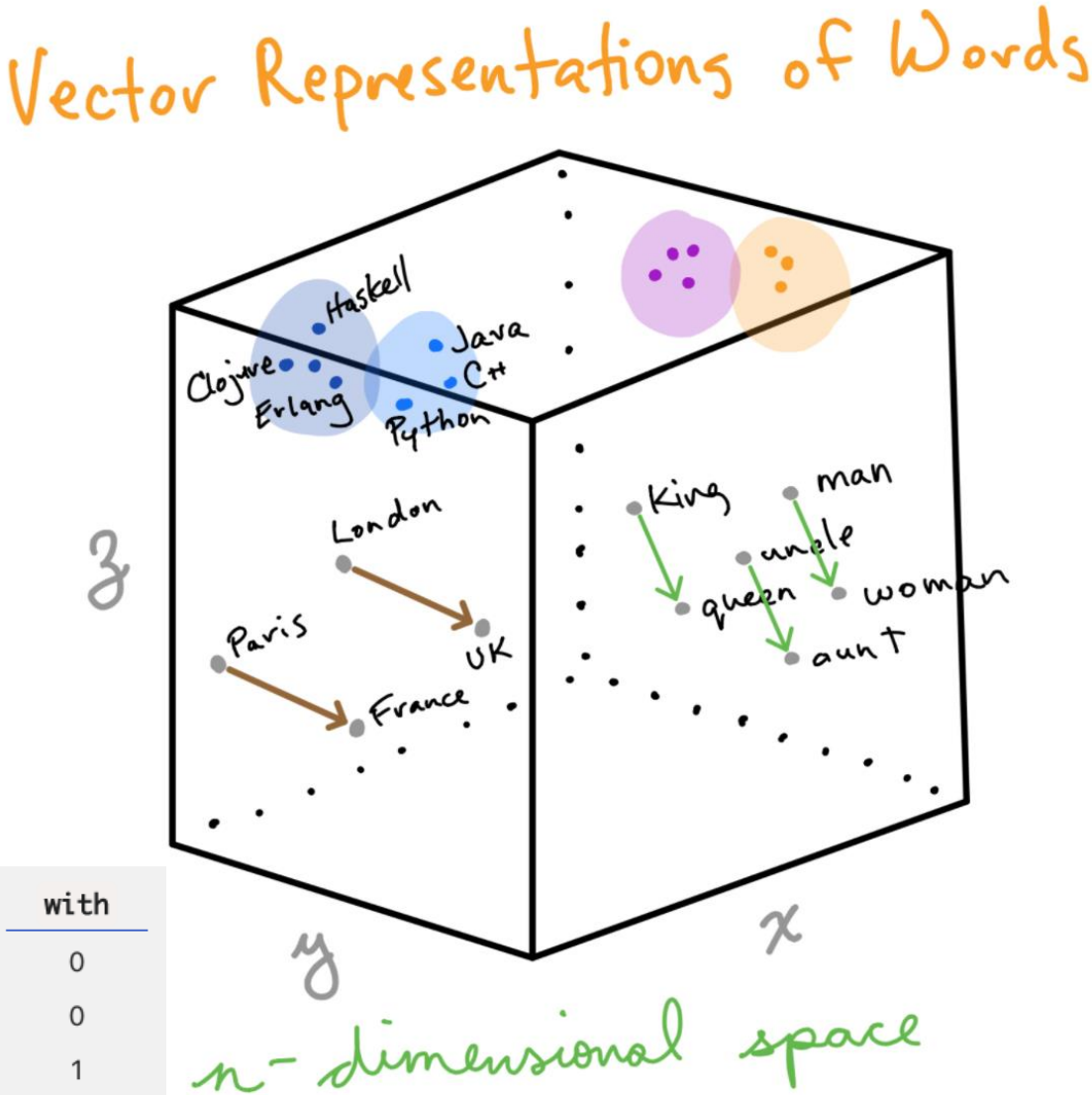
What Is Natural Language Processing?

- Natural Language Processing (NLP) is a field at the intersection of computer science, artificial intelligence, and linguistics. It concerns building systems that can process and understand human language.

- NLP transforms natural language into numbers so it can be searched and computed, taking advantage of all the tools of mathematics and machine learning

- Various methods are available to transform words into vectors of numbers, known as word embeddings. ***Bag of words*** (below) makes a vector of the entire corpus, tracking word frequency in a sentence with a count. It loses word order. **Word2Vec** (right) creates vectors in such a way that words with similar meaning are clustered together, so **"Queen"** can be derived from mathematical operation *King – man + woman*

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1



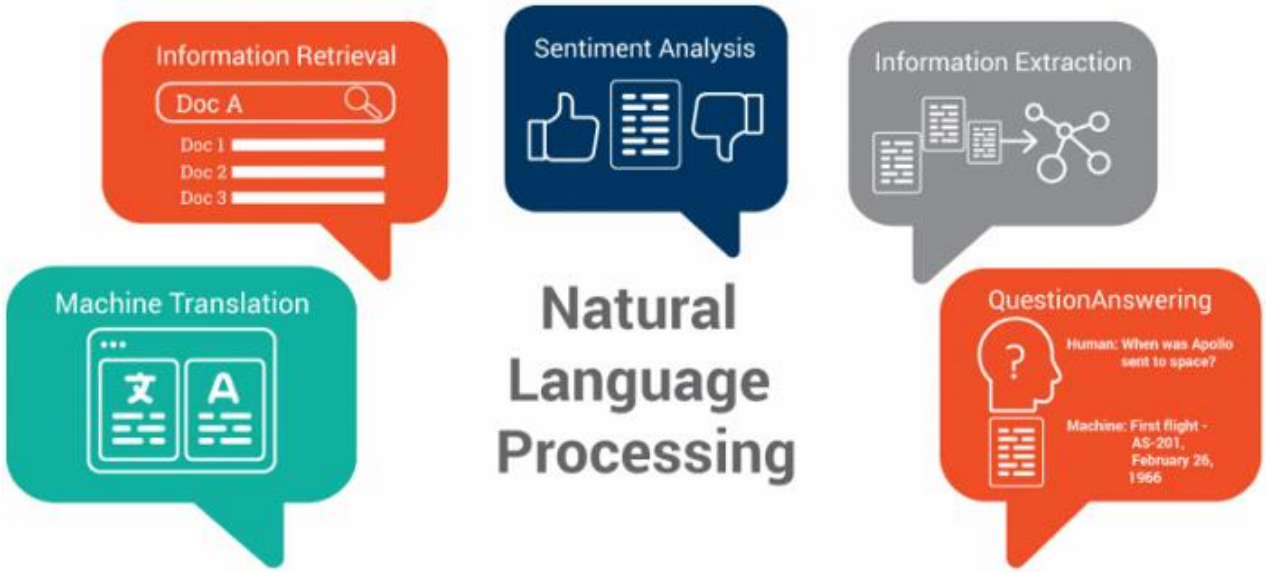
NLP Already Touches Your Everyday Life

- **Search engines** like Google use NLP for tasks including query understanding, question answering, auto correction and completion, information retrieval with ranking and grouping of results

- **Email tools** filter spam and classify mails into primary, social or promotions (Gmail)

- **Smart voice assistants** like Alexa and Siri understand our speech and respond (most of the time) with relevant answers

- **Language Translation** translates words, phrases and web pages between languages



Why is NLP Challenging?

- **Natural language is complex!** A system of components including characters, words and sentences, with building blocks that include phonemes, morphemes and lexemes, syntax and context

- **Word order matters** for context and meaning

- **Ambiguity** (multiple meaning and interpretation)

  examples ”*I made her duck*”, “*The man couldn’t lift his son because he was so heavy*”

- **Creativity** with a huge variety of languages, styles, dialects, genres and variations

- **World knowledge and common sense** that humans have but machines do not

  example “*Dog bit man*” vs “*Man bit dog*”. Which is the more likely?

NLP Applied To Life Sciences

The Challenge and Possible Solutions

- As an example, Covid-19 is associated with a rapidly changing knowledge landscape
- There is an ever-increasing amount of knowledge in the public domain
- It is a challenge to catch and digest the latest developments. Missing potentially important information can affect decision making
- A huge collection of NLP tools, predominately for use with the popular programming language Python, made freely available including from industry experts including Google and Facebook (e.g. BERT and RoBERTa for language modelling)
- Cloud computing makes compute-intense modelling of huge corpus of knowledge possible

NLP Applied To Life Sciences

Start Simple With Heuristics For Analysis And Information Extraction

- A regular expression (regex) is a character pattern used to match and find substrings in text. It is deterministic meaning it either finds a match or not.

- Example regex to identify published papers that reference clinical trial ids

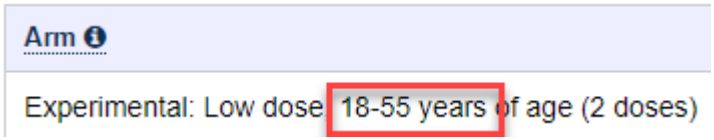
  NCT[0-9]{8}

  EUCTR[0-9]{4}-[0-9]{6}-[0-9]{2}-[A-Z]{2}

MP12, a highly attenuated (by 5-fluorouracil treatment in cell culture) human virus isolate of RVFV (Caplen et al., 1985, Vialat et al., 1997), has recently been tested in a phase II safety/efficacy clinical trial (ClinicalTrials.gov identifier: NCT00415051) to determine if it is safe to give to humans (results not yet published). MP-12 also has potential veterinary applications (Hunter et al., 2002).

- Example regex looking for age group eligibility in arms/intervention information of clinical trial

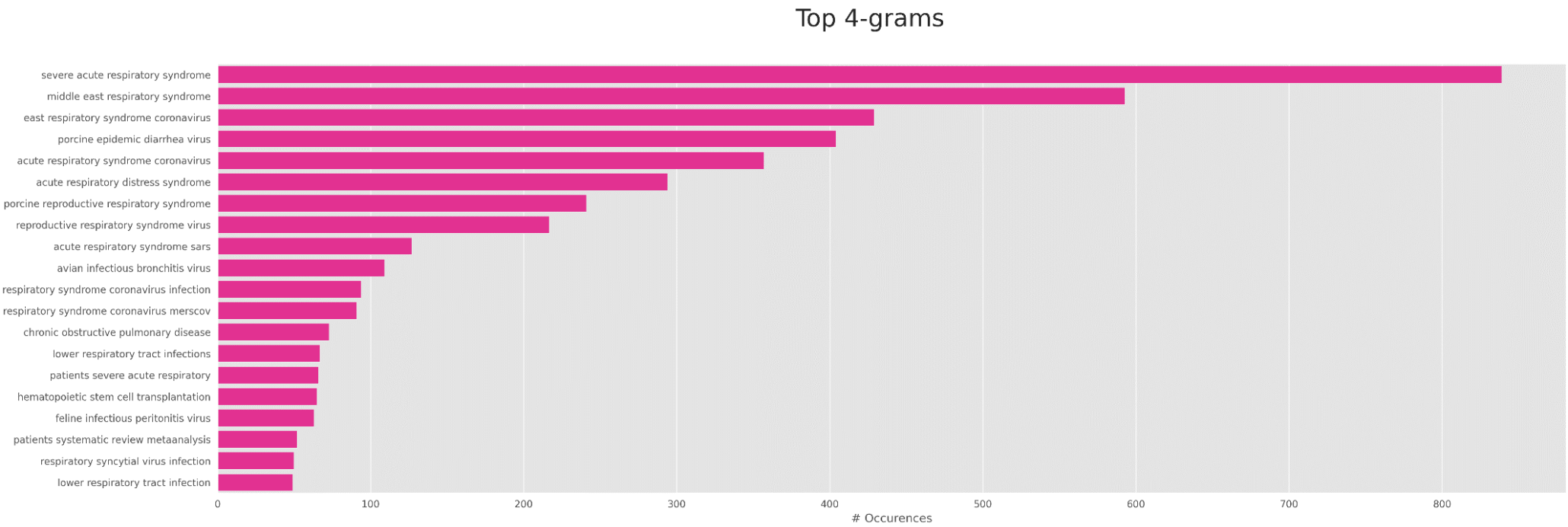
  \d\d-\d\d\syears



NLP Applied To Life Sciences

N-Grams For Text Analysis

- An N-gram is a sequence of words. “Pandemic” is a unigram, “respiratory syndrome” is a bi-gram and so on
- What individual words are used the most? What are the underlying topics?
- Highlights dominant topics and dominant words within a topic

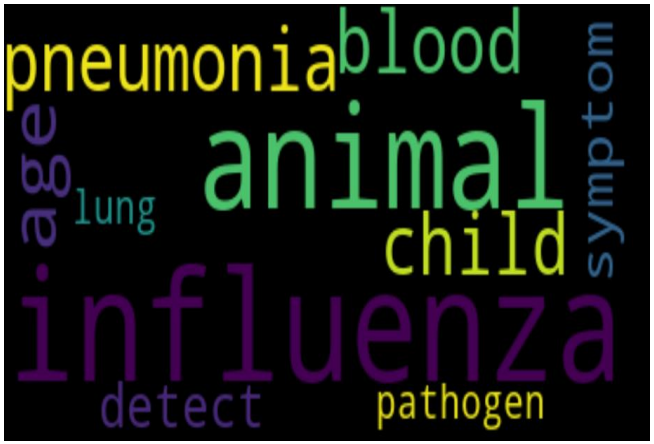


NLP Applied To Life Sciences

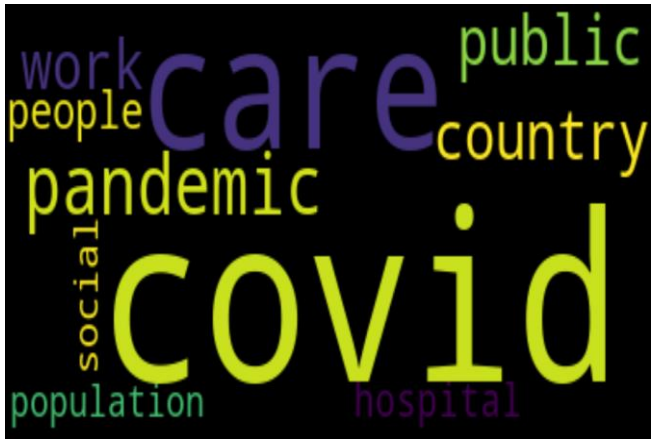
Classify Articles With Topic Modelling

- Organize articles into a set of topics, uncovering the topical structure the document collection
- Uses unsupervised learning methods, such as LDA (Latent Dirichlet Allocation), to find hidden patterns without any ground-truth to learn from
- Highlights dominant topics and significant words within a topic

Topic 1



Topic 2



Topic 3

NLP Applied To Life Sciences

Extract Article Entity Information With Named Entity Recognition

- Useful in providing a summary overview of articles and cluster/tag them by their entities

Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor: molecular mechanisms and potential therapeutic target. SARS-CoV-2 **CORONAVIRUS** has been sequenced [3 **CARDINAL**] . A phylogenetic **EVOLUTION** analysis [3 **CARDINAL** , 4 **CARDINAL**] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS** . There is a diversity of possible intermediate hosts for SARS-CoV-2 **CORONAVIRUS** , including pangolins **WILDLIFE** , but not mice **EUKARYOTE** and rats **EUKARYOTE** [5 **CARDINAL**] . There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS** . Using computer modeling , Xu et al . [6 **CARDINAL**] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE** . SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity to human ACE2 **GENE_OR_GENOME** , based on biochemical interaction studies and crystal structure analysis [7 **CARDINAL**] . SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and ……

NLP Applied To Life Sciences

Finding Relevant Documents With Search Models

- State-of-the art approach is using a BERT (Bi-Directional Encoder Representation from Transformers) model
- BERT learns language and context by encoding the words, the sentence in which the word is located, and the position of the word within the sentence. Adding these 3 vectors together gives an embedding vector that preserves word order and meaning.
- Another benefit is the ability to fine tune BERT models with supervised training on your own dataset e.g. CORD-19 (Covid-19 open source dataset), so it can learn the semantics of a specialized domain
- Can also be applied to other use cases including language translation and summarisation

Question What mutations have been identified in the Receptor Binding Motif (RBM) region of the S-glycoprotein Receptor Binding Domain (RBD) of the SARS-CoV-2 virus?

Paper ID PMC2443636
Title Mutation in murine coronavirus replication protein nsp4 alters assembly of double membrane vesicles
Rank 1
Score 82.96
Authors Clementz, Mark A.; Kanjanahaluethai, Amornrat; O'Brien, Timothy E.; Baker, Susan C.
Answer Excerpt However the mechanism by which this substitution in nsp4 causes the defect in RNA synthesis in Alb ts6 is not known. A schematic diagram of nsp4 topology indicating the position of the two asparagine residues modified by Nlinked glycosylation and an asparagine to threonine change predicted to be responsible for the temperature sensitive phenotype are depicted in Fig 1D. To determine if nsp4N176 nsp4N237 or nsp4N258 is important for nsp4 function we generated virus encoding each specific substitution. Each substitution was introduced into the MHVA59 genome using a reverse genetics approach pioneered by Yount et al. 2002 as described in the Materials and methods

Evaluation

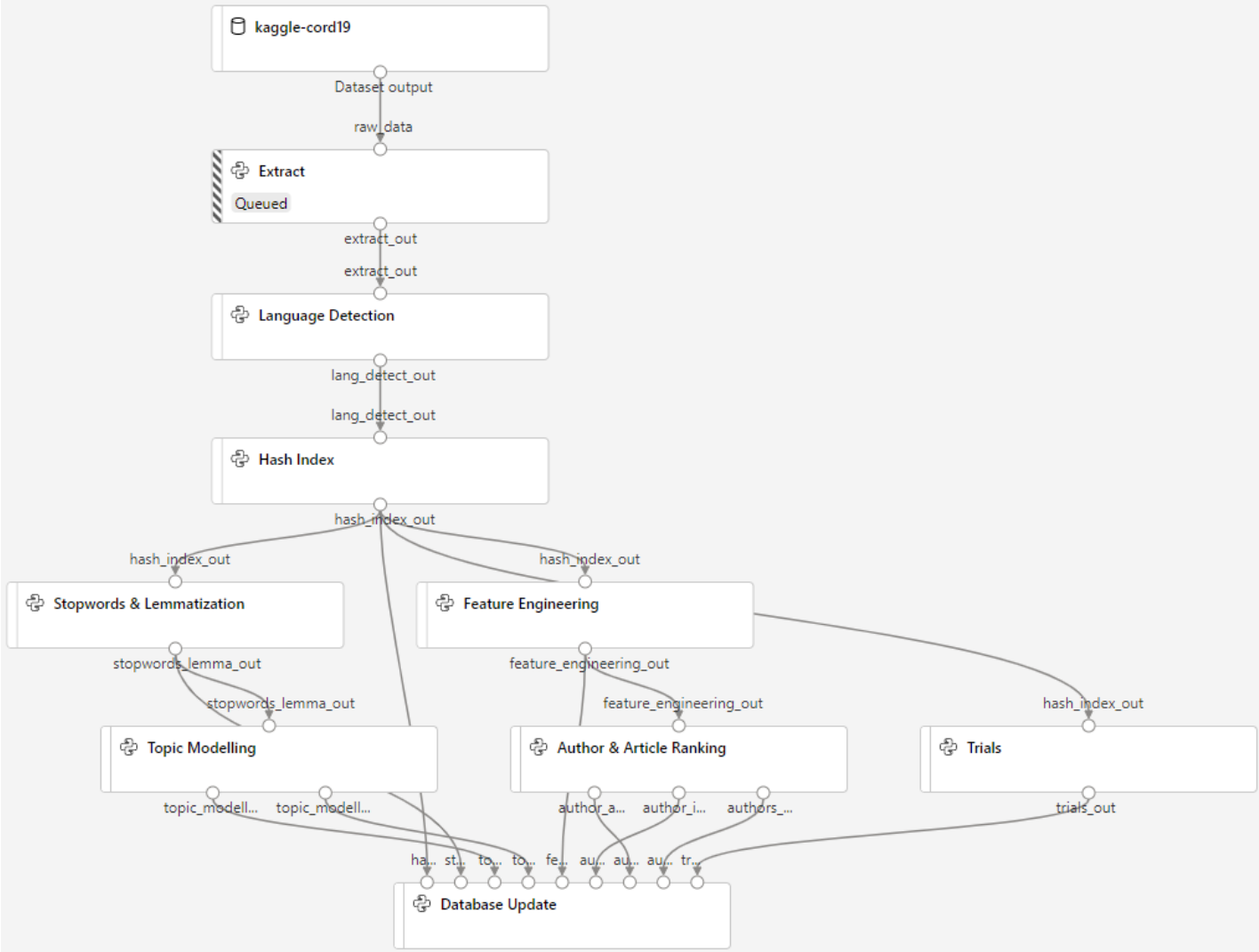
How To Measure How Good NLP Solutions Are?

- "*Good*" can have multiple meanings and is especially subjective in NLP
- Intrinsic evaluation of tasks like search can use existing performance benchmarks. It can be automated and set up by the technology team
- Extrinsic evaluation focuses on the human side and how well a solution solves a business problem. The metrics and process for evaluation typically established by the human expert
- If a solution performs well with intrinsic evaluation but fails to address business requirements, is it still considered a success?

An NLP Pipeline

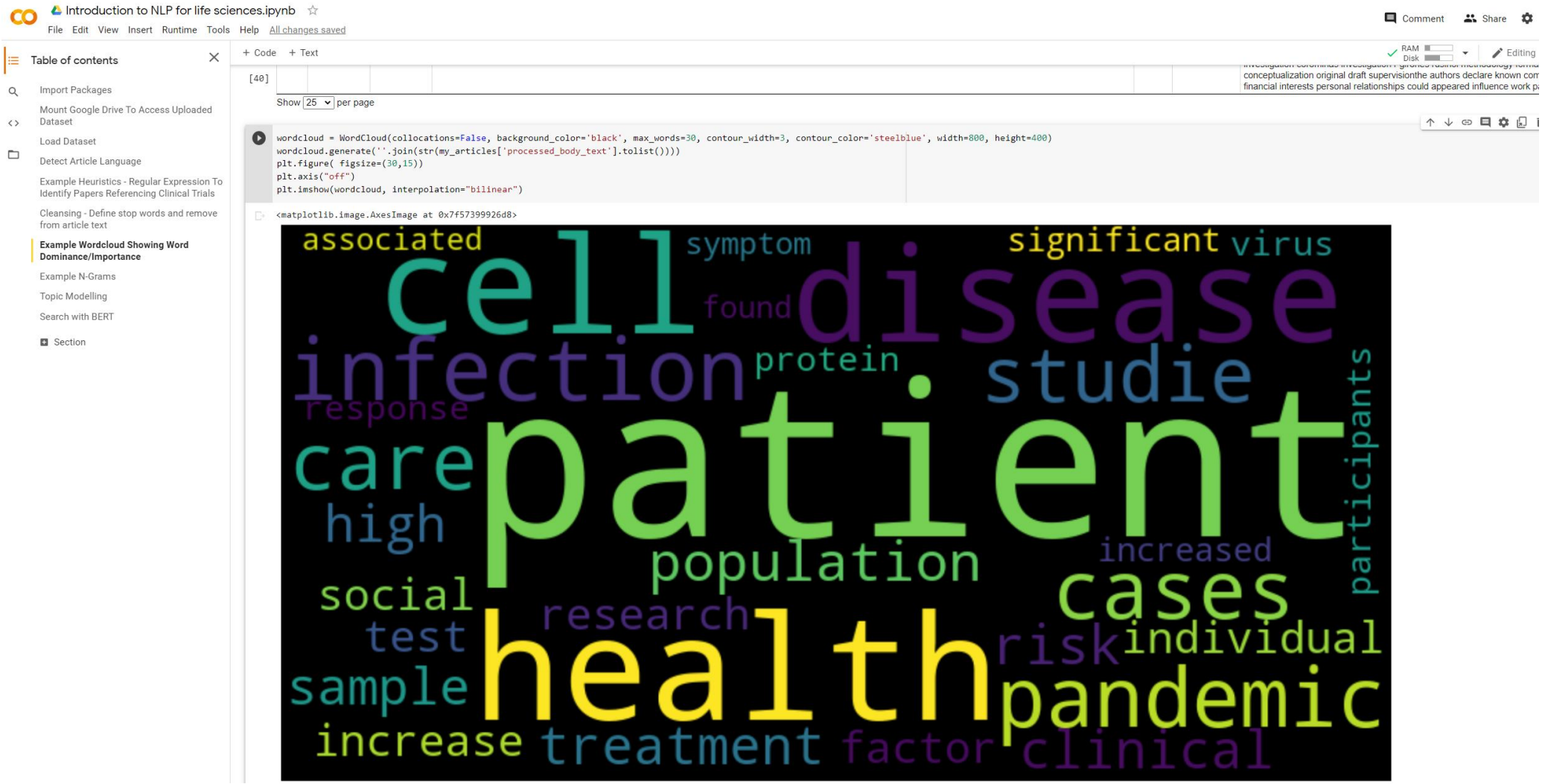
Decomposing the text problem into several sub-problems, step-by-step

- An NLP processing system is often referred to as a pipeline because natural language flows in one end, is processed through several steps, and the output flows out the other end



Demo

- Under the hood demo, showing a little of what can be done right now in the domain of Covid-19 literature



Further Thoughts

- What do you think? What are your big issues you face as a researcher where NLP might help?

Links

- Google Colab for development of NLP/Machine Learning within a browser environment
- Anaconda Data Science Toolkit
- Example Introduction to NLP for Life Sciences notebook with Cobvid-19 dataset