Capstone Datasets Explored

1)

[TensorFlow 2.0 Question Answering | Kaggle](https://www.kaggle.com/competitions/tensorflow2-question-answering/data)

Dataset files: v1.0-simplified\_simplified-nq-train.jsonl.gz

columns/fields

dict\_keys(['annotations', 'document\_html', 'document\_title', 'document\_tokens', 'document\_url', 'example\_id', 'long\_answer\_candidates', 'question\_text', 'question\_tokens'])

Nores: It has answer candidates but not labels.  Is this meant to use other pretrained models?

It is the HTML from wiki pages.  They give start and end character locations for the candidates.

2)

[Google QUEST Q&A Labeling | Kaggle](https://www.kaggle.com/competitions/google-quest-challenge)

Dataset files: google-quest-challenge.zip

In this competition, you’re challenged to use this new dataset to build predictive algorithms for different subjective aspects of question-answering. The question-answer pairs were gathered from nearly 70 different websites, in a "common-sense" fashion. Our raters received minimal guidance and training, and relied largely on their subjective interpretation of the prompts. As such, each prompt was crafted in the most intuitive fashion so that raters could simply use their common-sense to complete the task. By lessening our dependency on complicated and opaque rating guidelines, we hope to increase the re-use value of this data set. What you see is what you get!

Columns/fields

Index(['qa\_id', 'question\_title', 'question\_body', 'question\_user\_name',

       'question\_user\_page', 'answer', 'answer\_user\_name', 'answer\_user\_page',

       'url', 'category', 'host', 'question\_asker\_intent\_understanding',

       'question\_body\_critical', 'question\_conversational',

       'question\_expect\_short\_answer', 'question\_fact\_seeking',

       'question\_has\_commonly\_accepted\_answer',

       'question\_interestingness\_others', 'question\_interestingness\_self',

       'question\_multi\_intent', 'question\_not\_really\_a\_question',

       'question\_opinion\_seeking', 'question\_type\_choice',

       'question\_type\_compare', 'question\_type\_consequence',

       'question\_type\_definition', 'question\_type\_entity',

       'question\_type\_instructions', 'question\_type\_procedure',

       'question\_type\_reason\_explanation', 'question\_type\_spelling',

       'question\_well\_written', 'answer\_helpful',

       'answer\_level\_of\_information', 'answer\_plausible', 'answer\_relevance',

       'answer\_satisfaction', 'answer\_type\_instructions',

       'answer\_type\_procedure', 'answer\_type\_reason\_explanation',

       'answer\_well\_written'],

      dtype='object')

Computers are really good at answering questions with single, verifiable answers. But, humans are often still better at answering questions about opinions, recommendations, or personal experiences.

Humans are better at addressing subjective questions that require a deeper, multidimensional understanding of context - something computers aren't trained to do well…yet.. Questions can take many forms - some have multi-sentence elaborations, others may be simple curiosity or a fully developed problem. They can have multiple intents, or seek advice and opinions. Some may be helpful and others interesting. Some are simple right or wrong.

Unfortunately, it’s hard to build better subjective question-answering algorithms because of a lack of data and predictive models. That’s why the [CrowdSource](https://crowdsource.google.com/) team at Google Research, a group dedicated to advancing NLP and other types of ML science via crowdsourcing, has collected data on a number of these quality scoring aspects.

In this competition, you’re challenged to use this new dataset to build predictive algorithms for different subjective aspects of question-answering. The question-answer pairs were gathered from nearly 70 different websites, in a "common-sense" fashion. Our raters received minimal guidance and training, and relied largely on their subjective interpretation of the prompts. As such, each prompt was crafted in the most intuitive fashion so that raters could simply use their common-sense to complete the task. By lessening our dependency on complicated and opaque rating guidelines, we hope to increase the re-use value of this data set. What you see is what you get!

Demonstrating these subjective labels can be predicted reliably can shine a new light on this research area. Results from this competition will inform the way future intelligent Q&A systems will get built, hopefully contributing to them becoming more human-like.

3)

[U.S. Patent Phrase to Phrase Matching | Kaggle](https://www.kaggle.com/competitions/us-patent-phrase-to-phrase-matching/data)

Dataset file: us-patent-phrase-to-phrase-matching.zip

In this dataset, you are presented pairs of phrases (an anchor and a target phrase) and asked to rate how similar they are on a scale from 0 (not at all similar) to 1 (identical in meaning). This challenge differs from a standard semantic similarity task in that similarity has been scored here within a patent's context, specifically its [CPC classification (version 2021.05)](https://en.wikipedia.org/wiki/Cooperative_Patent_Classification), which indicates the subject to which the patent relates. For example, while the phrases "bird" and "Cape Cod" may have low semantic similarity in normal language, the likeness of their meaning is much closer if considered in the context of "house".

This is a code competition, in which you will submit code that will be run against an unseen test set. The unseen test set contains approximately 12k pairs of phrases. A small public test set has been provided for testing purposes, but is not used in scoring.

Information on the meaning of CPC codes may be found on the [USPTO website](https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html). The CPC version 2021.05 can be found on the [CPC archive website](https://www.cooperativepatentclassification.org/Archive).

Index(['id', 'anchor', 'target', 'context', 'score'], dtype='object')

Can you extract meaning from a large, text-based dataset derived from inventions? Here's your chance to do so.

The U.S. Patent and Trademark Office (USPTO) offers one of the largest repositories of scientific, technical, and commercial information in the world through its [Open Data Portal](https://developer.uspto.gov/about-open-data). Patents are a form of [intellectual property](https://www.uspto.gov/patents/basics/general-information-patents) granted in exchange for the public disclosure of new and useful inventions. Because patents undergo an intensive [vetting process](https://www.uspto.gov/sites/default/files/documents/InventionCon2020_Understanding_the_Patent_Examination_Process.pdf) prior to grant, and because the history of U.S. innovation spans over two centuries and 11 million patents, the U.S. patent archives stand as a rare combination of data volume, quality, and diversity.

“The USPTO serves an American innovation machine that never sleeps by granting patents, registering trademarks, and promoting intellectual property around the globe. The USPTO shares over 200 years' worth of human ingenuity with the world, from lightbulbs to quantum computers. Combined with creativity from the data science community, USPTO datasets carry unbounded potential to empower AI and ML models that will benefit the progress of science and society at large.”

— USPTO Chief Information Officer Jamie Holcombe

In this competition, you will train your models on a novel semantic similarity dataset to extract relevant information by matching key phrases in patent documents. Determining the semantic similarity between phrases is critically important during the patent search and examination process to determine if an invention has been described before. For example, if one invention claims "television set" and a prior publication describes "TV set", a model would ideally recognize these are the same and assist a patent attorney or examiner in retrieving relevant documents. This extends beyond paraphrase identification; if one invention claims a "strong material" and another uses "steel", that may also be a match. What counts as a "strong material" varies per domain (it may be steel in one domain and ripstop fabric in another, but you wouldn't want your parachute made of steel). We have included the Cooperative Patent Classification as the technical domain context as an additional feature to help you disambiguate these situations.

Can you build a model to match phrases in order to extract contextual information, thereby helping the patent community connect the dots between millions of patent documents?

## Score meanings

The scores are in the 0-1 range with increments of 0.25 with the following meanings:

* 1.0 - Very close match. This is typically an exact match except possibly for differences in conjugation, quantity (e.g. singular vs. plural), and addition or removal of stopwords (e.g. “the”, “and”, “or”).
* 0.75 - Close synonym, e.g. “mobile phone” vs. “cellphone”. This also includes abbreviations, e.g. "TCP" -> "transmission control protocol".
* 0.5 - Synonyms which don’t have the same meaning (same function, same properties). This includes broad-narrow (hyponym) and narrow-broad (hypernym) matches.
* 0.25 - Somewhat related, e.g. the two phrases are in the same high level domain but are not synonyms. This also includes antonyms.
* 0.0 - Unrelated.

4)

[UNIFESP X-ray Body Part Classifier Competition | Kaggle](https://www.kaggle.com/competitions/unifesp-x-ray-body-part-classifier/data)

Dataset file: unifesp-x-ray-body-part-classifier.zip

Classifying a body part from an x-ray image might seem silly, but having it automated can be a key for all the world around deep learning in medical imaging. In many hospitals, when a physician orders multiple imaging exams one accession number is created for each body part (eg. knee, ankle, and leg), but the registration for the correspondent images are often incorrect within each accession number, here is an example:

In this challenge, competitors are predicting the body part from a given x-ray.

When making predictions, competitors should predict as many body parts per image as they judge necessary

There should be 1 predicted column per image - and the labels are represented as integers that map each to one body part contained in the dataset.

A properly formatted row may look like any of the following.

SOPInstanceUID, 0 12

The labels are represented as integers that map to the following:

* Abdomen = 0
* Ankle = 1
* Cervical Spine = 2
* Chest = 3
* Clavicles = 4
* Elbow = 5
* Feet = 6
* Finger = 7
* Forearm = 8
* Hand = 9
* Hip = 10
* Knee = 11
* Lower Leg = 12
* Lumbar Spine = 13
* Others = 14
* Pelvis = 15
* Shoulder = 16
* Sinus = 17
* Skull = 18
* Thigh = 19
* Thoracic Spine = 20
* Wrist = 21

Note - Others indicates whether the sample contains image of non X-ray images that sometimes are misplaced in the PACS system as X-Ray (eg. esophagogram, densitometry).