**The problem Statement:**

To bifurcate the hard skills and soft skills from the given raw data of listed skills.

**Data**:

Raw\_skills\_data: List of up to 30,000 + skills which needs to bifurcated as per the requirements

Technical\_skills\_data: List of technical skills as examples

**Methodology adopted:**

Use the sentence transformers to generate the embeddings and find out the cosine similarity to know which ones are hard skills.

**Some keywords to help understand the project:**

1. Sentence transformer: Sentence Transformers is a Python framework for state-of-the-art sentence, text and image embeddings. You can use this framework to compute sentence / text embeddings for more than 100 languages. These embeddings can then be compared e.g. with cosine-similarity to find sentences with a similar meaning. This can be useful for semantic textual similar, semantic search, or paraphrase mining.

2. Sentence Transformer model built: ‘Distilbert base-nli-stsb-mean-tokens’ model in sentence transformer compute vector representations for sentences and paragraphs. The model is based on DistilBERT and perform state-of-the-art results in various tasks. Text is embedding in vector space such that similar text is close and can efficiently be found using cosine similarity.

3. Cosine Similarity**: Cosine similarity** is a metric, helpful in determining, how similar the data objects are irrespective of their size. We can measure the similarity between two sentences in Python using Cosine Similarity. In cosine similarity, data objects in a dataset are treated as a vector.

**Step-by-step Process:**

1. Get the raw skills and skills described in the example data i.e Examples\_skills\_technical.csv

2. Pass the series of list through the sentence model transformer built

3. Find the cosine similarity of each raw skill with every example skill present

4. Set a threshold to decide whether or not the skill mentioned is a hard skill

5. Based on the threshold define the similarity found out

6. Mode of the all the answers returned for every try would define whether or not the skill mentioned is a hard skill