Final Project

Unsupervised Learning for Solar Job Posting Analysis

https://github.com/igrosny/csca5632/blob/main/Final_Project.ipynb

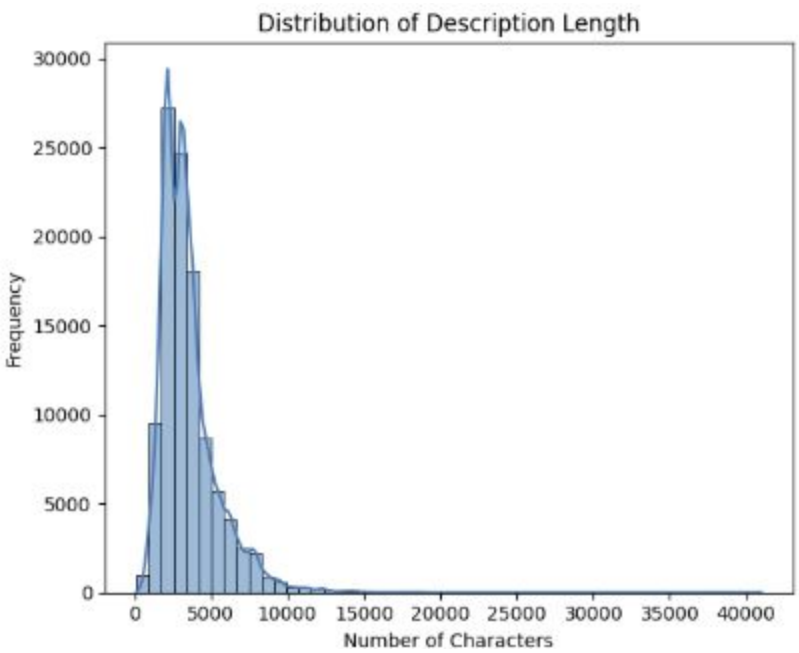
Introduction & Project Goal

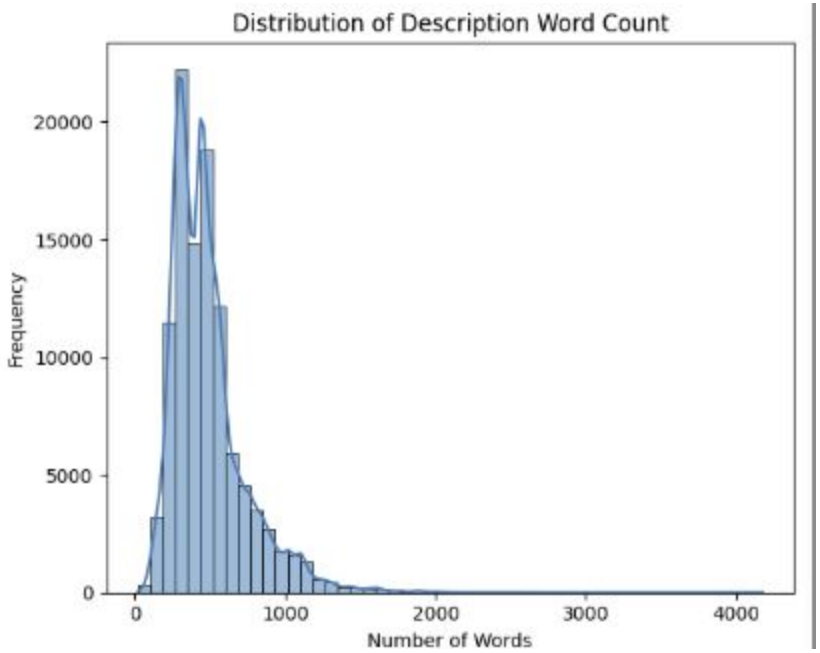
Goal: To analyze a dataset of job postings related to the solar industry using unsupervised machine learning techniques.

The Unsupervised Learning Problem

Analyzing large amounts of unstructured text data (job descriptions) to find meaningful patterns without manual labeling.

Exploratory Data Analysis (EDA)





1 - Clustering with Sentence Transformers & K-Means

```
NUM_CLUSTERS = 50
MODEL_NAME = 'all-MiniLM-L6-v2'
job_postings = df['doc_description'].tolist()

print(f"Loading pre-trained sentence transformer model: {MODEL_NAME}...")
model = SentenceTransformer(MODEL_NAME)
print(f"Generating embeddings for {len(job_postings)} job postings...")
embeddings = model.encode(job_postings, show_progress_bar=True)

print(f"Embeddings generated. Shape: {embeddings.shape}")

print(f"\nClustering embeddings into {NUM_CLUSTERS} groups using K-Means...")


kmeans = KMeans(n_clusters=NUM_CLUSTERS,
                random_state=42, # for reproducibility
                n_init='auto')

kmeans.fit(embeddings)

cluster_labels = kmeans.labels_
```

1 - Clustering with Sentence Transformers & K-Means

Group similar job descriptions together automatically.

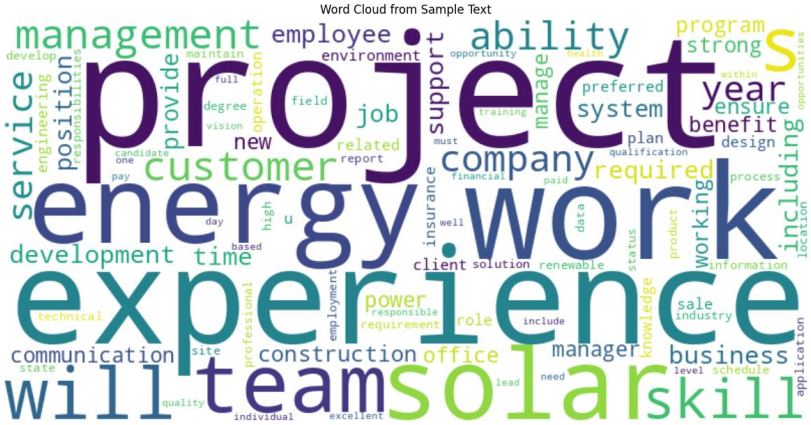
```
Cluster 0: Contains 3356 postings
-----------
  - Posting 47: Human Resources/Office Administrator.
  - Posting 48: SOLAR CONTRACTS ADMINISTRATOR...
  - Posting 105: Solar Operations Center Manager - Re
  - Posting 133: Logistics Coordinator...
  - Posting 172: Warehouse Assistant...
  - Posting 205: UX/UI Designer...
  - Posting 251: International Logistics Coordinator.
  - Posting 264: Accounts Payable Clerk...
  - Posting 268: Solar Field Tech Leader...
  - Posting 272: Director of O&M - Solar...
  ... (and 3346 more postings in this cluster)
```

```
Top 20 words in this cluster:
  - solar: 12793
  - experience: 9283
  - energy: 7736
  - all: 7735
  - ability: 7133
  - customer: 6053
  - other: 5786
  - team: 5242
  - company: 4959
  - skills: 4896
  - service: 4427
  - management: 4140
  - support: 4129
  - insurance: 4056
  - ensure: 4052
```

2 - Topic Modeling with TF-IDF & NMF

```
cleaned_postings = [clean_text(post) for post in job_postings]

print("Vectorizing text data using TF-IDF...")

vectorizer = TfidfVectorizer(max_df=0.95,
                             min_df=2,
                             stop_words='english',
                             ngram_range=(1,1))

dtm_tfidf = vectorizer.fit_transform(cleaned_postings)
feature_names = vectorizer.get_feature_names_out()

print(f"Applying NMF to find {NUM_TOPICS} topics...")


nmf_model = NMF(n_components=NUM_TOPICS,
                random_state=42,
                init='nndsvda',
                max_iter=500,
                l1_ratio=0.0,
                solver='cd')
```

Cluster visualization



Comparison & Challenges

K-Means Good at grouping entire documents based on overall similarity. Useful for finding comparable roles.

NMF: Effective at identifying underlying themes or skill sets that might appear across different types of jobs.

Overlap: Some clusters (K-Means) might strongly align with specific topics (NMF), but they offer different perspectives.

Challenge: Both methods output lists of keywords or example documents. Manually interpreting and assigning meaningful labels to these topics/clusters is time-consuming and subjective.

Conclusion

Unsupervised methods successfully revealed structure (clusters of similar jobs, latent topics) within the solar job posting data.

Key Insight: Demonstrated the ability to automatically categorize and theme large text datasets without prior labels.

Bottleneck: Manual labeling of topics/clusters.

Future Direction: Explore using Large Language Models (LLMs) to automate the generation of human-understandable topic/cluster labels based on the keywords and document examples, significantly reducing manual effort.